## Prolactin Levels During Long-Term Risperidone Treatment in Children and Adolescents: A Reanalysis of Data

**To the Editor:** We are writing in regard to our paper "Prolactin Levels During Long-Term Risperidone Treatment in Children and Adolescents,"[1] published in the November 2003 issue of *JCP*. The reason for this letter is to respond to concerns raised about this paper both during court proceedings and within the lay media. Janssen is a defendant in these legal cases. We are the 2 remaining living authors of this paper who were never employees of Janssen. A third, Thomas Moshang from Children's Hospital of Philadelphia, has since died.

To begin, at no time during our initial drafting of the paper did we observe any suggestion of inappropriate behavior. The first concerns about the content of this paper came to our attention almost a decade after the paper was published when we learned that the results of specific data analyses were not made available to us during the drafting of this paper.

We take the concerns that have been raised very seriously. Only recently did we obtain both the data sets and information necessary to do the requisite analyses, as well as obtain access to some of the original computer code used for the original analyses. The reanalysis that follows is entirely dependent on the integrity of the data provided to us by Janssen.

As part of this response, an independent statistician was identified—Dr Warren Bilker, who has worked closely with us to address these concerns. Janssen provided Dr Bilker with data sets and additional information about this paper's original analyses in order to address the issues that have been raised. Janssen funded Dr Bilker's efforts as part of a Data Access Agreement.

We have taken the following approach: first, we report findings of our independent analyses, which examined whether the results reported in the original paper could be replicated/verified. Second, we address the accusation that the manner in which the data were presented in this paper was misleading. Next, we address the assertion that an important safety signal pertaining to gynecomastia was not communicated to the *Journal* readership. Subsequently, we address the fact that during the original drafting of this paper, several statistical analyses were performed about which we were not aware until recently (as noted above). As a result of assertions that have been made publicly, we focused our analyses on the relationship between prolactin levels in risperidone-treated youths and the development of gynecomastia.

### 1. VERIFICATION OF RESULTS

Our independent analyses first examined whether the results reported in the paper could be replicated by Dr Bilker. Dr Bilker was able to review almost all of the data presented in the paper (see below). That is due to the fact that although all of the data from all of the studies involved were provided to Dr Bilker, some of the programs and methodology used to produce a small number of the results in the original paper by an outside consulting company were not available at the time of this reanalysis.

*The verification of results is listed in the same order and using the same headings as in the paper.*

#### Results—Patients and Treatment Information

- The statement "There was no statistical difference in gender, age, height, weight, body mass index (BMI), Tanner stage, IQ rating, or *DSM-IV* Axis II diagnosis of intellectual functioning"[1(p1365)] was verified with no differences found from the paper.
- The statements about the mean daily dose and mean duration of treatment for the intent-to-treat (ITT), primary analysis (PA), and non-PA populations were not verified due to the programs/methodology not being provided for these results at the time of this reanalysis.
- The last sentence of this section did have a few minor errors. They are noted in the text below using strikethrough for the replaced values.

> The PA populations included 489 males (82.6%) and 103 females (17.4%) with CD, oppositional defiant disorder, or DBD-NOS, with or without ADHD. The mean IQ of the patients was 65.1, and mental retardation was considered borderline in 40%, mild in 42%, and moderate in 18%. Patients had a mean age of ~~9.9~~ 9.4 years, and the majority ~~73%~~ 83.1% were in Tanner stage 1 of puberty when they began the study. Mean height was ~~137.8~~ 137.7 cm, mean weight was ~~35.4~~ 35.3 kg (~~78.7~~ 77.8 lb), mean BMI was ~~18.0~~ 18.2, and ~~80%~~ 80.2% of the patients were white.

### Results—Prolactin Levels

- All values in the first 2 paragraphs of this section were verified as correct. The third paragraph could not be verified due to the programs/methodology for discontinuation results not being available at the time of this reanalysis.
- The values presented in the subsections "By Gender," "By Age," and "By Gender and Age" were all verified with no differences found from the paper.
- All values in Table 1 were verified with no differences found from the paper.
- Figure 1 and all values included in Figure 1 were verified with no differences found from the paper.

### Results—Side Effects Hypothetically Attributable to Prolactin [SHAP(A)]

- Table 2. A single female patient was incorrectly identified as having gynecomastia. However, if this female is recategorized as having an adverse event of breast enlargement, several changes occur in Table 2. The changes are noted below with strikethrough text. The denominators used for male- and female-specific disorders in the paper were the full sample sizes and have been modified below to be the number of males or females, respectively. Additional information about SHAP(A) recovery is provided at the bottom of Table 2.

**Table 2. SHAP(A) Patients in the ITT, Primary Analysis (PA), and Non-PA Populations**

| Variable | ITT (N = 700) (515 males, 185 females) | PA (N = 592) (489 males, 103 females) | Non-PA (N = 108) (26 males, 82 females) |
|---|---|---|---|
| No. of patients with at least 1 SHAP | 34 (4.9) | 30 (5.1) | 4 (3.7) |
| Reports of SHAP (by preferred term) | | | |
| Gynecomastia (males) | ~~25 (3.6)~~ 24 (4.7) | 22 ~~(3.7)~~ (4.5) | ~~3 (2.8)~~ 2 (7.7) |
| Reproductive disorders, female | ~~9 (1.3)~~ 10 (5.4) | 8 ~~(1.4)~~ (7.8) | ~~1 (0.9)~~ 2 (2.4) |
| Amenorrhea | 4 ~~(0.6)~~ (2.2) | 3 ~~(0.5)~~ (2.9) | 1 ~~(0.9)~~ (1.2) |
| Menorrhagia | 3 ~~(0.4)~~ (1.6) | 3 ~~(0.5)~~ (2.9) | 0 (0.0) |
| Breast enlargement | ~~1 (0.1)~~ 2 (1.1) | 1 ~~(0.2)~~ (1.0) | ~~0 (0.0)~~ 1 (1.2) |
| Lactation nonpuerperal | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |
| Menstrual disorder | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |
| Vaginal hemorrhage | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |

*Note: All differences noted above are due to a single female coded as having gynecomastia, which should have been coded as "breast enlargement."*

*There are 592 subjects in the PA study. Of these, 30 had a SHAP(A) (5.1%). Of these, 14/30 (46.7%) recovered by the end of the study, and 16/30 (53.3%) did not resolve by the end of the study. Percentage with nonresolved SHAP(A) at end of study is 16/592 (2.7%) [1/8 females and 15/22 males had a nonresolved SHAP(A)].*

- Table 3. No differences were found in the number of events shown in Table 3 on reanalysis. However, the denominators used for male- and female-specific disorders in the paper were the full sample sizes and have been modified be to be the number of males or females, respectively (see next section of letter for modified table).
- All values in the text from below Table 3 to the bottom of the left column on page 1367 were verified as correct.
- Table 4. There were some differences identified in Table 4. They are noted in the table below with strikethrough of the modified values from verification. The values marked with an asterisk could not be verified with the data provided to us.

**Table 4. Comparison of SHAP Populations (primary analysis populations)**

| Parameter | SHAP(A) (N = 30) | SHAP(B) (N = 13) |
|---|---|---|
| Age of boys, mean, y | ~~11.4~~ 11.1 (n = 22) | ~~7.8~~ 7.4 (n = 5) |
| Age of girls, mean, y | ~~12.8~~ 12.1 (n = 8) | ~~12.8~~ 12.1 (n = 8) |
| Time to onset of first SHAP/without SHAP resolved | NA[a] | NA[a] |
| Risperidone dose with SHAP/without SHAP, mean, mg/d | NA[a] | NA[a] |
| Patients with SHAP resolved at study end, N | ~~17~~ 14 | 9 |
| Patients with SHAP and prolactin levels above ULN during any time period, range, % | 4.7–7.8 | 1.8–3.5 |
| Patients with SHAP and normal prolactin levels during any time period, range, % | 2.9–6.5 | 1.2–3.0 |

[a]*NA = not available because original programs/methodology were not available to confirm.*

- The values in the text below Table 4, page 1367 right side, through the end of the paragraph ending with "Further SHAP results refer to the SHAP(B) analysis" were all verified as correct.
- The paragraph beginning with "The mean (SD) daily dose of risperidone" could not be verified with the available data and documentation of the dose analysis programs.
- The paragraph beginning with "A total of 15 SHAP were reported" was verified as correct.
- In the next paragraph, there was 1 discrepancy, as noted with a strikethrough: "Only 1 of the patients with these prolactin levels, a ~~12.5~~ 12.0-year-old female, had SHAP."

### Results—Prolactin Levels and Extrapyramidal Symptoms (EPS)

In the next paragraph, the underlined text could not be verified with available data.

"Altogether, 129/592 patients in the PA population (21.8%) reported at least 1 EPS versus 18/108 (16.7%) in the non-PA population. The mean (SD) onset of the first EPS was 64.3 (99.3) days in the PA population and 40.6 (69.6) days in the non-PA population. There was no significant difference in the percentage of patients who experienced EPS with mean prolactin levels in the normal range (21.0%–24.5%) versus those at or above the upper limit of normal (ULN) (20.9%–~~24.3%~~ 27.6%)."

### Results—Prolactin Levels and Score on the Conduct Problem Subscale of the N-CBRF

It was reported that there was no significant correlation between prolactin levels and the improvement on the conduct problem subscale of the N-CBRF (correlations ranged from 0.10 to 0.02). These correlations appear to be taken from untransformed values of N-CBRF and prolactin levels, which yield correlations in the range −0.09 to −0.02. However, it is the change in N-CBRF, rather than the absolute N-CBRF, that is needed to consider improvement. The percent change in N-CBRF from baseline is highly left skewed, and thus Spearman correlations were applied. The Spearman correlations of the percent change in N-CBRF from baseline with prolactin levels over the time periods considered range from 0.03 to 0.09, with no significant correlations.

### Results—Prolactin Levels and Risperidone Dose

Could not be verified due to programs and methodology not being available at the time of this reanalysis.

From the data available to us for review, the changes/errata noted do not change the conclusions or interpretation of the original paper. In the spirit of transparency, we believe these new findings, although they are both modest and do not alter our interpretation of the study, should be communicated to readers. Based on the data provided to us from Janssen, we believe that no other revisions to our paper are indicated. Furthermore, as we have found no evidence of falsification, we do not believe this paper should be retracted from the medical literature.

## 2. MALES-ONLY ANALYSES

Since concerns have been raised specifically about our having combined the males and females in the original analysis, we believe it is important to present separate analyses for males.

Tables and a figure that include only male participants are listed next.

**Figure 1. Prolactin Levels in Children Receiving Long-Term Risperidone Treatment (as presented in paper—no differences found on reanalysis)**

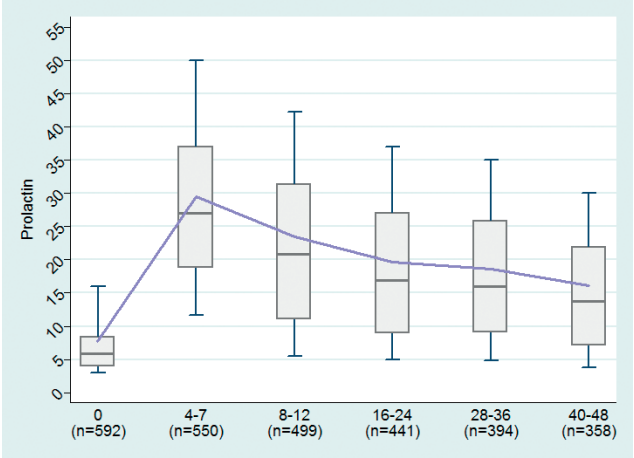


**Figure 1M. Prolactin Levels in Children Receiving Long-Term Risperidone Treatment: Males Only**

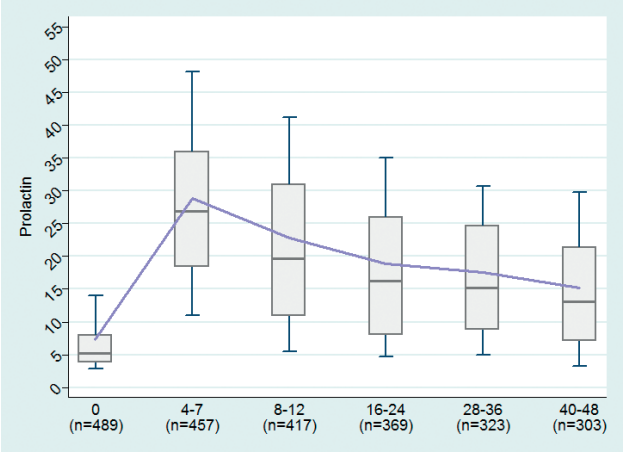



**Table 2M. SHAP(A) Patients in the ITT, Primary Analysis (PA), and Non-PA Populations: Males Only (excludes female disorders)**

| Variable | ITT (N = 515) | PA (N = 489) | Non-PA (N = 26) |
|---|---|---|---|
| No. of patients with at least 1 SHAP | 24 (4.7) | 22 (4.5) | 2 (7.7) |
| Reports of SHAP (by preferred term) | | | |
| Gynecomastia (males) | 24 (4.7) | 22 (4.5) | 2 (7.7) |

*Note: There are 489 subjects in the PA study. Of these, 22 had a SHAP(A) (4.5%). Seven of 22 (31.8%) resolved by the end of study participation, and 15/22 (68.2%) did not resolve by study's end. Percentage with nonresolved SHAP(A) at end of study is 15/489 (3.1%).*

No differences were found in the number of events shown in Table 3 on reanalysis. However, the denominators used for male- and female-specific disorders in the paper were the full sample sizes and have been modified below to be the number of males or females, respectively.

**Table 3. SHAP(B) Patients in the ITT, Primary Analysis (PA), and Non-PA Populations**

| Variable | ITT (N = 700) (515 males, 185 females) | PA (N = 592) (489 males, 103 females) | Non-PA (N = 108) (26 males, 82 females) |
|---|---|---|---|
| No. of patients with at least 1 SHAP | 14 (2.0) | 13 (2.2) | 1 (0.9) |
| Reports of SHAP (by preferred term) | | | |
| Gynecomastia (males) | 5 ~~(0.7)~~ (1.0) | 5 ~~(0.8)~~ (1.0) | 0 (0.0) |
| Reproductive disorders, female | 9 ~~(1.3)~~ (4.9) | 8 ~~(1.4)~~ (7.8) | 1 ~~(0.9)~~ (1.2) |
| Amenorrhea | 4 ~~(0.6)~~ (2.2) | 3 ~~(0.5)~~ (2.9) | 1 ~~(0.9)~~ (1.2) |
| Menorrhagia | 3 ~~(0.4)~~ (1.6) | 3 ~~(0.5)~~ (2.9) | 0 (0.0) |
| Breast enlargement | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |
| Lactation nonpuerperal | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |
| Menstrual disorder | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |
| Vaginal hemorrhage | 1 ~~(0.1)~~ (0.5) | 1 ~~(0.2)~~ (1.0) | 0 (0.0) |

*Note: There are 592 subjects in the PA study. Of these, 13 had a SHAP(B) (5.1%). Of these, 9/13 (69.2%) resolved by the end of the study, 4/13 (30.8%) did not resolve by the end of the study. Percentage with nonresolved SHAP(B) at end of study is 4/592 (0.7%) [1/8 females and 3/5 males had a nonresolved SHAP(B)].*

**Table 3M. SHAP(B) Patients in the ITT, Primary Analysis (PA), and Non-PA Populations: Males Only (excludes female disorders)**

| Variable | ITT (N = 515) | PA (N = 489) | Non-PA (N = 26) |
|---|---|---|---|
| No. of patients with at least 1 SHAP | 5 (1.0) | 5 (1.0) | 0 (0.0) |
| Reports of SHAP (by preferred term) | | | |
| Gynecomastia (males) | 5 (1.0) | 5 (1.0) | 0 (0.0) |

*Note: There are 489 subjects in the PA study. Of these, 5 had a SHAP(B) (1.0%). Of these 2/5 (40.0%) resolved by the end of the study, and 3/5 (60.0%) did not resolve by the end of the study. Percentage with nonresolved SHAP(B) at end of study is 3/489 (0.6%).*

We hope presenting these additional males-only tables provides greater clarity to the *Journal*'s readership.

## 3. ADDITIONAL GYNECOMASTIA DATA

The greatest concern to us was the assertion that there was an important safety signal about the risk of gynecomastia in males within the extant data that was not reported in this paper. We believe the additional males-only data that we have reported above address, in part, these concerns.

It should be noted that the original paper did not focus primarily on the issue of gynecomastia. For that reason, we did not focus on that specific issue in the original paper. However, due to the concerns noted during legal proceedings regarding the relationship between risperidone and gynecomastia, we further examined the data provided to us relating to gynecomastia.

First, we considered how often patients were discontinued from these clinical trials due to adverse events. We noted that 7 patients discontinued due to an adverse event. All of these 7 patients discontinued due to gynecomastia that was considered to be "moderate" in severity. These 7 patients were all participants in the international study (referred to as *INT* in the original paper).

Of particular concern to us was the severe and disfiguring gynecomastia that has been reported in the media. From the data provided to us, no participants had gynecomastia rated as being "severe." All reported gynecomastia events were either mild or moderate.

## 4. REEVALUATION OF UNPUBLISHED ANALYSES: PROLACTIN AND SIDE EFFECTS

During these legal proceedings, we learned that several statistical analyses were conducted about which we were not aware at the time the original paper was published. Concerns that have been raised in legal proceedings and the media have focused on 2 series of analyses.

The first, which has been referred to as "Table 21," was a series of analyses that examined the relationship between a patient's having a prolactin level above the upper limit of normal (ULN) at a specific time point and the presence of side effects hypothetically attributable to prolactin (SHAP[A]) for that specific patient at any time point during the course of study participation. These concerns pertain to whether or not gynecomastia occurring in children and youth treated with risperidone is related to the risperidone-induced changes in prolactin concentrations.

### Original Analyses of Table 21

From the data provided to us recently, the analysis originally done for Table 21 used 6 separate $2 \times 2$ tables and provided a *P* value for each table for the association between prolactin concentrations "above ULN" at the specific time period and the presence of a SHAP(A) at any time period. The typical procedure for testing the association across the 6 tables is to first test the homogeneity of the odds ratios (ORs) across the 6 tables, using a Mantel-Haenszel (MH) like test of homogeneity. If the ORs are heterogeneous, then the 6 tables are examined separately. Since there are 6 *P* values, 1 for each table, a multiple comparisons adjustment must be made to adjust for the inflated type I error rate.

The MH test assumes that the 6 groups used to form the 6 tables are independent, an assumption that is not valid in this case, since the majority of subjects have observations for multiple time periods and thus contribute to multiple tables. A Mantel-Haenszel like test of homogeneity of the ORs accommodating this clustering of observations was obtained using the logistic regression form of the MH test and applying a Huber-White variance adjustment, a model including "above ULN," time period, and the interaction of these variables. A test for the significance of the interaction provides a cluster adjusted extension of the homogeneity test. The extension of the test of H0: OR = 1 that accommodates the clustered observations developed by Begg[2] and implemented in SAS by Begg and Paykin[3] was used to obtain the correct test for this hypothesis.

All values in the original Table 21, which appears below, were verified as correct.

**Table 21. Original Version**

Table 21:

| | shap_a ever | | | | | |
|---|---|---|---|---|---|---|
| | aboveULN | 0 | 1 | Total | | |
| Pre-dose | | | | | OR = 1.415 | |
| | 0 | 535 | 28 | 563 | $\chi^2(1)$ = | 0.2120 Pr = 0.645 |
| | 1 | 27 | 2 | 29 | | |
| | Total | 562 | 30 | 592 | | |
| | aboveULN | 0 | 1 | Total | | |
| Weeks 4-7 | | | | | OR = 1.488 | |
| | 0 | 156 | 6 | 162 | $\chi^2(1)$ = | 0.7148 Pr = 0.398 |
| | 1 | 367 | 21 | 388 | | |
| | Total | 523 | 27 | 550 | | |
| | aboveULN | 0 | 1 | Total | | |
| Weeks 8-12 | | | | | OR = 2.833 | |
| | 0 | 235 | 7 | 242 | $\chi^2(1)$ = | 5.8221 Pr = 0.016 |
| | 1 | 237 | 20 | 257 | | |
| | Total | 472 | 27 | 499 | | |
| | aboveULN | 0 | 1 | Total | | |
| Weeks 16-24 | | | | | OR = 0.786 | |
| | 0 | 248 | 17 | 265 | $\chi^2(1)$ = | 0.3229 Pr = 0.570 |
| | 1 | 167 | 9 | 176 | | |
| | Total | 415 | 26 | 441 | | |
| | aboveULN | 0 | 1 | Total | | |
| Weeks 28-36 | | | | | OR = 0.714 | |
| | 0 | 230 | 16 | 246 | $\chi^2(1)$ = | 0.5293 Pr = 0.467 |
| | 1 | 141 | 7 | 148 | | |
| | Total | 371 | 23 | 394 | | |
| | aboveULN | 0 | 1 | Total | | |
| Weeks 40-48 | | | | | OR = 0.964 | |
| | 0 | 234 | 14 | 248 | $\chi^2(1)$ = | 0.0052 Pr = 0.942 |
| | 1 | 104 | 6 | 110 | | |
| | Total | 338 | 20 | 358 | | |

The MH test of homogeneity of the ORs, accommodating the clustering, is not rejected ($\chi^2(5) = 10.67$, $P = .0582$). The MH combined OR is 1.218, and the hypothesis that the MH combined OR is equal to 1 is not rejected (cluster adjusted $\chi^2(1) = 0.6120$, $P = .4340$). Thus, it cannot be stated that any of the ORs across the 6 time periods are different from one another or that the overall (combined) OR is different from 1.

No testing would generally be done on the individual tables for the individual periods due to the above findings. However, the individual table *P* values for association were considered, as reported in Table 21, even though the MH test indicated homogeneity of the ORs.

In such instances, adjustments for multiple comparisons should be applied. There are a variety of possible multiple comparisons adjustments. To assure that the multiple comparisons adjustments were not driven by which test was selected, 7 different tests were applied: Bonferroni, stepdown Bonferroni, Sidak, stepdown Sidak, Hochberg, false discovery rate, and Benjamini-Hochberg-Yekutieli. The results of the Hochberg, false discovery rate, and Benjamini-Hochberg-Yekutieli procedures are presented. However, in no case did the conclusion differ for other procedures.

**Multiple Comparisons Testing of 6 Individual Table *P* Values for Table 21**

| Time Period | Unadjusted *P* Value | *P* Value Adjusted by Hochberg Procedure | *P* Value Adjusted by False Discovery Rate Procedure | *P* Value Adjusted by Benjamini, Hochberg, and Yekutieli Procedure |
|---|---|---|---|---|
| Pre-dose | .6452 | .9422 | .7742 | 1.0000 |
| Weeks 4–7 | .3979 | .9422 | .7742 | 1.0000 |
| Weeks 8–12 | .0158 | .0948 | .0948 | .2322 |
| Weeks 16–24 | .5699 | .9422 | .7742 | 1.0000 |
| Weeks 28–36 | .4669 | .9422 | .7742 | 1.0000 |
| Weeks 40–48 | .9422 | .9422 | .7742 | 1.0000 |

Thus, considering the 6 individual tables, after adjustment for multiple comparisons, there are no significant differences in the risk of SHAP(A) for those above the ULN versus those not above the ULN in any of the 6 time periods.

### Analyses With Males Only—Focus on Gynecomastia

Table 21 included both males and females. It is helpful to consider a males-only version of Table 21. Table 21M specifically considers the gynecomastia adverse event in males only.

**Table 21M. Males-Only Analysis**

Table 21M:

| | aboveULN | shap_a ever 0 | 1 | Total | | | |
|---|---|---|---|---|---|---|---|
| Pre-dose | 0 | 441 | 21 | 462 | OR = 0.807 | | |
| | 1 | 26 | 1 | 27 | $\chi^2(1)$ = | 0.0421 | Pr = 0.837 |
| | Total | 467 | 22 | 489 | | | |
| Weeks 4-7 | 0 | 110 | 4 | 114 | OR = 1.346 | | |
| | 1 | 327 | 16 | 343 | $\chi^2(1)$ = | 0.2732 | Pr = 0.601 |
| | Total | 437 | 20 | 457 | | | |
| Weeks 8-12 | 0 | 187 | 4 | 191 | OR = 3.323 | | |
| | 1 | 211 | 15 | 226 | $\chi^2(1)$ = | 4.9126 | Pr = 0.027 |
| | Total | 398 | 19 | 417 | | | |
| Weeks 16-24 | 0 | 202 | 12 | 214 | OR = 0.678 | | |
| | 1 | 149 | 6 | 155 | $\chi^2(1)$ = | 0.5842 | Pr = 0.445 |
| | Total | 351 | 18 | 369 | | | |
| Weeks 28-36 | 0 | 181 | 12 | 193 | OR = 0.479 | | |
| | 1 | 126 | 4 | 130 | $\chi^2(1)$ = | 1.6274 | Pr = 0.202 |
| | Total | 307 | 16 | 323 | | | |
| Weeks 40-48 | 0 | 192 | 11 | 203 | OR = 0.919 | | |
| | 1 | 95 | 5 | 100 | $\chi^2(1)$ = | 0.0235 | Pr = 0.878 |
| | Total | 287 | 16 | 303 | | | |

The MH test of homogeneity of the ORs is rejected ($\chi^2(5) = 14.35$, $P = .0135$). This indicates that for at least 1 pair of tables the association between "above ULN" at the specific time period and the presence of a SHAP(A) at any time period is different, as measured by the OR, but not necessarily that any of the ORs are different from 1. The ORs range from 0.479 to 3.323. Since statistically significant heterogeneity of the odds ratios was detected, the MH combined OR is not presented. It is helpful to consider the individual table odds ratios and *P* values.

There are 6 *P* values in the above analysis for Table 21M. The same multiple comparisons procedures were applied as in Table 21.

**Multiple Comparisons Testing of 6 Individual Table *P* Values for Table 21M**

| Time Period | Unadjusted *P* Value | *P* Value Adjusted by Hochberg Procedure | *P* Value Adjusted by False Discovery Rate Procedure | *P* Value Adjusted by Benjamini, Hochberg, and Yekutieli Procedure |
|---|---|---|---|---|
| Pre-dose | .8375 | .8782 | .8782 | 1.0000 |
| Weeks 4–7 | .6012 | .8782 | .8782 | 1.0000 |
| Weeks 8–12 | .0267 | .1602 | .1602 | .3925 |
| Weeks 16–24 | .4447 | .8782 | .8782 | 1.0000 |
| Weeks 28–36 | .2021 | .8782 | .6063 | 1.0000 |
| Weeks 40–48 | .8782 | .8782 | .8782 | 1.0000 |

Thus, even if one were to consider testing the 6 individual tables, after adjustment for multiple comparisons, there are no significant differences in the risk of gynecomastia for those above the ULN versus those not above the ULN in any of the 6 time periods.

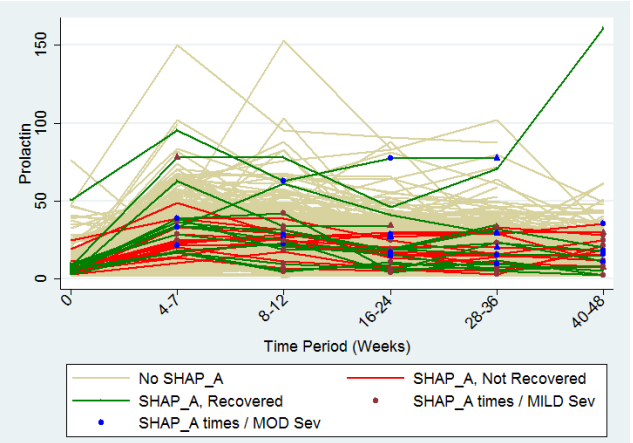
### Longitudinal Approach

Table 21 considers what are really longitudinal measurements of prolactin level and SHAP(A) in 6 separate cross-sectional analyses. A more informative approach to assessing the prolactin data is to consider longitudinal graphics and analysis of the relationship between prolactin levels over time and their effect on SHAP(A) presentation over time.

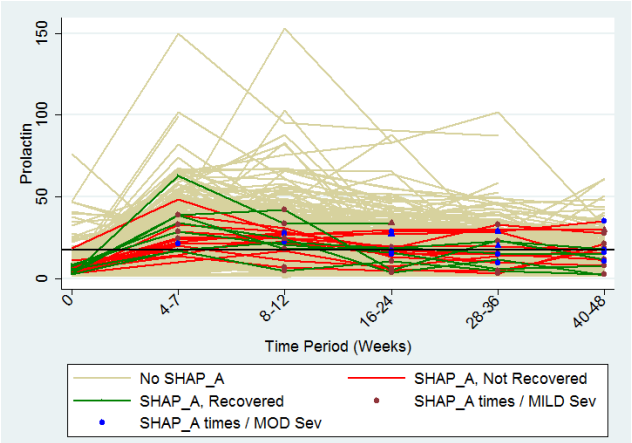
The following plots show the values of prolactin at each time measured for each subject in the group being represented in each plot. There is a separate connected line for each subject. The lines are color coded, with gold used for each subject that was never diagnosed with a SHAP(A) during the course of the study, green used for each subject that had at least 1 SHAP(A) diagnosis but had recovered from the SHAP(A) by the end of the study follow-up period, and red used for each subject that had at least 1 SHAP(A) diagnosis and had not recovered from the SHAP(A) by the end of the study follow-up period. Each prolactin level for a subject that was associated with a SHAP(A) diagnosis is denoted by either a blue or red marker. A red marker denotes a SHAP(A) of mild severity, while a blue marker denotes a SHAP(A) of moderate severity. Note that there were no SHAP(A)s that were of severe severity. There were a small number of SHAP(A)s at specific time periods for which no prolactin measurements were available. Values for prolactin were imputed in these cases to facilitate placing markers on the plots to represent these SHAP(A) diagnoses. A circle marker was used for actual values, while a triangle marker was used for imputed values. In cases in which imputed values were used, the higher of the prolactin values before and after the missing prolactin value was used. Plots are presented for the PA here, but were also completed for the ITT and non-PA groups. Plots are provided with the following sets of subjects: "all," "subjects with SHAP(A) who recovered by end of study," and "subjects with SHAP(A) who did not recover by end of study." These 4 plots are then repeated considering males only. For the males-only plots, a horizontal line at a prolactin level of 18 ng/mL, the ULN for males, is included.

It is seen that the subjects with SHAP(A) do not appear to have patterns of prolactin levels that are higher than those of subjects that did not have a SHAP(A). The same is true of the males-only longitudinal plots.

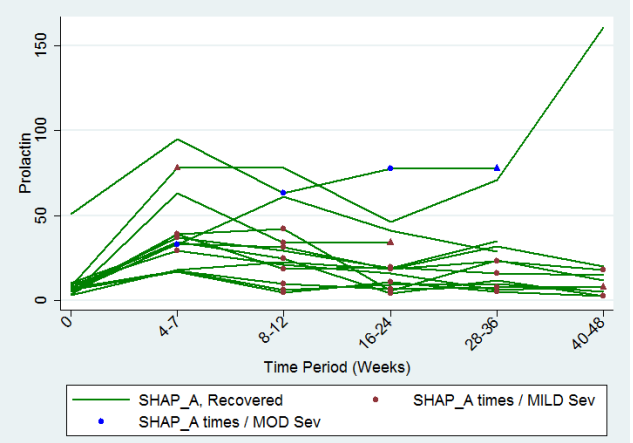
PA Group—All



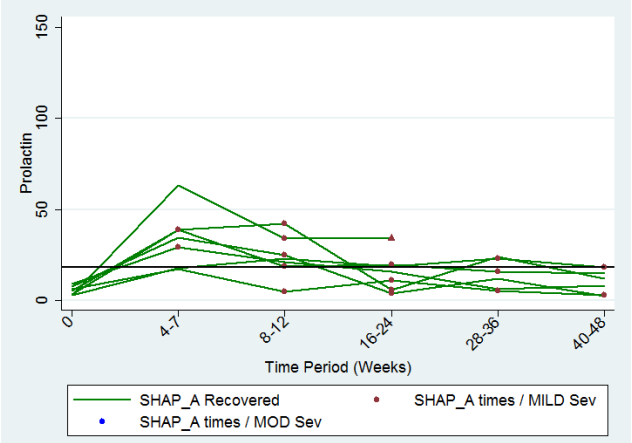

PA Group—Males Only



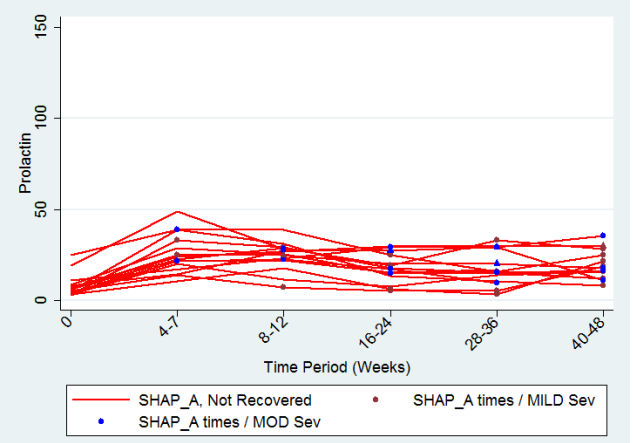

PA Group—Subjects With SHAP(A) Who Recovered by End of Study



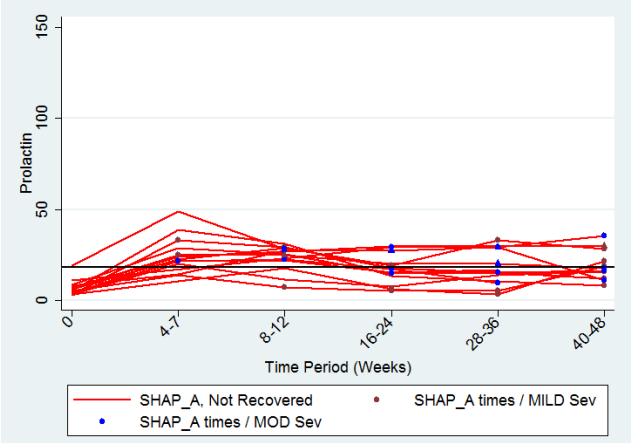

PA Group—Subjects With SHAP(A) Who Recovered by End of Study: Males Only




PA Group—Subjects With SHAP(A) Who Did Not Recover by End of Study




PA Group—Subjects With SHAP(A) Who Did Not Recover by End of Study: Males Only



## Longitudinal Data Analysis

The Table 21 analysis considered SHAP(A) at any time period as the outcome, rather than SHAP(A) at each time period separately. Additionally, in the Table 21 analysis, the prolactin measurement was binary—above ULN, yes or no—rather than considering the continuous prolactin level. A more informative analysis of the data is to consider the continuous prolactin level and SHAP(A) (yes/no) for each time period for every subject. Thus, each subject can contribute up to 6 observations, 1 for each time period. The multiple observations per subject are nonindependent, and this feature of the data is accommodated in the logistic regressions used to model the data (cluster option for logistic regression in STATA 14, which applies the Huber-White adjustment).

First, consider all subjects from the PA population. Note that the "pre-dose" time period was excluded from the models since there were no SHAP(A)s during this time period, leaving 5 time periods. The "weeks 4–7" period was considered the baseline period in the models. First, consider a model with prolactin level (continuous), time period (5-level factor), and the prolactin-by–time period interaction. The interaction was not significant ($P = .7083$) and thus was dropped. The model including prolactin level (continuous) and time period (5-level factor) had a significant global (Wald) test for time period ($P = .033$), while prolactin level was not significant ($P = .344$). The final model considered is thus a model including prolactin level and time period.

| Covariate | OR | *P* Value | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| Prolactin | 1.009 | .344 | 0.991 | 1.027 |
| Weeks 4–7 (baseline) | 1.000 | … | … | … |
| Weeks 8–12 | 2.098 | .047 | 1.009 | 4.365 |
| Weeks 16–24 | 2.625 | .011 | 1.248 | 5.524 |
| Weeks 28–36 | 2.427 | .039 | 1.048 | 5.623 |
| Weeks 40–48 | 2.952 | .007 | 1.353 | 6.442 |
| Time period (global test) | | .033 | | |

Since this appears to show an increasing trend in the risk of SHAP(A) across time periods, models were fit to assess this trend. In the following models, both prolactin and time period were considered continuous to test the trend over time period. First, a model including prolactin level (continuous), time period (continuous, 1–5), and the prolactin-by–time period interaction was fit. The interaction was not significant ($P = .999$), and thus it was dropped. The model, shown below, including prolactin level and time period had a significant time period effect ($P = .008$) and no prolactin level effect ($P = .402$). Thus, the risk of SHAP(A) appears not to be related to prolactin, but rather to time period, with an increased rate over time periods. This may be due to a detection bias or possibly increased obesity over time or some other factor(s).

| Covariate | OR | *P* Value | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| Prolactin | 1.008 | .402 | 0.990 | 1.026 |
| Time period | 1.234 | .008 | 1.056 | 1.441 |

Next, consider repeating the above analyses with males only from the PA population. First, consider a model with prolactin level (continuous), time period (5-level factor), and the prolactin-by–time period interaction. The interaction was not significant ($P = .493$) and thus was dropped. The model including prolactin level (continuous) and time period (5-level factor) had a significant global (Wald) test for time period ($P = .037$), while prolactin level was not significant ($P = .856$). The final model considered is thus a model including prolactin level and time period.

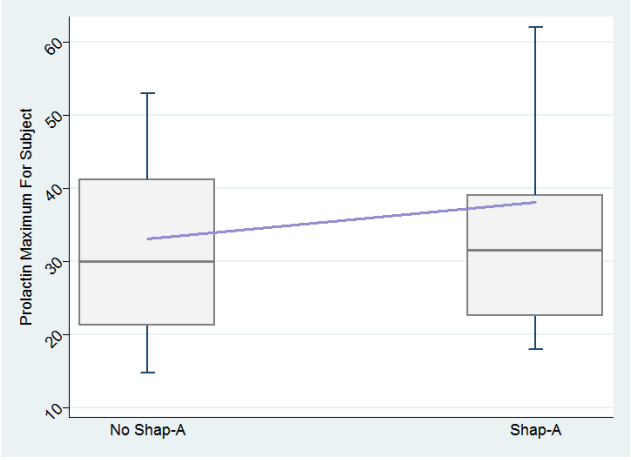
| Covariate | OR | *P* Value | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| Prolactin | 1.002 | .856 | 0.983 | 1.020 |
| Weeks 4–7 (baseline) | 1.000 | … | … | … |
| Weeks 8–12 | 2.699 | .022 | 1.153 | 6.319 |
| Weeks 16–24 | 2.339 | .011 | 1.317 | 8.464 |
| Weeks 28–36 | 3.229 | .033 | 1.098 | 9.490 |
| Weeks 40–48 | 4.134 | .006 | 1.505 | 11.362 |
| Time period (global test) | | .037 | | |

Since this appears to show an increasing trend over time period, models were fit to assess this possibility. In the following models, both prolactin and time period were considered continuous to test the trend over time period. First, a model including prolactin level (continuous), time period (continuous, 1–5), and the prolactin-by–time period interaction was fit. The interaction was not significant ($P = .275$), and thus it was dropped. The model, shown below, including prolactin level and time period had a significant time period effect ($P = .008$) and no prolactin level effect ($P = .965$). Thus, the risk of SHAP(A) for males only appears not to be related to prolactin, but rather to be related to time period with an increased rate over time periods. This may be due to a detection bias or possibly increased obesity over time or some other factor(s).

| Covariate | OR | *P* Value | Lower 95% CI | Upper 95% CI |
|---|---|---|---|---|
| Prolactin | 1.0004 | .965 | 0.981 | 1.020 |
| Time period | 1.291 | .008 | 1.070 | 1.557 |

In order to graphically assess the impact of large prolactin values on the risk of SHAP(A), boxes were created, similar to Figure 1 in the original paper. The boxes show maximum prolactin levels across all time periods within subject, separately for those that had a SHAP(A) event at any time and those that did not have a SHAP(A) event. The horizontal line within the box represents the median across all subjects of the subject-specific maximum prolactin level. The line that connects the boxes connects the mean of the maximum prolactin levels. Boxes are shown for all in PA and for males only in PA.

**Boxes Representing Maximum Prolactin Levels Within Subject**




**Boxes Representing Maximum Prolactin Levels Within Subject: Males Only**


These plots do not indicate that the maximum prolactin values for each subject play a key role in their having a SHAP(A), either overall or for males only.

To summarize, when prolactin levels and SHAP(A) are examined using multiple techniques, we were not able to find a relationship between the magnitude of prolactin concentrations and SHAP(A) in general and gynecomastia in particular. It is possible that the gynecomastia observed in these individuals may have been due to weight gain and the resulting enhanced androgen-to-estrogen conversion in the increased body fat[4,5] or some other unknown determinant(s).

## 5. REEVALUATION OF UNPUBLISHED ANALYSES: PROLACTIN AND BEHAVIORAL RESPONSE

The other set of analyses about which concerns were raised has been referred to in legal proceedings as "Table 34." As with "Table 21," we were not aware of this table's existence at the time the original paper was published.

### Table 34 Analyses

The primary efficacy measure for these clinical trials was the Nisonger Child Behavior Rating Form (N-CBRF). The analysis originally done for Table 34, "Responders on the Conduct Problem Subscale of the N-CBRF by Prolactin Levels (PAP – As Observed): Frequency Tables," used 5 separate 2 × 2 tables and provided a *P* value for each table using 3 different response criteria. Thus, a total of 15 analyses were completed (5 analyses for each response criterion). The association between percent decrease in the N-CBRF (indicative of reduced symptomatology) and an elevated prolactin level above the ULN at a specific time period was examined.

The original analyses suggested that either a ≥ 25% or a ≥ 35% decrement in the N-CBRF was associated with having a prolactin level above the ULN at weeks 16–24. Statistically, the same issues discussed for Table 21 apply to Table 34. The multiple comparisons corrected analyses for the original table are presented here, with separate multiple comparisons adjustments made for each response criteria.

**Multiple Comparisons Testing of 5 Individual Table *P* Values for Table 34: Separate Multiple Comparisons Adjustments Were Made for Response Criteria**

| Time Period | Unadjusted *P* Value | *P* Value Adjusted by Hochberg Procedure | *P* Value Adjusted by False Discovery Rate Procedure | *P* Value Adjusted by Benjamini, Hochberg, and Yekutieli Procedure |
|---|---|---|---|---|
| **≥ 25% vs 25%** | | | | |
| Weeks 4–7 | .6236 | .6762 | .6762 | 1.0000 |
| Weeks 8–12 | .6762 | .6762 | .6762 | 1.0000 |
| Weeks 16–24 | .0358 | .1790 | .1790 | .4087 |
| Weeks 28–36 | .1616 | .6464 | .4040 | .9225 |
| Weeks 40–48 | .5286 | .6762 | .6762 | 1.0000 |
| **≥ 35% vs 35%** | | | | |
| Weeks 4–7 | .7673 | .7821 | .7821 | 1.0000 |
| Weeks 8–12 | .7821 | .7821 | .7821 | 1.0000 |
| Weeks 16–24 | .0411 | .2055 | .2055 | .4692 |
| Weeks 28–36 | .7077 | .7821 | .7821 | 1.0000 |
| Weeks 40–48 | .4187 | .7821 | .7821 | 1.0000 |
| **≥ 50% vs 50%** | | | | |
| Weeks 4–7 | .5358 | .8457 | .7125 | 1.0000 |
| Weeks 8–12 | .8457 | .8457 | .8457 | 1.0000 |
| Weeks 16–24 | .2410 | .8457 | .7125 | 1.0000 |
| Weeks 28–36 | .5049 | .8457 | .7125 | 1.0000 |
| Weeks 40–48 | .5700 | .8457 | .7125 | 1.0000 |

Thus, considering the 5 individual tables separately for each of the 3 response criteria, after adjustment for multiple comparisons, there are no significant differences in the response rate, based on the change in N-CBRF, for those with a prolactin level above ULN versus those without a prolactin level above the ULN in any of the 5 time periods for any of the 3 response criteria.

All values in the original Table 34 were verified as correct. A version of Table 34 including males only, Table 34M, is included below.

Table 34M:

Weeks 4 to 7, Response Criteria >=25% vs <25%

| improve25 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 0.874 |
| 1 | 178 | 56 | 234 | $\chi^2(1) = 0.1300$ Pr = 0.7184 |
| 2 | 40 | 11 | 51 | |
| Total | 218 | 67 | 285 | |

Weeks 8 to 12, Response Criteria >=25% vs <25%

| improve25 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 0.878 |
| 1 | 131 | 95 | 226 | $\chi^2(1) = 0.1778$ Pr = 0.6732 |
| 2 | 33 | 21 | 54 | |
| Total | 164 | 116 | 280 | |

Weeks 16 to 24, Response Criteria >=25% vs <25%

| improve25 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 1.756 |
| 1 | 96 | 119 | 215 | $\chi^2(1) = 3.0728$ Pr = 0.0796 |
| 2 | 17 | 37 | 54 | |
| Total | 113 | 156 | 269 | |

Weeks 28 to 36, Response Criteria >=25% vs <25%

| improve25 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 1.487 |
| 1 | 89 | 123 | 212 | $\chi^2(1) = 1.5572$ Pr = 0.2121 |
| 2 | 18 | 37 | 55 | |
| Total | 107 | 160 | 267 | |

Weeks 40 to 48, Response Criteria >=25% vs <25%

| improve25 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 0.833 |
| 1 | 78 | 165 | 243 | $\chi^2(1) = 0.3580$ Pr = 0.5496 |
| 2 | 21 | 37 | 58 | |
| Total | 99 | 202 | 301 | |

Weeks 4 to 7, Response Criteria >=35% vs <35%

| improve35 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 0.990 |
| 1 | 159 | 49 | 208 | $\chi^2(1) = 0.0010$ Pr = 0.9745 |
| 2 | 59 | 18 | 77 | |
| Total | 218 | 67 | 285 | |

*(continued)*

Weeks 8 to 12, Response Criteria >=35% vs <35%

| improve35 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 117 | 87 | 204 |
| 2 | 47 | 29 | 76 |
| Total | 164 | 116 | 280 |

OR = 0.830
$\chi^2(1) = 0.4599$ Pr = 0.4977

Weeks 16 to 24, Response Criteria >=35% vs <35%

| improve35 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 87 | 106 | 193 |
| 2 | 26 | 50 | 76 |
| Total | 113 | 156 | 269 |

OR = 1.578
$\chi^2(1) = 2.6434$ Pr = 0.1040

Weeks 28 to 36, Response Criteria >=35% vs <35%

| improve35 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 76 | 113 | 189 |
| 2 | 31 | 47 | 78 |
| Total | 107 | 160 | 267 |

OR = 1.020
$\chi^2(1) = 0.0050$ Pr = 0.9434

Weeks 40 to 48, Response Criteria >=35% vs <35%

| improve35 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 71 | 146 | 217 |
| 2 | 28 | 56 | 84 |
| Total | 99 | 202 | 301 |

OR = 0.973
$\chi^2(1) = 0.0104$ Pr = 0.9189

Weeks 4 to 7, Response Criteria >=50% vs <50%

| improve50 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 136 | 39 | 175 |
| 2 | 82 | 28 | 110 |
| Total | 218 | 67 | 285 |

OR = 1.191
$\chi^2(1) = 0.3772$ Pr = 0.5391

Weeks 8 to 12, Response Criteria >=50% vs <50%

| improve50 | aboveULN 1 | 2 | Total |
|---|---|---|---|
| 1 | 96 | 72 | 168 |
| 2 | 68 | 44 | 112 |
| Total | 164 | 116 | 280 |

OR = 0.863
$\chi^2(1) = 0.3532$ Pr = 0.5523

*(continued)*

Weeks 16 to 24, Response Criteria >=50% vs <50%

| improve50 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 1.161 |
| 1 | 65 | 84 | 149 | $\chi^2(1) = 0.3584$ Pr = 0.5494 |
| 2 | 48 | 72 | 120 | |
| Total | 113 | 156 | 269 | |

Weeks 28 to 36, Response Criteria >=50% vs <50%

| improve50 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 0.800 |
| 1 | 57 | 94 | 151 | $\chi^2(1) = 0.7834$ Pr = 0.3761 |
| 2 | 50 | 66 | 116 | |
| Total | 107 | 160 | 267 | |

Weeks 40 to 48, Response Criteria >=50% vs <50%

| improve50 | aboveULN 1 | 2 | Total | |
|---|---|---|---|---|
| | | | | OR = 1.324 |
| 1 | 63 | 115 | 178 | $\chi^2(1) = 1.2363$ Pr = 0.2662 |
| 2 | 36 | 87 | 123 | |
| Total | 99 | 202 | 301 | |

There are no significant associations in Table 34 for males only without consideration of multiple comparisons and thus none with consideration of multiple comparisons.

## 6. AUTHORSHIP

Finally, concerns about authorship have been raised. To the best of our recollection, we met all 4 criteria for authorship according to International Committee of Medical Journal Editors guidelines. At the time of this paper's publication, we believed that nonauthor contributors were appropriately acknowledged. However, during the course of legal proceedings and reports in the media, we have learned that there may have been nonauthor contributors to this paper who were unknown to us at the time of the paper's publication. If there were nonauthor contributors, we do not know their identities or their specific contributions.

## 7. SUMMARY

To summarize, the results of our reanalysis support our statements in the manuscript. Of particular note, our finding that there was no direct correlation between prolactin elevation and SHAP is supported by the data in the reanalysis. In addition, our Conclusion section remains accurate.

### References


1. Findling RL, Kusumakar V, Daneman D, et al. Prolactin levels during long-term risperidone treatment in children and adolescents. *J Clin Psychiatry*. 2003;64(11):1362–1369.
2. Begg MD. Analyzing *k* (2 x 2) tables under cluster sampling. *Biometrics*. 1999;55(1):302–307.
3. Begg MD, Paykin AB. Performance of and software for a modified Mantel-Haenszel statistic for correlated data. *J Stat Comput Simul*. 2001;70(2):175–195.
4. Crocker MK, Stern EA, Sedaka NM, et al. Sexual dimorphisms in the associations of BMI and body fat with indices of pubertal development in girls and boys. *J Clin Endocrinol Metab*. 2014;99(8):E1519–E1529.
5. Kley HK, Deselaers T, Peerenboom H, et al. Enhanced conversion of androstenedione to estrogens in obese males. *J Clin Endocrinol Metab*. 1980;51(5):1128–1132.



**Robert L. Findling, MD, MBA**[a]
rfindli1@jhmi.edu
**Denis Daneman, MD**[b]

[a]Department of Psychiatry & Behavioral Sciences, Johns Hopkins University, Baltimore, Maryland

[b]Department of Pediatrics, University of Toronto and the Hospital for Sick Children, Toronto, Ontario, Canada

***Potential conflicts of interest:*** **Dr Findling** receives or has received research support from, acted as a consultant for, and/or served on a speakers bureau for Alcobra, American Academy of Child & Adolescent Psychiatry, American Physician Institute, American Psychiatric Press, AstraZeneca, Bracket, Bristol-Myers Squibb, CogCubed, Cognition Group, Coronado Biosciences, Dana Foundation, Elsevier, Forest, GlaxoSmithKline, Guilford Press, Johns Hopkins University Press, Johnson & Johnson, Jubilant Clinsys, KemPharm, Lilly, Lundbeck, Merck, National Institutes of Health, Neurim, Novartis, Noven, Otsuka, Oxford University Press, Pfizer, Physicians Postgraduate Press, Purdue, Rhodes Pharmaceuticals, Roche, Sage, Shire, Sunovion, Supernus Pharmaceuticals, Transcept Pharmaceuticals, Validus, and WebMD. **Dr Daneman** has no potential conflicts of interest to disclose.

***Funding/support:*** The original study was supported by Janssen-Ortho Inc. Janssen funded Dr Bilker's work on the reanalyses for this letter to the editor.

*J Clin Psychiatry 2016;77(2):e155–e170*
*dx.doi.org/10.4088/JCP.15l10500*



It is illegal to post this copyrighted PDF on any website.