6.0 PERFORMANCE CHARACTERISTICS OF FETAX

The performance characteristics (i.e., accuracy, sensitivity, specificity, positive predictivity, negative predictivity, false positive rate, and false negative rate) of FETAX compared to either rat, mice, and/or rabbit teratogenicity test results or human teratogenicity study results were determined by NICEATM. FETAX studies that did not follow the ASTM FETAX Guideline (1991, 1998), especially in regard to data presentation and analysis, were excluded from consideration of performance characteristics. The decision criteria used in determining the performance characteristics of FETAX included:

- single decision criteria (TI >1.5; MCIG/LC_{50} <0.30) for identifying teratogenic potential, as defined by the ASTM FETAX Guideline (1991, 1998);

- modified single decision criterion (TI >3.0) for identifying teratogenic potential, as used in a recent study by Fort et al. (2000a);

- multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.30) for identifying teratogenic potential, as used in FETAX Validation Study Phase III.3 (Bantle et al., 1999); and

- multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.30) for identifying teratogenic potential, as used in a recent study by Fort et al. (2000a).

In the ASTM FETAX Guideline (1991, 1998), a TI value greater than 1.5, an MCIG/LC_{50} ratio less than 0.30, or the presence of severe malformations was considered to be indicative of teratogenic activity. In the FETAX Phase III.3 Validation Study (Bantle et al., 1999), multiple as well as single criteria were used. When multiple criteria were used, test substances were classified as positive when both the TI value was greater than 1.5 and the MCIG/LC_{50} ratio was less than 0.3, equivocal when either but not both criteria were positive, and negative when neither criteria was positive. Where results were classified as equivocal, information on the severity of the observed malformations was used to potentially resolve the classification. In the Fort et al. (2000a) study, single and multiple criteria were used as described in the Phase III.3

Validation Study, except that the critical TI decision value was increased from 1.5 to 3.0; values between 1.5 and 3.0 were considered to be suggestive. In the performance analysis conducted by NICEATM, information on the types and incidence of malformations induced in *X. laevis* embryos were excluded from the evaluations due to the almost complete absence of quantitative data on malformations in the published FETAX literature. For substances that were evaluated multiple times in FETAX, the NICEATM consensus FETAX result was based on a simple weight-of-evidence approach; test substances with an equal number of positive and negative studies were classified as equivocal and were excluded from the performance calculations. In the performance calculations presented herein, the numbers in parenthesis after a percentage value are the number of correct results divided by the total number of test substances considered. Differences in the total number of FETAX test substances considered under apparently identical conditions are due to differences in available data or from the exclusion of test substances with an equivocal classification for that particular decision criteria. Where multiple criteria, equivocal FETAX results were encountered, performance characteristics were calculated excluding equivocal FETAX results, including equivocal FETAX results as positive, or including equivocal FETAX results as negative. The FETAX, laboratory mammal, and human teratogenicity results used in these analyses are summarized in **Appendix 5**.

6.1 Performance Characteristics of FETAX compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX compared to combined rat, mouse, and rabbit teratogenicity test results were determined using three approaches. Performance characteristics were calculated based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. In the latter analysis, a substance tested with and without metabolic activation was classified as positive in FETAX if a consensus positive response was obtained either with or without metabolic activation. A test substance tested with and without metabolic activation was classified as a FETAX negative only if a consensus positive response was not obtained using either exposure condition. In addition to these analysis conducted using the total FETAX database, the performance characteristics were determined by chemical and product class for

FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results. For the evaluation of FETAX compared to teratogenicity data obtained from combined rat, mouse, and rabbit studies, a substance was classified as a laboratory mammal teratogen if a positive result was reported for any of the three species. In contrast, test substances positive in one, but not another, species were classified as equivocal by the investigators in the FETAX Phase III.3 Validation Study (Bantle et al., 1999) and in the comparative study conducted by Fort et al. (2000a).

6.1.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 5**).

<u>Single Decision Criteria:</u> Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC_{50} <0.3),

- accuracy varied from 54% (40/74) to 63% (57/90),
- sensitivity from 40% (16/40) to 78% (39/50),
- specificity from 45% (18/40) to 71% (24/34),
- positive predictivity from 62% (23/37 and 16/26) to 64% (39/61),
- negative predictivity from 50% (24/48 and 26/52) to 62% (18/29),
- false positive rate from 29% (10/34) to 55% (22/40), and
- false negative rate from 22% (11/58) to 60% (24/40).

Maximal accuracy and sensitivity, but minimal specificity, occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), and when equivocal results were excluded from the evaluation,

- accuracy was 63% (31/49),
- sensitivity was 67% (16/24),
- specificity was 60% (15/25),
- positive predictivity was 62% (16/26),
- negative predictivity was 65% (15/23),
- false positive rate was 40% (10/25), and
- false negative rate was 33% (8/24).

When equivocal results were re-classified as positives and included in the analysis,

- accuracy was 63% (46/73),
- sensitivity was 79% (31/39),
- specificity was 44% (15/34),
- positive predictivity was 62% (31/50),
- negative predictivity was 65% (15/23),
- false positive rate was 56% (19/34), and
- false negative rate was 21% (8/39).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 55% (40/73),
- sensitivity was 41% (16/39),
- specificity was 71% (24/34),
- positive predictivity was 62% (16/26),
- negative predictivity was 51% (24/47),
- false positive rate was 29% (10/34), and
- false negative rate was 59% (23/39).

Sensitivity was increased when equivocal results were re-classified as positives and included in the analysis, while specificity was increased when equivocal results were re-classified as negatives and included in the analysis. Accuracy was not increased when equivocal calls were re-classified as positives or negatives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 58% (36/62),
- sensitivity was 47% (14/30),
- specificity was 69% (22/32),
- positive predictivity was 58% (14/24),
- negative predictivity was 58% (22/38),
- false positive rate was 31% (10/32), and
- false negative rate was 53% (16/30).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 61% (44/72),
- sensitivity was 58% (22/38),
- specificity was 65% (22/34),
- positive predictivity was 65% (22/34),
- negative predictivity was 58% (22/38),
- false positive rate was 35% (12/34), and
- false negative rate was 42% (16/38).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 53% (38/72),
- sensitivity was 37% (14/38),
- specificity was 71% (24/34),
- positive predictivity was 58% (14/24),
- negative predictivity was 50% (24/48),
- false positive rate was 29% (10/34), and
- false negative rate was 63% (24/38).

Accuracy appeared to be optimal when equivocal responses were re-classified as positives and included in the analysis, while sensitivity and specificity were optimal when equivocal responses were re-classified as positives or negative, respectively, and included in the analysis.

The performance characteristics for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were generally not improved by using multiple decision criteria. The use of a single decision criterion based on a TI value greater than 1.5 resulted in increased accuracy and sensitivity over one based on a TI value greater than 3.0.

6.1.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX, with metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 6**).

Single Decision Criteria: Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 42% (11/26) to 56% (15/27),
- sensitivity ranged from 20% (2/10) to 87% (13/15),
- specificity from 17% (2/12) to 70% (7/10),
- positive predictivity from 40% (2/5) to 57% (13/23),
- negative predictivity from 40% (6/15) to 50% (2/4),
- false positive rate from 30% (3/10) to 83% (10/12), and
- false negative rate from 13% (2/15) to 80% (8/10).

Maximal accuracy and sensitivity occurred when the single decision criterion was a TI value greater than 1.5. However, specificity was highest when an MCIG/LC$_{50}$ ratio less than 0.3 was used as the single decision criterion.

Multiple Decision Criteria: Using multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 50% (4/8),
- sensitivity was 67% (2/3),
- specificity was 40% (2/5),
- positive predictivity was 40% (2/5),
- negative predictivity was 67% (2/3),
- false positive rate was 60% (3/5), and
- false negative rate was 33% (1/3).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 55% (11/20),
- sensitivity was 90% (9/10),
- specificity was 20% (2/10),
- positive predictivity was 53% (9/17),
- negative predictivity was 67% (2/3),
- false positive rate was 80% (8/10), and
- false negative rate was 10% (1/10).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 45% (9/20),
- sensitivity was 20% (2/10),
- specificity was 70% (7/10),
- positive predictivity was 40% (2/5),
- negative predictivity was 47% (7/15),
- false positive rate was 30% (3/10), and
- false negative rate was 80% (8/10).

Accuracy and sensitivity but not specificity were maximal when equivocal calls were re-classified as positives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 40% (6/15),
- sensitivity was 13% (1/8),
- specificity was 71% (5/7),
- positive predictivity was 33% (1/3),
- negative predictivity was 42% (5/12),
- false positive rate was 29% (2/7), and
- false negative rate was 88% (7/8).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 37% (7/19),
- sensitivity was 22% (2/9),
- specificity was 50% (5/10),
- positive predictivity was 29% (2/7),
- negative predictivity was 42% (5/12),
- false positive rate was 50% (5/10), and
- false negative rate was 78% (7/9).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 47% (9/19),
- sensitivity was11% (1/9),
- specificity was 80% (8/10),
- positive predictivity was 33% (1/3),
- negative predictivity was 50% (8/16),
- false positive rate was 20% (2/10), and
- false negative rate was 89% (8/9).

Accuracy and specificity were slightly better when equivocal results were classified as positive and included in the analysis.

Accuracy and sensitivity but not specificity were improved compared to combined rat, mouse, and rabbit teratogenicity results when a TI value greater than 1.5 rather than 3.0 was used as the decision criteria. Performance was not generally improved when multiple decision criteria were used.

6.1.3 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The overall performance characteristics of FETAX, with and without metabolic activation, compared to the combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 7**).

<u>Single Decision Criteria:</u> Based on the use of a single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 53% (48/90) to 61% (55/90),
- sensitivity from 43% (17/40) to 82% (41/50),
- specificity from 35% (14/40) to 68% (23/34),
- positive predictivity was 61% (17/28, 23/38, and 41/67),
- negative predictivity from 48% (23/46) to 61% (14/23),
- false positive rate from 32% (11/34) to 65% (26/40), and
- false negative rate from 18% (9/50) to 58% (23/40).

Maximal accuracy and sensitivity occurred when the single decision criterion was a TI value greater than 1.5, while maximal specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio of less than 0.3.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 63% (29/46),
- sensitivity was 74% (17/23),
- specificity was 52% (12/23),
- positive predictivity was 61% (17/28),
- negative predictivity was 67% (12/18),
- false positive rate was 48% (11/23), and
- false negative rate was 26% (6/23).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 62% (45/73),
- sensitivity was 85% (33/39),
- specificity was 35% (12/34),
- positive predictivity was 60% (33/55),
- negative predictivity was 67% (12/18),
- false positive rate was 65% (22/34), and
- false negative rate was 15% (6/39).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 55% (40/73),
- sensitivity was 44% (17/39),
- specificity was 68% (23/34),
- positive predictivity was 61% (17/28),
- negative predictivity was 51% (23/45),
- false positive rate was 32% (11/34), and
- false negative rate was 56% (22/39).

Accuracy and sensitivity were similar when equivocal response were excluded from the analysis or re-classified as positives and included in the analysis; specificity was optimal when equivocal responses were re-classified as positives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 58% (35/60),
- sensitivity was 48% (14/29),
- specificity was 68% (21/31),
- positive predictivity was 58% (14/24),
- negative predictivity was 58% (21/36),
- false positive rate was 32% (10/31), and
- false negative rate was 52% (15/29).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 62% (45/73),
- sensitivity was 62% (24/39),
- specificity was 62% (21/34),
- positive predictivity was 65% (24/37),
- negative predictivity was 58% (21/36),
- false positive rate was 38% (13/34), and
- false negative rate was 38% (15/39).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 52% (38/73),
- sensitivity was 36% (14/39),
- specificity was 71% (24/34),
- positive predictivity was 58% (14/24),
- negative predictivity was 49% (24/49),
- false positive rate was 29% (10/34), and
- false negative rate was 64% (25/39).

With the exception of specificity, performance appeared to be optimal when equivocal calls were re-classified as positives and included in the analysis. In general, a FETAX decision criteria based on the use of a TI value greater than 3.0 was not as accurate as one based on using a TI value greater than 1.5.

Based on an analysis of the performance characteristics for FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results, the use of single decision criterion based on a TI value greater than 1.5 rather than 3.0 appeared to provide the most optimal approach in terms of accuracy and sensitivity. The use of multiple decision criteria did not appreciable improve FETAX performance.

6.1.4 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Combined Rat, Mouse, or Rabbit Teratogenicity Test Results by Chemical and Product Class

The most numerically prevalent classes were alcohols (including glycols), amides, amines, halogenated organic compounds, esters, heavy metals and their salts, hydrazides and hydrazines, nitrogen heterocyclic compounds, organic (phenolic and carboxylic) acids, and salts (see **Section 3.3**). The most common product classes tested in FETAX were antimicrobials, chemical synthesis materials, cosmetics, dyes, food additives, fossil fuels, pesticides, pharmaceuticals, photographic chemicals, and polymers (including monomers). The performance characteristics of FETAX, using with and without metabolic activation studies combined were compared to combined rat, mouse, and rabbit teratogenicity results by chemical and product class using single decision criteria (i.e., TI >1.5, TI >3.0, MCIG/LC_{50} <0.3) (**Tables 8**, **9**, and **10**, respectively). Analyses were limited to those chemical and product classes that included a minimum of 15 substances tested in FETAX for which there was also laboratory mammal teratogenicity test results. For comparative purposes, the corresponding performance characteristics when all FETAX data were considered are included in each table.

<u>Amides plus Hydrazides:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 56% (9/16),
- sensitivity was 78% (7/9),
- specificity was 29% (2/7),
- positive predictivity was 58% (7/12),
- negative predictivity was 50% (2/4),
- false positive rate was 71% (5/7), and
- false negative rate was 22% (2/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 44% (7/16),
- sensitivity was 44% (4/9),
- specificity was 43% (3/7),
- positive predictivity was 50% (4/8),
- negative predictivity was 38% (3/8),
- false positive rate was 57% (4/7), and
- false negative rate was 56% (5/9).

Due to the absence of a sufficient database, performance characteristics using a decision criterion based on an MCIG/LC_{50} ratio of less than 0.3 were not determined.

<u>Amines:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 89% (8/9),
- specificity was 17% (1/6),
- positive predictivity was 62% (8/13),
- negative predictivity was 50% (1/2),
- false positive rate was 83% (5/6), and
- false negative rate was 11% (1/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 67% (6/9),
- specificity was 50% (3/6),
- positive predictivity was 67% (6/9),
- negative predictivity was 50% (3/6),
- false positive rate was 50% (3/6), and
- false negative rate was 33% (3/9).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 53% (8/15),
- sensitivity was 56% (5/9),
- specificity was 50% (3/6),
- positive predictivity was 63% (5/8),
- negative predictivity was 43% (3/7),
- false positive rate was 50% (3/6), and
- false negative rate was 44% (4/9).

<u>Nitrogen Heterocyclic Compounds:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 70% (21/30),
- sensitivity was 80% (16/20),
- specificity was 50% (5/10),
- positive predictivity was 76% (16/21),
- negative predictivity was 56% (5/9),
- false positive rate was 50% (5/10), and
- false negative rate was 20% (4/20).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 53% (16/30),
- sensitivity was 50% (10/20),
- specificity was 60% (6/10),
- positive predictivity was 71% (10/14),
- negative predictivity was 38% (6/16),
- false positive rate was 40% (4/10), and
- false negative rate was 50% (10/20).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 48% (11/23),
- sensitivity was 53% (8/15),
- specificity was 50% (4/8),
- positive predictivity was 67% (8/12),
- negative predictivity was 36% (4/11),
- false positive rate was 50% (4/8), and
- false negative rate was 47% (7/15).

Organic (Phenolic and Carboxylic) Acids: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 80% (16/20),
- sensitivity was 86% (12/14),
- specificity was 67% (4/6),
- positive predictivity was 86% (12/14),
- negative predictivity was 67% (4/6),
- false positive rate was 33% (2/6), and
- false negative rate was 14% (2/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 60% (12/20),
- sensitivity was 43% (6/14),
- specificity was 100% (6/6),
- positive predictivity was 100% (6/6),
- negative predictivity was 43% (6/14),
- false positive rate was 0% (0/6), and
- false negative rate was 57% (8/14).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 40% (4/10),
- specificity was 100% (5/5),
- positive predictivity was 100% (4/4),
- negative predictivity was 45% (5/11),
- false positive rate was 0% (0/5), and
- false negative rate was 60% (6/10).

Pharmaceuticals: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 54% (21/39),
- sensitivity was 82% (18/22),
- specificity was 18% (3/17),
- positive predictivity was 56% (18/32),
- negative predictivity was 43% (3/7),
- false positive rate was 82% (14/17), and
- false negative rate was 18% (4/22).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 50% (19/38),
- sensitivity was 43% (9/21),
- specificity was 59% (10/17),
- positive predictivity was 56% (9/16),
- negative predictivity was 45% (10/22),
- false positive rate was 41% (7/17), and
- false negative rate was 57% (12/21).

Using an MCIG/LC$_{50}$ ratio less than 0.3 as the single decision criterion,

- accuracy was 53% (18/34),
- sensitivity was 47% (9/19),
- specificity was 60% (9/15),
- positive predictivity was 60% (9/15),
- negative predictivity was 47% (9/19),
- false positive rate was 40% (6/15), and
- false negative rate was 53% (10/19).

Due to the limited FETAX database, only five chemical classes and one product class were evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results. Among the chemical and product classes evaluated, a decision criterion based on a TI value greater than 1.5 generally provided greater accuracy and sensitivity, but less specificity, than one based on either on a TI value greater than 3.0 or on an MCIG/LC$_{50}$ ratio less than 0.3. In general, the accuracy of FETAX compared to laboratory mammal teratogenicity test results was somewhat improved for nitrogen heterocyclic compounds, and phenolic and carboxylic acids. Performance characteristics for the other chemical classes and the single product class evaluated were not appreciable different from the performance of FETAX compared to the total database.

6.2 Performance Characteristics of FETAX compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX compared to individual rat, mouse, and rabbit species teratogenicity test results were calculated using single TI decision criteria (TI >1.5, TI >3.0) only. Comparisons using other decision criteria (i.e., MCIG/LC$_{50}$ <0.30, various multiple decision criteria) were not conducted because of the inadequate numbers of comparisons available for the analysis. In this analysis, performance characteristics were determined based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. Performance characteristics based on chemical and product class for FETAX compared to individual rat, mouse, and rabbit species teratogenicity test results were not determined due to the paucity of the data. For the evaluation of FETAX compared to teratogenicity data obtained from combined rat, mouse, and rabbit studies, a substance was classified as a laboratory mammal teratogen if a positive result was reported for any of the three species.

6.2.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, without metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are provided in **Table 11**.

<u>FETAX versus Rat:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 61% (46/75),
- sensitivity was 77% (30/39),
- specificity was 44% (16/36),
- positive predictivity was 60% (30/50),
- negative predictivity was 64% (16/25),
- false positive rate was 56% (20/36), and
- false negative rate was 23% (9/39).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 51% (37/73),
- sensitivity was 43% (16/37),
- specificity was 58% (21/36),
- positive predictivity was 52% (16/31),
- negative predictivity was 50% (21/42),
- false positive rate was 42% (15/36), and
- false negative rate was 57% (21/37).

<u>FETAX versus Mouse:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 68% (45/66),
- sensitivity was 83% (33/40),
- specificity was 46% (12/26),
- positive predictivity was 70% (33/47),
- negative predictivity was 63% (12/19),
- false positive rate was 54% (14/26), and
- false negative rate was 18% (7/40).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 57% (37/65),
- sensitivity was 51% (20/39),
- specificity was 65% (17/26),
- positive predictivity was 69% (20/29),
- negative predictivity was 47% (17/36),
- false positive rate was 35% (9/26), and
- false negative rate was 49% (19/39).

<u>FETAX versus Rabbit:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 47% (16/34),
- sensitivity was 64% (9/14),
- specificity was 35% (7/20),
- positive predictivity was 41% (9/22),
- negative predictivity was 58% (7/12),
- false positive rate was 65% (13/20), and
- false negative rate was 36% (5/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 48% (16/33),
- sensitivity was 31% (4/13),
- specificity was 60% (12/20),
- positive predictivity was 33% (4/12),
- negative predictivity was 57% (12/21),
- false positive rate was 40% (8/20), and
- false negative rate was 69% (9/13).

Using either TI decision criteria value, the performance characteristics of FETAX, without metabolic activation, compared to teratogenicity data for rats and mice were quite similar, while that for rabbits appeared to be reduced. Furthermore, the performance characteristics compared to rats and mice were not different from the corresponding performance characteristics based on combined rat, mouse, and rabbit teratogenicity data (**Table 5**). Comparing the performance characteristics for each species as a function of the TI decision criterion value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0.

6.2.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, with metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are shown in **Table 12**.

<u>FETAX versus Rat:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 65% (15/23),
- sensitivity was 100% (11/11),
- specificity was 33% (4/12),
- positive predictivity was 58% (11/19),
- negative predictivity was 100% (4/4),
- false positive rate was 67% (8/12), and
- false negative rate was 0% (0/11).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 36% (8/22),
- sensitivity was 30% (3/10),
- specificity was 42% (5/12),
- positive predictivity was 42% (5/12),
- negative predictivity was 58% (7/12),
- false positive rate was 70% (7/10), and
- false negative rate was 30% (3/10).

<u>FETAX versus Mouse:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 52% (11/21),
- sensitivity was 85% (11/13),
- specificity was 0% (0/8),
- positive predictivity was 58% (11/19),
- negative predictivity was 0% (0/2),
- false positive rate was 100% (8/8), and
- false negative rate was 15% (2/13).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 40% (8/20),
- sensitivity was 42% (5/12),
- specificity was 38% (3/8),
- positive predictivity was 50% (5/10),
- negative predictivity was 30% (3/10),
- false positive rate was 63% (5/8), and
- false negative rate was 58% (7/12).

<u>FETAX versus Rabbit:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 53% (8/15),
- sensitivity was 100% (7/7),
- specificity was 13% (1/8),
- positive predictivity was 50% (7/14),
- negative predictivity was 100% (1/1),
- false positive rate was 88% (7/8), and
- false negative rate was 0% (0/7).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 50% (7/14),
- sensitivity was 33% (2/6),
- specificity was 63% (5/8),
- positive predictivity was 40% (3/5),
- negative predictivity was 56% (5/9),
- false positive rate was 38% (3/8), and
- false negative rate was 67% (4/6).

Using either TI decision criterion value, the performance characteristics of FETAX, with metabolic activation, compared to teratogenicity data for all three-laboratory species appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI decision criterion value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0. However, the validity of these conclusions is suspect because of the very limited number of substances tested with metabolic activation.

6.2.3 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, with and without metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are presented in **Table 13**.

<u>FETAX versus Rat:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 61% (46/75),
- sensitivity was 82% (32/39),
- specificity was 39% (14/36),
- positive predictivity was 59% (32/54),
- negative predictivity was 67% (14/21),
- false positive rate was 61% (22/36), and
- false negative rate was 18% (7/39).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 49% (36/74),
- sensitivity was 42% (16/38),
- specificity was 56% (20/36),
- positive predictivity was 50% (16/32),
- negative predictivity was 48% (20/42),
- false positive rate was 44% (16/36), and
- false negative rate was 58% (22/38).

<u>FETAX versus Mouse:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 64% (42/66),
- sensitivity was 85% (34/40),
- specificity was 31% (8/26),
- positive predictivity was 65% (34/52),
- negative predictivity was 57% (8/14),
- false positive rate was 69% (18/26), and
- false negative rate was 15% (8/40).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 55% (36/66),
- sensitivity was 50% (20/40),
- specificity was 62% (16/26),
- positive predictivity was 67% (20/30),
- negative predictivity was 44% (16/36),
- false positive rate was 38% (10/26), and
- false negative rate was 50% (20/40).

<u>FETAX versus Rabbit:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 44% (15/34),
- sensitivity was 71% (10/14),
- specificity was 25% (5/20),
- positive predictivity was 40% (10/25),
- negative predictivity was 56% (5/9),
- false positive rate was 75% (15/20), and
- false negative rate was 29% (4/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 47% (16/34),
- sensitivity was 29% (4/14),
- specificity was 60% (12/20),
- positive predictivity was 33% (4/12),
- negative predictivity was 55% (12/22),
- false positive rate was 40% (8/20), and
- false negative rate was 71% (10/14).

Using either TI decision criterion value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, and rabbits appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI decision criteria value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0.

6.3 Performance Characteristics of FETAX compared to Human Teratogenicity Study Results

The performance characteristics of FETAX compared to human teratogenicity study results were determined also using three approaches. Performance characteristics were calculated based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. In addition to these analysis conducted using the total FETAX database, the performance characteristics were determined, where feasible, by chemical and product class for FETAX, with and without metabolic activation combined, compared to human teratogenicity study results.

6.3.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, without metabolic activation, compared to human teratogenicity study results were calculated using both single and multiple decision criteria (**Table 14**).

Single Decision Criteria: Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC_{50} <0.3),

- accuracy varied from 48% (15/31) to 63% (17/27 and 19/30),
- sensitivity from 47% (7/15) to 67% (10/15),
- specificity from 31% (5/16) to 80% (12/15),
- positive predictivity from 48% (10/21) to 70% (7/10),
- negative predictivity from 50% (5/10) to 65% (11/17),
- false positive rate from 20% (3/15) to 69% (11/16), and
- false negative rate from 33% (5/15) to 53% (8/15).

Maximal accuracy and specificity occurred when the single decision criterion was an MCIG/LC_{50} ratio less than 0.3. Maximal sensitivity occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 61% (11/18),
- sensitivity was 67% (6/9),
- specificity was 56% (5/9),
- positive predictivity was 60% (6/10),
- negative predictivity was 63% (5/8),
- false positive rate was 44% (4/9), and
- false negative rate was 33% (3/9).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 52% (14/27),
- sensitivity was 75% (9/12),
- specificity was 33% (5/15),
- positive predictivity was 47% (9/19),
- negative predictivity was 63% (5/8),
- false positive rate was 67% (10/15), and
- false negative rate was 25% (3/12).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 63% (17/27),
- sensitivity was 50% (6/12),
- specificity was 73% (11/15),
- positive predictivity was 60% (6/10),
- negative predictivity was 65% (11/17),
- false positive rate was 27% (4/15), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 73% (16/22),
- sensitivity was 60% (6/10),
- specificity was 83% (10/12),
- positive predictivity was 75% (6/8),
- negative predictivity was 71% (10/14),
- false positive rate was 17% (2/12), and
- false negative rate was 40% (4/10).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 65% (17/26),
- sensitivity was 58% (7/12),
- specificity was 71% (10/14),
- positive predictivity was 64% (7/11),
- negative predictivity was 67% (10/15),
- false positive rate was 29% (4/14), and
- false negative rate was 42% (5/12).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 69% (18/26),
- sensitivity was 50% (6/12),
- specificity was 86% (12/14),
- positive predictivity was 75% (6/8),
- negative predictivity was 67% (12/18),
- false positive rate was 14% (2/14), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were excluded or were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were excluded or were re-classified as negative and included in the analysis

The performance characteristics of FETAX, without metabolic activation, compared to human teratogenicity study results were maximal and similar when either a TI value greater than 3.0 or a MCIG/LC$_{50}$ ratio less than 0.3 were used as the single decision criterion. In general, the use of multiple criteria did not increase the performance of FETAX for predicting human teratogenicity.

6.3.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, with metabolic activation, compared to human teratogenicity results were calculated using both single and multiple decision criteria (**Table 15**). The validity of this analysis is questionable considering the very limited number of substances tested with metabolic activation in FETAX for which there were relevant human data also.

<u>Single Decision Criteria:</u> Based on the use of single decision criterion (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 40% (4/10) to 100% (8/8),
- sensitivity from 50% (1/2) to 100% (3/3),
- specificity from 14% (1/7) to 100% (5/5),
- positive predictivity from 33% (3/9) to 100% (3/3),
- negative predictivity from 86% (6/7) to 100% (1/1 and 5/5),
- false positive rate from 0% (0/5) to 86% (6/7), and
- false negative rate from 0% (0/3) to 50% (1/2).

Maximal accuracy, sensitivity, and specificity occurred when the single decision criterion was an MCIG/LC_{50} ratio less than 0.3.

<u>Multiple Decision Criteria:</u> Using the multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 100% (4/4),
- sensitivity was 100% (3/3),
- specificity was 100% (1/1),
- positive predictivity was 100% (3/3),
- negative predictivity was 100% (1/1),
- false positive rate was 0% (0/1), and
- false negative rate was 0% (0/3).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 50% (4/8),
- sensitivity was 100% (3/3),
- specificity was 20% (1/5),
- positive predictivity was 43% (3/7),
- negative predictivity was 100% (1/1),
- false positive rate was 80% (4/5), and
- false negative rate was 0% (0/3).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 100% (8/8),
- sensitivity was 100% (3/3),
- specificity was 100% (5/5),
- positive predictivity was 100% (3/3),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/3).

Maximal performance characteristics occurred when equivocal results were excluded from analysis or were re-classified as negative results and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 100% (6/6),
- sensitivity was 100% (1/1),
- specificity was 100% (5/5),
- positive predictivity was 100% (1/1),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/1).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 100% (7/7),
- sensitivity was 100% (2/2),
- specificity was 100% (5/5),
- positive predictivity was 100% (2/2),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/2).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 86% (6/7),
- sensitivity was 50% (1/2),
- specificity was 100% (5/5),
- positive predictivity was 100% (1/1),
- negative predictivity was 83% (5/6),
- false positive rate was 0% (0/5), and
- false negative rate was 50% (1/2).

Maximal performance characteristics occurred when equivocal results were excluded from analysis or were re-classified as negative results and included in the analysis.

The performance characteristics of FETAX, with metabolic activation, compared to human teratogenicity study results were maximal and similar when an MCIG/LC_{50} ratio less than 0.3 were used as the single decision criterion. In general, the use of multiple criteria did not increase the performance of FETAX for predicting human teratogenicity.

6.3.3 Performance of FETAX, With and Without Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, with and without metabolic activation, compared to human teratogenicity results were calculated using both single and multiple decision criteria (**Table 16**).

Single Decision Criteria: Based on the use of single decision criterion (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 48% (15/31) to 70% (19/27),
- sensitivity from 47% (7/15) to 80% (12/15),
- specificity from 19% (3/16) to 81% (13/16),
- positive predictivity from 48% (12/25) to 70% (7/10),
- negative predictivity from 50% (3/6) to 73% (11/15),
- false positive rate from 19% (3/16) to 81% (13/16), and
- false negative rate from 20% (3/15) to 53% (8/15).

Maximal accuracy and specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio less than 0.3. Maximal sensitivity occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 69% (11/16),
- sensitivity was 89% (8/9),
- specificity was 43% (3/7),
- positive predictivity was 67% (8/12),
- negative predictivity was 75% (3/4),
- false positive rate was 57% (4/7), and
- false negative rate was11% (1/9).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 52% (14/27),
- sensitivity was 92% (11/12),
- specificity was 20% (3/15),
- positive predictivity was 48% (11/23),
- negative predictivity was 75% (3/4),
- false positive rate was 80% (12/15), and
- false negative rate was 8% (1/12).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 67% (8/13),
- specificity was 73% (11/15),
- positive predictivity was 67% (8/12),
- negative predictivity was 73% (11/15),
- false positive rate was 27% (4/15), and
- false negative rate was 33% (4/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were excluded from analysis or were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 76% (16/21),
- sensitivity was 67% (6/9),
- specificity was 83% (10/12),
- positive predictivity was 75% (6/8),
- negative predictivity was 77% (10/13),
- false positive rate was 17% (2/12), and
- false negative rate was 33% (3/9).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 75% (9/12),
- specificity was 67% (10/15),
- positive predictivity was 64% (9/14),
- negative predictivity was 77% (10/13),
- false positive rate was 33% (5/15), and
- false negative rate was 25% (3/12).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 50% (6/12),
- specificity was 87% (13/15),
- positive predictivity was 75% (6/8),
- negative predictivity was 68% (13/19),
- false positive rate was 13% (2/15), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis. Maximal sensitivity occurred when equivocal results were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis.

In general, among single decision criteria, the use of a criterion based on an MCIG/LC_{50} ratio less than 0.3 resulted in the greatest accuracy and specificity, while a TI value greater than 1.5 resulted in the greatest sensitivity for identifying human teratogenicity responses. The use of multiple decision criteria did not have an appreciable effect on the performance characteristics of FETAX.

6.3.4 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Human Teratogenicity Study Results by Chemical and Product Class

The most numerically prevalent chemical classes were alcohols (including glycols); amides; amines; halogenated organic compounds; esters; heavy metals and their salts; hydrazides and hydrazines; nitrogen heterocyclic compounds; organic (phenolic and carboxylic) acids; and salts (see **Section 3.3**). The most common product classes tested in FETAX were antimicrobials, chemical synthesis, cosmetics, dyes, food additives, fossil fuels, pesticides, pharmaceuticals, photographic chemicals, and polymers (including monomers). The performance characteristics of FETAX, with and without metabolic activation, compared to human teratogenicity study results were determined by chemical and product class using single decision criteria (i.e., TI >1.5, TI >3.0, MCIG/LC_{50} <0.3) only (**Table 17**). Analyses were limited to those chemical and product classes that included a minimum of 15 substances tested in FETAX for which there was also human teratogenicity study results. For comparative purposes, the corresponding performance characteristics when all FETAX data were considered are included in **Table 17**.

Nitrogen Heterocyclic Compounds: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 56% (9/16),
- sensitivity was 78% (7/9),
- specificity was 29% (2/7),
- positive predictivity was 58% (7/12),
- negative predictivity was 50% (2/4),
- false positive rate was 71% (5/7), and
- false negative rate was 22% (2/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 69% (11/16),
- sensitivity was 56% (5/9),
- specificity was 86% (6/7),
- positive predictivity was 83% (5/6),
- negative predictivity was 60% (6/10),
- false positive rate was 14% (1/7), and
- false negative rate was 44% (4/9).

Due to the absence of a sufficient database, performance characteristics using a decision criterion based on an MCIG/LC_{50} ratio of less than 0.3 were not determined.

Pharmaceuticals: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 43% (9/21),
- sensitivity was 80% (8/10),
- specificity was 9% (1/11),
- positive predictivity was 44% (8/18),
- negative predictivity was 33% (1/3),
- false positive rate was 91% (10/11), and
- false negative rate was 20% (2/10).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 67% (14/21),
- sensitivity was 50% (5/10),
- specificity was 82% (9/11),
- positive predictivity was 71% (5/7),
- negative predictivity was 64% (9/14),
- false positive rate was 18% (2/11), and
- false negative rate was 50% (5/10).

Using an MCIG/LC$_{50}$ ratio less than 0.3 as the single decision criterion,

- accuracy was 68% (13/19),
- sensitivity was 67% (6/9),
- specificity was 70% (7/10),
- positive predictivity was 67% (6/9),
- negative predictivity was 70% (7/10),
- false positive rate was 30% (3/10), and
- false negative rate was 33% (3/9).

Due to the limited FETAX database with corresponding human teratogenicity study results, only one chemical class and one product class were evaluated for performance characteristics compared to human teratogenicity study results. The performance characteristics of FETAX compared to human teratogenicity study results were not improved for these chemical and product classes compared to that for the total database.

6.4 Performance Characteristics of Rat, Mouse, and/or Rabbit Teratogenicity Test Results compared to Human Teratogenicity Study Results

The performance characteristics for combined rat, mouse, and rabbit teratogenicity results, as well as for each individual species, compared to human teratogenicity responses were calculated (for comparative purposes compared to FETAX with and/or without metabolic activation, these data are presented in **Tables 14** through **16**).

For combined laboratory mammal results,

- accuracy was 63% (19/30),
- sensitivity was 71% (10/14),
- specificity was 56% (9/16),
- positive predictivity was 59% (10/17),
- negative predictivity was 69% (9/13),
- false positive rate was 44% (7/16), and
- false negative rate was 29% (4/14).

When the performance characteristics for rat compared to human teratogenicity results only were determined,

- accuracy was 65% (17/26),
- sensitivity was 75% (9/12),
- specificity was 57% (8/14),
- positive predictivity was 60% (9/15),
- negative predictivity was 73% (8/11),
- false positive rate was 43% (6/14), and
- false negative rate was 25% (3/12).

When the performance characteristics for mouse compared to human teratogenicity results only were calculated,

- accuracy was 68% (19/28),
- sensitivity was 71% (10/14),
- specificity was 64% (9/14),
- positive predictivity was 67% (10/15),
- negative predictivity was 69% (9/13),
- false positive rate was 36% (5/14), and
- false negative rate was 29% (4/14)

When the performance characteristics for rabbit compared to human teratogenicity results only were calculated,

- accuracy was 53% (8/15),
- sensitivity was 50% (4/8),
- specificity was 57% (4/7),
- positive predictivity was 57% (4/7),
- negative predictivity was 50% (4/8),
- false positive rate was 43% (3/7), and
- false negative rate was 50% (4/8).

Maximal performance were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally reduced compared to that for the other two species, but may reflect the limited database available for substances also tested in FETAX. The rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity study results appeared to be not much improved compared to that calculated for FETAX, with and without metabolic activation, using the MCIG/LC_{50} ratio of less than 0.3 as the single decision criterion.

6.5 FETAX Results Discordant with Reference Laboratory Mammal or Human Teratogenicity Study Results

The substances tested in FETAX that are discordant with the teratogenicity results obtained for laboratory mammals and humans are listed in **Table 18**. For the purpose of collecting these data, a substance was classified as positive in FETAX based on the most commonly used decision criterion (i.e., TI >1.5) only. Furthermore, if tested with and without metabolic activation, a substance was classified as a FETAX positive if a positive response was obtained using either exposure condition, and as a FETAX negative only if negative results were obtained with and without metabolic activation. Classification of a laboratory mammal teratogenicity result as positive was based on the presence of at least one positive rat, mouse, and/or rabbit study.

Using these classification parameters:

- Twenty-four substances were discordant with laboratory mammal teratogenicity results (seven substances were FETAX positive and laboratory mammal negative; seventeen substances were FETAX negative and laboratory mammal positive);

- Eight substances were concordant with laboratory mammal teratogenicity data but discordant with human teratogenicity results (one substance was FETAX/laboratory mammal negative and human positive; seven substances were FETAX/laboratory mammal positive and human negative); and

- Eight substances were discordant with laboratory mammal and human teratogenicity results (two substances were FETAX negative and laboratory mammal/human positive; six substances were FETAX positive and laboratory mammal/human negative);

- three substances were discordant with laboratory mammal but concordant with human teratogenicity results (no substance was a FETAX/human negative and laboratory mammal positive; three substances were FETAX/human positive and laboratory mammal negative).

Table 18. FETAX Results Discordant with Reference Laboratory Mammal Data and/or Human Teratogenicity Results*

Substance	TI>1.5	Laboratory Mammal	Human
Substances Discordant with Laboratory Mammal Teratogenicity Results			
alpha.-Chaconine	-	+	
Actinomycin D	-	+	
Cycloheximide	-	+	
Dichloroacetic acid	-	+	
Formamide	-	+	
Glycerol formal	-	+	
N-Nitrosodimethylamine	-	+	
2-Butyne-1,4-diol	+	-	
Acrylamide	+	-	
Amaranth	+	-	
Atrazine	+	-	
Benzo[a]pyrene	+	-	
Cobalt chloride	+	-	
Copper chloride	+	-	
Cotinine	+	-	
Diethylene glycol	+	-	
Glycerol	+	-	
Hydrazine	+	-	
Monosodium glutamate	+	-	
Permethrin	+	-	
Propylene glycol	+	-	
Sodium acetate	+	-	
Sodium selenate	+	-	
Trichloroethylene	+	-	
Substances Concordant with Laboratory Mammal but Discordant with Human Teratogenicity Results			
p-Hydroxydilantin	-	-	+
Boric Acid	+	+	-
Cadmium chloride	+	+	-
Caffeine	+	+	-
Dichloroacetate	+	+	-
Phenytoin	+	+	-
Theophylline	+	+	-
Trichloroacetic acid	+	+	-

Substances Discordant with Laboratory Mammal and Human Teratogenicity Results			
Ethanol (L)	-	+	+
m-Hydroxydilantin	-	+	+
Acetaminophen	+	-	-
Acetone	+	-	-
Ascorbic acid	+	-	-
Diphenhydramine hydrochloride	+	-	-
Doxylamine succinate	+	-	-
Furazolidone	+	-	-
Substances Discordant with Laboratory Mammal but Concordant with Human Teratogenicity Results			
4-Hydroxycoumarin	+	-	+
Coumarin	+	-	+
Isoniazid	+	-	+

*If tested with and without metabolic activation, a substance was classified as a FETAX positive if a positive response was obtained using either exposure condition, and as a FETAX negative only if negative results were obtained with and without metabolic activation. Classification of a laboratory mammal teratogenicity result as positive was based on the presence of at least one positive rat, mouse, and/or rabbit study.
The symbols "-" and "+" signify a negative and positive response, respectively.

The bases for the discordant results (e.g., mechanistic, the use of a less than optimal decision criteria) between FETAX and the combined laboratory mammal and/or the human teratogenicity results remains to be determined.

6.6 NICEATM Analysis of FETAX Decision Criteria

The use of a single decision criterion based on a TI value greater than 1.5 appeared to provide the optimal approach in terms of accuracy and sensitivity for predicting combined laboratory mammal teratogenicity data. The use of a TI value greater than 3.0 as the single decision criterion resulted in increased specificity, but decreased sensitivity. The use of multiple decision criteria had no appreciable effect on accuracy or sensitivity but increased specificity when equivocal results were excluded from the analysis. Using either TI decision criterion value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, or rabbits individually appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a decision criterion based on a TI value greater than 1.5 rather than 3.0.

In general, the use of a single decision criterion based on an MCIG/LC_{50} ratio less than 0.3 appeared to provide the optimal approach for predicting human teratogenicity data. The use of multiple decision criterion increased sensitivity when equivocal results were classified as positive, and increased specificity when equivocal results were classified as negative.

Maximal performance characteristics for laboratory mammal data compared to human results were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally poor compared to that for the other two species. In general, the rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity results appeared to be similar to that calculated for FETAX using the MCIG/LC_{50} ratio less than 0.3 as the single decision criterion. However, the database for this comparison was limited to substances tested in FETAX only.

Limiting the analysis of the performance characteristics to substances for which there were, in each case, FETAX, laboratory mammal, and human results does not alter these conclusions.

6.6.1 Evaluation for the Optimal FETAX Single Decision Criterion

In an attempt to identify the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data, NICEATM assessed the relationship between different TI values or MCIG/LC_{50} ratios and performance characteristics. Accuracy, sensitivity, and specificity were calculated for FETAX, without metabolic activation, compared to combined laboratory mammal (rat, mouse, and rabbit) or human teratogenicity results. In conducting these analysis, the median TI value or median MCIG/LC_{50} ratio was used for test substances where multiple studies had been conducted. The use of a median value may result in performance characteristics for

FETAX that are different from those calculated in **Sections 6.1** through **6.3**. FETAX performance characteristics in those sections were based on a weight-of-evidence approach that only evaluated whether a TI value or an MCIG/LC_{50} ratio was above or below the selected decision point.

The optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion for identifying teratogens in FETAX depends on whether the assay is to be used as a replacement for an existing *in vivo* laboratory mammal assay, or as a screen to identify substances expected to be positive in laboratory mammal assays or in humans. If used as a replacement assay, accuracy (i.e., the ability to correctly identify both positive and negative teratogens) is probably the most important performance characteristic on which to evaluate the data. In contrast, for screening purposes, sensitivity (i.e., the proportion of all positive substances that are correctly identified as positive; sensitivity is also the inverse of the false negative rate) may be the performance characteristic of primary interest.

6.6.1.1 Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

Optimal TI Value: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using TI values ranging from 0 to 49 as the single decision criterion, compared to combined rat, mouse, and rabbit teratogenicity results are presented graphically in **Figure 1**.

Maximal accuracy for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results was ~60% at TI values between 0 and ~2.1. At TI values between 2.1 and ~22, accuracy steady decreased to ~40% and then remained relatively constant at this value as the TI increased. Sensitivity was 85% at a TI value of 1.42; the corresponding specificity was 40%.

Optimal MCIG/LC_{50} Ratio: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using MCIG/LC_{50} ratios ranging from 0 to 1.5 as the single decision criterion, compared to combined rat, mouse, and rabbit teratogenicity test results are presented graphically in **Figure 2**.

Maximal accuracy for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results was ~58% at MCIG/LC_{50} ratios between 0 and 0.2. At MCIG/LC_{50} ratios between 0.2 and 0.4, accuracy steadily decreased to ~40% and then remained relatively constant at MCIG/LC_{50} ratios up to 1.5. Sensitivity was 85% at an MCIG/LC_{50} ratio of 0.08; the corresponding specificity was 13%.

When compared to combined rat, mouse, and rabbit teratogenicity results, accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%. This level of accuracy does not support the use of FETAX, as currently conducted, as a possible replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 40%. This low specificity corresponds to a false positive rate of greater than 60%. The poor specificity at a sensitivity of 85% raises concerns about the use of FETAX as a screening assay.

6.6.1.2 Human Teratogenicity Study Results

Optimal TI Value: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using TI values ranging from 0 to 49 as the single decision criterion, compared to human teratogenicity study results are presented graphically in **Figure 3**.

Maximal accuracy for FETAX, without metabolic activation, compared to human teratogenicity study results was ~60% at TI values around 3.0. Accuracy then decreased to ~50% at higher TI values. Sensitivity was 85% at a TI value of 1.0; the corresponding specificity was 8%.

Optimal MCIG/LC_{50} Ratio: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using MCIG/LC_{50} ratios ranging from 0 to 1.5 as the single decision criterion, compared to human teratogenicity study results are presented graphically in **Figure 4**.

Maximal accuracy for FETAX, without metabolic activation, compared to human teratogenicity study results was ~50% at MCIG/LC_{50} ratios between 0 and 0.06 or between 1.2 and 1.5. Sensitivity was 85% at an MCIG/LC_{50} ratio between 0.06 and 0.07; the corresponding specificity was 8%.

When compared to human teratogenicity results, maximum accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion was never greater than about 50%. This value is lower than the previously reported accuracy of 64% calculated using an MCIG/LC_{50} ratio of less than 0.3 as the decision criterion for FETAX, without metabolic activation, compared to human teratogenicity study results (**Table 14**). This difference presumably reflects the use of median values in this analysis versus the weight-of-evidence approach used to generate the data for **Table 14**. This level of accuracy does not support the use of FETAX, as currently conducted, as apotential replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 10%. This low specificity corresponds to a false positive rate of greater than 90%. This poor specificity at a sensitivity of 85% raises concerns about the use of FETAX as a screening assay.

6.6.2 Characteristic Malformations Induced in *X. laevis* Embryos

Qualitative information on the types of malformations was reported for 35 substances (**Appendices 2** and **3)**. Three of these were environmental samples, while the remaining 32 were individual substances. Malformations reported most commonly (i.e., reported for at least ten substances) included gut miscoiling, craniofacial malformations, and microencephaly. Substances inducing such malformations are provided in **Table 19**.

Table 19. Substances Inducing Gut Miscoiling, Craniofacial Malformations, or Microencephaly in *X. laevis* Embryos

Substance	FETAX Malformation(s) Induced
5-Azacytidine	Gut miscoiling; craniofacial malformations; microencephaly
5-Fluorouracil	Gut miscoiling; microencephaly
Amaranth	Gut miscoiling; craniofacial malformations
Bisphenol A	Craniofacial malformations
Copper (1)	Gut miscoiling; craniofacial malformations; microencephaly
Copper (2)	Gut miscoiling; craniofacial malformations; microencephaly
Copper sulfate	Gut miscoiling; craniofacial malformations; microencephaly
Desisopropyl atrazine	Microencephaly
Diethylene glycol	Gut miscoiling
Glycerol	Gut miscoiling; craniofacial malformations
Hydroxyurea	Microencephaly
Maneb	Craniofacial malformations
Methotrexate	Gut miscoiling; microencephaly
Nickel chloride	Gut miscoiling; craniofacial malformations
Pentachlorophenol	Gut miscoiling; craniofacial malformations; microencephaly
Permethrin	Microencephaly
Phthalic acid	Gut miscoiling
Propylthiourea	Craniofacial malformations
Pseudoephedrine	Gut miscoiling; craniofacial malformations

Sodium arsenite	Gut miscoiling; craniofacial malformations
Sodium iodoacetate	Gut miscoiling
Zinc (1)	Gut miscoiling; craniofacial malformations; microencephaly
Zinc (2)	Gut miscoiling; craniofacial malformations; microencephaly
Zinc sulfate heptahydrate	Gut miscoiling; craniofacial malformations; microencephaly

Other malformations reported less frequently are as follows, in decreasing order of occurrence:

- microopthalmia,
- opthalmic malformations,
- pericardial edema,
- mouth deformities,
- visceral edema,
- muscular kinking,
- facial abnormalities,
- gut malformations,
- edema,
- skeletal kinking,
- blistering of the dorsal fin,
- eye malformations,
- head anomalies,
- abnormal heart coiling,
- bent tail,
- curved tail tip,
- notocord defects,
- brain abnormalities,
- improper skin pigmentation,
- visceral hemorrhage,
- anencephaly,
- dermal blisters,
- incomplete gut coiling,
- hunchback,
- hydrocephaly,
- rupture of the eye pigment vesicle,
- opthalmic edema,
- axial skeletal anomalies,
- failure of the choriod to fuse,
- hypopigmented eyes,
- fin expansion,
- malformed fins,
- heart anomalies,
- enlarged heart, and
- vertebral fusions.

In the FETAX Phase III.3 Validation Study, Bantle et al. (1999) evaluated study results based on both single and multiple decision criteria. Using multiple decision criteria, test substances were classified as equivocal when either a TI value greater than 1.5 or an MCIG/LC_{50} ratio less than 0.30 was obtained. In such situations, the types and severity of malformations in *X. laevis* embryos were examined for guidance in assessing teratogenic hazard. However, due to the subjectivity of malformation identification, a decision was made that this approach should not be made a permanent part of the decision criteria by the investigators.

Dr. D. Fort (personal communication) has recently re-evaluated the FETAX Phase III.3 Validation Study results based on limiting the analysis of the EC_{50} to malformations deemed characteristic for the substance tested, rather than using data on all malformations as described in the ASTM FETAX Guideline (1991, 1998). Using the preserved embryos, Dr. D. Fort (personal communication) has recently re-evaluated the types and incidences of malformations in the various studies conducted in the FETAX Phase III.3 Validation Study. Subsequently, Dr. Fort then limited the analysis of the EC_{50} to malformations deemed characteristic for the substance tested, rather than using data on all malformations as described in the ASTM FETAX Guideline (1991, 1998). The embryos were re-evaluated to ensure the use of a uniform criteria in identifying malformations. The premise behind the use of characteristic malformations to evaluate the potential teratogenic hazard of a test substance is that any given teratogenic agent induces a syndrome characteristic of that substance. Non-specific, or background, malformations are also found in any given study. Malformations that are characteristic of the test substance should increase in frequency and possibly severity with increasing concentrations of the test substance. Malformations that occur sporadically and do not increase in frequency or severity with respect to test substance concentration are not likely directly due to the test material itself. To evaluate FETAX studies using this criterion, both characteristic and non-characteristic malformations are determined. However, statistical evaluation of the malformation data is limited to characteristic malformation data only. Because an evaluation of malformations is subjective, a secondary review of the scoring process is recommended (D. Fort, personal communication).

A preliminary assessment of the results of the re-analysis indicated that the use of the characteristic malformation criterion resulted in decreased intra- and inter-laboratory variability, a decreased number of equivocal test calls, and increased endpoint precision. Further, since this approach considers the syndrome associated with exposure to a given substance, it provides a more accurate means of comparing results between species. This approach may or may not increase the predictive accuracy of FETAX since that depends on the responsiveness of *Xenopus* to the test material. The disadvantages include, increased time required to evaluate each test, greater knowledge required by the technical staff, and a rigid QA/QC program to enforce secondary data review. However, in this re-analysis, all data on characteristic malformations were collected by the same scorer, which would be inherently expected to reduce inter-laboratory variability. NICEATM suggests that this approach has merit and that the process by which characteristic malformations is recognized *posthoc* needs to be evaluated across multiple laboratories.

Another aspect of characteristic malformations in FETAX that has yet to be critically explored is the correlation between the types of agent-specific malformations induced in *X. laevis* and those induced by the same agent in rats, mice, and rabbits, or in humans. A very limited assessment by Sabourin and Faulk (1987) and one more recently by Fort et al. (2000a) suggested a positive correlation between the types of malformations induced in laboratory mammals and in *Xenopus* embryos. A more extensive evaluation of the correlation between the types of malformations induced in laboratory mammals and in *Xenopus* embryos is currently in progress by NTP using data collected in the FETAX Phase III.3 Validation Study. The results of this assessment may support the validity of additional research in this area.

6.6.3 Evaluation of Growth Inhibition

In FETAX, the ratio between the MCIG and the LC_{50} is used as one criterion for identifying teratogens. The MCIG is the minimal concentration to inhibit growth, as determined by comparing the mean head-to-tail length at each test concentration compared to the appropriate control value, using student's t-test. However, because an assessment of growth is not required for range-finding tests (ASTM, 1991; 1998), the test concentrations selected for the definitive tests are frequently not conducive to an adequate assessment of the MCIG. As a consequence, the MCIG has been associated with the greatest inter-laboratory variability (see **Section 7**). Dr. D. Fort (personal communication) has suggested that a point estimate for growth inhibition, rather than the MCIG, would enhance the performance characteristics of FETAX. The possible effect of this modification to the decision criteria for FETAX on performance and the possible protocol changes needed for implementation have not yet been determined.

6.6.4 The Use of Confidence Intervals

Dr. D. Fort (personal communication) has suggested that the FETAX performance characteristics would be increased if 95% confidence intervals were used for statistically identifying TI values (and other point estimates) that are significantly greater than the decision point. This approach would allow for the variability among the replicate definitive tests to be considered when identifying a positive response in FETAX. The utility of this approach has yet to be evaluated.

6.6.5 Performance of FETAX with Metabolic Activation

In the FETAX Phase III.2 Validation Study, caffeine and CP were evaluated for their teratogenic activity in both the absence and presence of an exogenous MAS. This validation study was conducted because the investigators recognized the importance of including the capacity for metabolic activation. Based on the results of this validation study, the investigators concluded that the inclusion of metabolic activation in the assay was essential if FETAX was to be used to predict developmental hazard in mammals (including humans) but that the methodology required further development. The FETAX Phase III.3 Validation Study extended the Phase III.2 Validation Study results by testing 12 substances (acrylamide, boric acid, dichloroacetate, diethylene glycol, ethylene glycol, glycerol, phthalic acid, sodium arsenite, sodium bromate, sodium iodoacetate, tribromoacetic acid, and triethylene glycol dimethylether), with and without metabolic activation, in three laboratories with extensive FETAX experience. The rationale for the selection of the test substances was not provided in the validation report. However, it is likely that selection was based on the availability of relevant laboratory mammal data and the suitability of the test substance for testing in FETAX (e.g., water solubility, lack of volatility). It does not appear that selection was based on the known or suspected requirement for metabolic activation to be a teratogen. NICEATM evaluated the possible metabolic activation requiring status of all substances tested in FETAX with an MAS. Identification of the possible involvement of metabolic activation was based on whether the substance was positive in one or more *in vitro* genetic toxicological tests (generally the *Salmonella typhimurium* reverse mutation assay) in the presence of metabolic activation only. *In vitro* genetic toxicology data were obtained from the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database. This method for identifying substances that may require metabolic activation to be teratogenic *in vitro* assumes a common mechanism between mutagenicity and teratogenicity that may not be valid. The results of this determination are presented in **Table 20**, with substances ranked by the increasing ratio of the TI with metabolic activation to the TI without metabolic activation. Also provided in **Table 20** is the FETAX result for studies conducted with and without metabolic activation, based on the single decision criterion of a TI greater than 1.5.

Of the 35 substances tested with metabolic activation in FETAX, useful *in vitro* genetic toxicology data were located on 15 substances (43%). Of these 15 substances, 11 were genotoxic in the absence of metabolic activation and four were only genotoxic with metabolic activation. In FETAX, with and without metabolic activation, three of the 35 substances were classified as negative under both metabolic conditions, seven were positive with metabolic activation only, three were positive without metabolic activation only, and 22 were positive under both metabolic conditions. Of the four substances requiring metabolic activation to be genotoxic *in vitro*, two substances were positive in FETAX with metabolic activation only while the other two substances were active in FETAX with and without metabolic activation. Of the eleven substances that are genotoxic *in vitro* without metabolic activation, two substances were positive in FETAX with metabolic activation only, two were positive in FETAX without metabolic activation only, and the remaining seven substances were positive in FETAX with and without metabolic activation.

The information in **Table 20** was also evaluated based on the assumption that a ratio of the median TI with metabolic activation to the median TI without metabolic activation of

Table 20. Substances Tested in FETAX With Metabolic Activation: Identification of Possible Metabolic Activation Requiring Substances

Substance	Requires MA*	Result Without MA	Result With MA	TI With MA/ TI Without MA
Doxylamine succinate	No	+	+	0.01
Nicotine		+	+	0.01
Hydrazine	No	+	+	0.03
Acetylhydrazide		+	+	0.06
4-Bromobenzene		+	-	0.13
Cytochalasin D	No	+	-	0.38
Sodium iodoacetate		-	-	0.42
Theophylline	No	+	-+	0.46
Caffeine	No	+	+	0.68
Isoniazid		+	+	0.70
Sodium bromate		+	+	0.72
Solanine		-	-	0.76
Triethylene glycol dimethyl ether		+	+	0.78
Phenytoin		+	+	0.80
Isonicotinic acid		+	+	0.84
N-Ethyl-N-nitrosourea	No	+	+	0.88
Boric Acid		+	+	0.89
Tribromoacetic acid		+	+	0.92
7-Hydroxycoumarin		+	+	1.00
Ethylene glycol	No	+	+	1.00

Table 20. Substances Tested in FETAX With Metabolic Activation: Identification of Possible Metabolic Activation Requiring Substances (Continued)

Substance	Requires MA*	Result Without MA	Result With MA	TI With MA/ TI Without MA
Acrylamide	No	+	+	1.01
Phthalic acid		-	-	1.04
3-Methylxanthine		+	+	1.07
Diethylene glycol		+	+	1.08
Dichloroacetate	No	-	+	1.11
1-Methylxanthine		+	+	1.12
Glycerol		-	+	1.18
Sodium arsenite	No	-	+	1.28
2-Acetylaminofluorene	Yes	+	+	1.34
4-Hydroxycoumarin		-	+	1.67
CP	Yes	-	+	1.85
Acetaminophen		-	+	1.92
Trichloroethylene	No	+	+	2.10
Urethane	Yes	+	+	3.91
Benzo[a]pyrene	Yes	-	+	6.83

The terms “No” and “Yes” indicates chemicals that do not or do appear to require metabolic activation, respectively, to induce a positive response in an *in vitro* genetic toxicological test according to the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database. MA = metabolic activation.

*Indicates substances without relevant metabolic activation-requiring information in these two databases.

[1] Classification of the test substance in FETAX based on a weight-of-evidence approach where multiple studies had been conducted, using a TI >1.5 as the single decision criterion.

[2] Ratio of median TI value with metabolic activation to median TI value without metabolic activation.

approximately one indicates independence of metabolism, while a ratio below 0.5 indicates decreased activity with metabolic activation and a ratio above 1.5 indicates increased activity with metabolic activation. Eight of the 35 substances tested with metabolic activation exhibited a with metabolic activation/without metabolic activation TI ratio below 0.5. One of these eight substances was negative in FETAX with and without metabolic activation, three were positive in FETAX without metabolic activation only, and four were positive in FETAX with and without metabolic activation. Six of the 35 substances tested with metabolic activation exhibited a with metabolic activation/without metabolic activation TI ratio greater than 1.5. Four of these six substances were positive in FETAX with metabolic activation only, and two were positive in FETAX with and without metabolic activation.

This evaluation revealed that most of the 35 substances tested with metabolic activation were not known to require metabolic activation to be active *in vitro*, but that there was a tendency towards increased activity in FETAX with metabolic activation for those substances that required metabolic activation to be genotoxic *in vitro*. Based on the limited database, additional studies to validate the role of metabolic activation in FETAX appear to be justified.

6.7 Strengths and Limitations of FETAX in Terms of Performance Characteristics

FETAX is a 96-hour *in vitro* whole-embryo test developed to determine the teratogenic and developmental toxicity potential of chemicals, metals, and complex mixtures (ASTM, 1991; 1998; Finch, 1994). It is essentially an organogenesis test, and organogenesis is highly conserved across amphibians and laboratory mammals. The first 96 hours of embryonic development in *Xenopus* parallel many of the major processes of human organogenesis (ASTM, 1991; 1998). Thus, it was anticipated that FETAX should be useful in predicting potential human developmental toxicants and teratogens (ASTM, 1991; 1998). Due to the nature of the endpoints assessed, FETAX does not provide information on substances that may induce functional developmental deficits in mammals. Because FETAX has been concluded by the developers to be easy, rapid, reliable, and inexpensive, the test (with and without metabolic activation) has been proposed as a screening assay for potential human teratogens and developmental toxicants (ASTM, 1991; 1998). As a screening test, a positive FETAX response would indicate a potential human hazard while a negative FETAX response would not indicate the absence of a hazard. In the role of a screening assay, a negative response would be followed by *in vivo* laboratory mammal testing, while a positive response would require no further testing unless the investigator is concerned about a potential false positive response.

NICEATM evaluated the performance characteristics of FETAX, with and/or without metabolic activation, compared to teratogenicity test results in rats, mice, and/or rabbits, and compared to human teratogenicity study results. In this analysis, different decision criteria (i.e., single decision criteria based on a TI value greater than 1.5 or 3.0, or an MCIG/LC_{50} ratio less than 0.30; multiple decision criteria based on a TI value greater than 1.5 or 3.0 plus an MCIG/LC_{50} ratio less than 0.30) reported in the literature for identifying teratogenic potential in FETAX were evaluated. When the performance for FETAX, with and without metabolic activation, was determined compared to combined rat, mouse, and rabbit teratogenicity results, maximal accuracy was 60%, maximal sensitivity was 80%, and maximal specificity was 56%. These values occurred using different decision criteria. When the performance for FETAX, with and without metabolic activation, was determined compared to human teratogenicity study results, maximal accuracy was 73%, maximal sensitivity was 93%, and maximal specificity was 79%. Again, each maximal value occurred using different decision criteria.

NICEATM also evaluated the performance characteristics of FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results and human teratogenicity study results by chemical and product class using single decision criteria (i.e., TI >1.5, TIMCIG/LC_{50} <0.3). Analyses were limited to chemical and product classes containing a minimum of 15 FETAX test substances with corresponding animal or human teratogenicity data. Only five chemical classes and one product class were evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results, while only one chemical class and one product class were evaluated for performance characteristics compared to human teratogenicity study results. The accuracy of FETAX compared to laboratory mammal teratogenicity test results was somewhat improved compared to that for the total database for amides, nitrogen heterocyclic compounds, and organic (phenolic and carboxylic) acids. Performance for the other chemical classes and the single product class evaluated were not different from the performance of FETAX compared to the total database. The performance characteristics of FETAX compared to human teratogenicity study results were not improved for nitrogen heterocyclic compounds and pharmaceuticals compared to that for the total database.

In response to these results, NICEATM attempted to identify the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data. Performance characteristics (accuracy, sensitivity, specificity) were detemined for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results or compared to human teratogenicity study results. When compared to laboratory mammal or human data, maximum accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%. This level of accuracy does not support the use of FETAX, as currently conducted, as a replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 30%. The poor specificity at a sensitivity of 85% raises concems about the use of FETAX as a screening assay.

Based on these analyses, additional efforts to optimize the decision criteria appear to be warranted. Several modifications that are potentially useful (e.g., use of characteristic malformations, use of confidence intervals) were discussed in **Section 6.6**.

6.8 Data Interpretation Issues

As specified by the ASTM FETAX Guideline (1991, 1998), three separate decision criteria (TI>1.5; MCIG/LC_{50}<0.3, and severity of malformation) have been used to identify potential human teratogens. The ASTM FETAX Guideline (1991, 1998) concludes that any single decision criterion is sufficient to identify a potential teratogenic hazard, and that these three decision criteria are based on empirical evidence resulting from over 100 materials tested (without metabolic activation) in FETAX. In the NICEATM analysis of the performance characteristics of FETAX compared to either laboratory mammal or human teratogenicity results, these as well as multiple decision criteria were considered. The multiple decision criteria (TI >1.5 or TI >3.0 plus MCIG/LC_{50} <0.3) evaluated were those used in the most recent FETAX Validation Study (Bantle et al., 1999) and in a comparative FETAX-rat study conducted by Fort et al. (2000a). This analysis indicates that the use of a TI value greater than 1.5 and an MCIG/LC_{50} ratio below 0.3 as the single decision criteria results in the maximum accuracy for laboratory mammal and human teratogenicity results, respectively. The use of multiple decision criteria did not significantly increase the ability of FETAX to correctly identify mammalian (including human) teratogens. These analyses suggest that additional effort is warranted to investigate and optimize the methods by which FETAX data are collected and interpreted.

6.9 Section 6 Conclusions

The use of single decision criterion based on a TI value greater than 1.5 appeared to provide the most optimal approach in terms of accuracy and sensitivity for predicting combined laboratory mammal teratogenicity data. The use of multiple decision criteria had no appreciable effect on accuracy or sensitivity but increased specificity when equivocal results were excluded from the analysis.

Using either TI decision criteria value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, or rabbits appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI value, increased accuracy and sensitivity but decreased specificity was associated with the use of a TI value greater than 1.5 rather than 3.0.

In general, the use of single decision criterion based on an MCIG/LC_{50} ratio lower than 0.3 appeared to provide the most optimal approach for predicting human teratogenicity data. The use of multiple decision criteria increased sensitivity when equivocal results were classified as positive and specificity when equivocal results were classified as negative.

Five chemical classes and one product class was evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results. Among the chemical and product classes evaluated, a decision criterion based on a TI value greater than 1.5 generally provided greater accuracy and sensitivity, but less specificity, than one based on either on a TI value greater than 3.0 or on an MCIG/LC_{50} ratio of less than 0.3. The accuracy of FETAX compared to laboratory mammal teratogenicity test results for nitrogen heterocyclic compounds, and organic (phenolic and carboxylic) acids was somewhat improved compared to that for the total database. Performance compared to the other chemical classes and the single product class (pharmaceuticals) evaluated were not different from the performance of FETAX compared to the total database.

Maximal performance characteristics for laboratory mammal data compared to human results were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally poor compared to that for the other two species. The rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity results appeared to be slightly but consistently improved over the performance of FETAX when TI was used as the single decision criterion.

NICEATM conducted an evaluation for the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data. Performance characteristics (accuracy, sensitivity, specificity) were calculated for FETAX compared to combined laboratory mammal (rat, mouse, and rabbit) or human teratogenicity results. When compared to combined laboratory mammal or human teratogenicity results, accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of 85% was accompanied by specificity of 40% or less. The magnitude of these values suggests that FETAX is not appropriate as a replacement for *in vivo* laboratory mammal teratogenicity tests, and that its use as a screen, based on current decision criterion, is problematic.

An analysis of FETAX database revealed 43 substances that were discordant with laboratory mammal teratogenicity results and/or human teratogenicity results, seven substances that were concordant with laboratory mammal teratogenicity data but discordant with human teratogenicity results, and three substances that were discordant with laboratory mammal but concordant with human teratogenicity results. The bases for these discordant results are not known.

The inclusion of an exogenous MAS in FETAX is considered to be essential for predicting developmental hazard in humans (ASTM, 1991; 1998). Two FETAX validation studies (Phase III.2 and Phase III.3) were conducted in which substances were tested with and without metabolic activation. However, selection of the substances tested did not appear to have been based on whether or not metabolic activation was required for teratogenic activity *in vitro*. NICEATM evaluated the possible metabolic activation requiring status of these and other substances tested in FETAX with metabolic activation. Identification of the possible involvement of metabolic activation was based on whether the substance was positive in one or more *in vitro* genetic toxicological tests (generally the *Salmonella typhimurium* reverse mutation assay) with, but not without, metabolic activation. This method for identifying substances that may require metabolic activation to be teratogenic *in vitro* assumes a common mechanism between mutagenicity and teratogenicity that may not be valid. This evaluation revealed that most of the 35 substances tested with metabolic activation were not known to require metabolic activation to be active *in vitro*, but that there was a tendency towards increased activity in FETAX with metabolic activation for those substances that required metabolic activation to be genotoxic *in vitro*. Based on the limited database, additional studies to validate the role of metabolic activation in FETAX appear to be justified.

Several approaches have been suggested for modifying the decision criteria used to distinguish between a positive and a negative FETAX response. These approaches include an evaluation of the EC_{50} based on characteristic malformations only, a point estimate rather than an MCIG for growth inhibition, and 95% confidence intervals for statistically identifying TI values (and other point estimates) that are significantly greater than the decision point. The effects of these suggested approaches on the performance characteristics of FETAX have not yet been evaluated.

Another aspect of characteristic malformations in FETAX that has yet to be critically explored in the correlation between the types of agent-specific malformations induced in *X. laevis* and those induced by the same agent in rats, mice, and/or rabbits, or in humans. An evaluation is in progress by NTP using data collected in the FETAX Phase III.3 Validation Study. The results of this assessment may indicate the appropriateness of additional research in this area.