

DECLARATION OF JANE E. (“BETH”) MORGAN
In Support of the Citizen Petition of GlaxoSmithKline

Docket No. 2004P-0239

I, Jane E. (“Beth”) Morgan, under penalty of perjury, declare as follows:

1. I work at GlaxoSmithKline (GSK) as a statistical consultant. I received a Ph.D. in Statistics from North Carolina State University in 1996, a M.S. in Statistics in 1991 from Virginia Tech, and a B.S. in Biostatistics from the University of North Carolina at Chapel Hill in 1988. In various capacities, I have worked in my field at GSK or predecessor companies more or less continuously since 1992, except for a period of months in 1999 when I was a visiting faculty member at North Carolina State University and taught a graduate level course entitled “Experimental Statistics for Biological Sciences II.”

2. My current responsibilities at GSK focus on the application of statistical methods to support the development, validation, and (as needed) site-transfer of processes for manufacturing and testing pharmaceutical products. The products I support include relatively complex drug-device combination products, such as inhalers and sprays for respiratory or intranasal use. It is critical that laboratory studies that measure the technical performance of such complex products, both in absolute quality terms, and also in relative terms (for instance, comparing former and current versions after a change in manufacturing method or

site), be designed and analyzed according to valid statistical methods. Over the years, I have also spent considerable time supporting colleagues engaged in early stage product development, including formulation scientists working in the respiratory field.

3. My outside professional activities include active participation in a Working Group of the Product Quality Research Institute (PQRI), and co-authorship of an impending paper summarizing the Working Group's deliberations. PQRI brings together professionals from the Food and Drug Administration's (FDA) Center for Drug Evaluation and Research, industry and academia to collaborate on research that generates scientific information to support regulatory policy. Specifically, the research efforts are directed to identifying the types of product quality information that should be submitted in regulatory filings to the FDA, and in streamlining that information if appropriate. I participate in the "Particle Size Distribution Profile Comparisons Working Group." It seeks to define appropriate statistical methods for evaluating one of the key *in vitro* tests that reflect the technical performance of inhaled and intranasal products. This work supports the overall development of appropriate *in vitro* tests that should be considered to assess the bioavailability and bioequivalence of nasal and inhalation drug products.

4. For products that are not intended to work by circulating in the bloodstream (*e.g.*, certain nasal sprays and inhalation products), bioequivalence means that there is no "significant difference" in the rate and extent to which the

active ingredient becomes available at the site of drug action. Comparisons of how “pioneer” and proposed generic products perform in this regard can be made using *in vitro* tests, *in vivo* tests, or a combination.

5. No matter what the selected bioequivalence test(s), a critical part is the statistical methodology used to analyze the results. The statistical methodology, along with the agreed-upon acceptance criteria, determine whether a valid conclusion can be drawn that the two products being compared are equivalent in their performance. Without defining, *in advance*, the statistical methods and criteria that will be applied, it is not possible to reach scientifically valid conclusions about relative product performance. If the criteria are not defined in advance, and applied fully and faithfully, evaluation of the results carries the scientifically unacceptable risk of being subjective and, ultimately, unreliable.

6. In May 1999, FDA issued a draft guidance document proposing a battery of *in vitro* tests (and for products formulated as suspensions, *in vivo* tests as well) that would, collectively, allow for a determination of bioavailability and bioequivalence for nasal aerosols and nasal sprays for local action. In April 2003, the agency issued a revised version of that guidance, again in draft. Notably, the April 2003 draft guidance was published without including information on the statistical methodologies and acceptance criteria to be used in evaluating the results of comparative measures taken in the various proposed tests (both *in vitro* and *in vivo*) for establishing the bioequivalence of nasal spray products. Instead,

the agency included placeholders for a series of statistical documents which, the agency stated, would be made available for comment. To date, FDA has not published any revised statistical documents, although some (outmoded) documents dating from the 1999 draft publication are available.

7. GSK, as the sponsor of pioneering intranasal medicines such as FLONASE® (fluticasone propionate) Nasal Spray, and BECONASE AQ® (beclomethasone dipropionate, monohydrate) Nasal Spray, has supported FDA's public deliberative process for defining tests sufficient to establish the bioequivalence of intranasal drug products. After FDA's publication of the initial draft in 1999, and then again in response to the updated April 2003 draft, GSK submitted significant comments, with a goal of contributing to the definition of scientifically robust standards in which patients and prescribers can have confidence.

8. In May 2004, GSK submitted a citizen petition urging FDA to finalize the guidance document and resolve certain fundamental scientific issues before approving generic versions of FLONASE®. GSK urged that prompt publication for comment of proposed statistical methods and criteria would be a necessary step toward completion of FDA's guidance. That petition is still pending with the agency, along with a supplemental petition seeking the same relief for proposed generic versions of BECONASE AQ®.

9. As discussed below, I have analyzed publicly available materials underlying a number of approvals in recent years of locally acting generic nasal spray products. The focus of my review was on the statistical methods and criteria used by FDA to determine whether the proposed generic products are bioequivalent to the brand-name pioneer products on which they are based. My purpose was to assess whether the statistical approaches supporting the bioequivalence assessments in these approvals were clear, were consistent from one application to the next, and were implemented in such a way as to correctly reach a conclusion of *in vitro* equivalence. I have concluded that they were not, as explained further below. This reinforces the need for FDA to establish valid bioequivalence methods for intranasal products by completing the guidance development process, including the supporting statistical documentation, *before* the agency reviews any new generic nasal spray products.

10. The approval documentation I reviewed pertains to the following abbreviated new drug applications (ANDAs):

- ANDA 76-156 for IPRATROPIUM BROMIDE 0.021MG SPRAY, sponsored by Apotex and approved April 18, 2003;
- ANDA 75-824 for BUTORPHANOL TARTRATE 1MG SPRAY, METERED, sponsored by Roxane and approved March 12, 2002;
- ANDA 75-759 for BUTORPHANOL TARTRATE 1MG SPRAY, sponsored by Mylan and approved August 8, 2001;

- ANDA 74-800 for CROMOLYN SODIUM 5.2MG NASAL SPRAY, sponsored by A.L. Pharma and approved July 26, 2001;
- ANDA 75-702 for CROMOLYN SODIUM 5.2MG SPRAY, sponsored by Bausch & Lomb and approved July 3, 2001; and
- ANDA 74-830 for DESMOPRESSIN ACETATE NASAL SOLUTION .01%, sponsored by Bausch & Lomb and approved Jan. 25, 1999.

11. Each of the ANDAs I reviewed is for a product that is formulated as a solution, whereas FLONASE® and BECONASE AQ® are formulated as suspensions. To my knowledge, FDA has yet to approve a generic version of a locally acting nasal spray product formulated as a suspension and has acknowledged that a showing of bioequivalence for suspension products is more challenging, and will require *in vivo* as well as *in vitro* studies.

12. Overall, I found that the applications did not contain complete statistical information that scientifically proved *in vitro* equivalence. Additionally, even when the incorrect approach was followed for showing *in vitro* equivalence, there were numerous examples in which statistically significant differences were observed but overlooked without any additional information proving equivalence. Details underlying my conclusions are presented below.

I. RECENTLY APPROVED ANDAS FOR NASAL SPRAY PRODUCTS HAVE ASSUMED EQUIVALENCE BETWEEN PRODUCTS UNLESS COMPELLING DATA PROVE OTHERWISE, WHEREAS THE SCIENTIFICALLY CORRECT WAY TO PROVE EQUIVALENCE IS TO ASSUME *INEQUIVALENCE* BETWEEN PRODUCTS AND *DISPROVE* THIS ASSUMPTION

13. FDA's draft guidance on bioavailability and bioequivalence studies for locally acting nasal spray products (June 1999, April 2003) would require that sponsors, at a minimum, conduct a battery of *in vitro* studies to demonstrate "equivalent performance" of a proposed generic product (the "test" product) and the approved pioneer (the "reference" product). These studies include, for example, comparative tests of the droplet sizes and spray patterns emitted from the test and reference products.

14. To reach a conclusion, statistically, based on a set of data, the first step is to set forth the "null" and "alternative" hypotheses. These are two statements that describe the true nature of the comparison. Because the null hypothesis is the starting default assumption, it is only the statement under the alternative hypothesis that can be statistically proven or stated with confidence. The null hypothesis describes the default or assumed condition between the two groups of data. The alternative hypothesis is the conclusion that the sponsor or investigator would like to make, provided the data support this conclusion. The correct null and alternative hypotheses for proving equivalence are stated in the following paragraph. The statement made under the null hypothesis cannot be statistically

under an assumption of inequivalence, the probability of having observed such a small difference in the data was less than 5%. This means that, if the measured difference were taken over and over again and if the products were truly not equivalent, then the likelihood of such a small difference occurring is quite low (for example, less than 5%).

16. Based on the ANDA documents I reviewed, it appears the agency has accepted a very different and less demanding approach of adopting a null hypothesis of equality. That is, the apparent default premise is that the test product (*i.e.*, the proposed generic) and the reference product (*i.e.*, the approved pioneer) are actually equivalent, that is, they have means that are equal for a specific *in vitro* measurement. In other words, it appears that FDA has approved ANDAs in which the null and alternative hypotheses stated were incorrectly reversed. With this incorrect approach, the default hypothesis of equality or equivalence is rejected only if a difference is observed between the test and reference products that is large enough to be highly unlikely to have occurred as a result of chance alone, *i.e.*, under the assumption of equality (typically, a probability of less than 5% is set as the threshold). If, however, the observed difference is not that great in light of the variability in the data, then the default null hypothesis of equality or equivalence cannot be rejected. I will refer to this less demanding approach, for convenience, as the “default equivalence” approach. Use of this approach does not constitute scientific proof of equivalence since large variability,

small sample sizes, or poor experimental design, can all lead to a failure to reject the null hypothesis of equality or equivalent performance between products.

17. The term “p value” is used as a shorthand reference to the probability of a result occurring under the null hypothesis, if the measurements were taken over and over again. As noted above, a probability threshold of less than 5% is typical for deciding whether to reject a null hypothesis. The statistical test is constructed, therefore, to control the risk of falsely concluding the alternative hypothesis to be 5% or less. In the “default equivalence” paradigm, if a “p value” for a measured difference was less than .05, then the measured difference would be considered significant enough to reject the null hypothesis of equality.

18. Starting with a null hypothesis of equality can, potentially, reward a sponsor or company for a poorly designed equivalence study that has small sample sizes or large variability because these factors can lead to failure to reject the null hypothesis of equality. For example, in the default equivalence approach, an observed p value of 0.06 would lead to a failure to reject the hypothesis of equality. A p value of 0.06 means that this difference, given the variability in the data, has only a 6% chance of occurring if the products had the same mean, or average, *in vitro* performance. However, a p value of 0.06 (or a 6% likelihood under an assumption of equality) should not be considered convincing evidence of equivalence or equal performance because, again, a poorly designed study can easily mask potential differences and lead to p values greater than 0.05.

19. Compare this “default equivalence” approach to the approach FDA generally applies for bioequivalence determinations. As described above, FDA’s standard methodology and the correct method for proving equivalence tests a null hypothesis that the generic product is not equivalent to the pioneer, *i.e.*, that the true difference between the products in the characteristic being measured exceeds a limit of acceptable difference. This hypothesis is rejected in favor of the alternative hypothesis of *in vitro* equivalence not only if the observed difference is below the limit of acceptable difference, but also if a difference this small is shown to have low probability of occurring if the products, in fact, did differ by more than the acceptable amount. The standard default position, in other words, is that the two products are *not* the same, and only compelling data to the contrary will support a conclusion of equivalence. This encourages sponsors to implement well designed studies; because, with this correct problem statement, “sloppy” experimentation is penalized by leading to a default conclusion of *inequivalence*.

20. The ANDA review documentation I analyzed suggests that, at times, sponsors have shown that the *observed* average difference between products lies within a pre-defined limit of acceptable difference, but this analysis is incomplete. It is not scientific proof of equivalence; it does not incorporate an assessment of the observed difference relative to the variability in the data. A small difference in the average is not proof of true *in vitro* product equivalence if large variability exists. Definitive, scientific proof of equivalence involves also quantifying the *likelihood* of

observing such a small difference, given the variability in the data, if the products were, in truth, inequivalent.

21. Based on my review of ANDAs for locally acting nasal sprays, I conclude that the FDA is allowing the use of a statistical approach that is at odds with the agency's usual approach, and at odds with the approach to which pioneering companies like GSK are typically held. The approach applied to these nasal spray products shifts from a presumption of *inequivalence* to a presumption of equivalence, in a manner that is likely to mask actual differences between the proposed generic product and the approved pioneer. Generic sponsors should be rewarded for developing truly equivalent products; unfortunately, the methodology put forth in recent ANDA approvals can reward a generic sponsor for submitting poor quality or highly variable data for an *inequivalent* product.

II. EVEN WITHIN ITS "DEFAULT EQUIVALENCE" APPROACH, FDA HAS REPEATEDLY REJECTED RESULTS THAT SUGGESTED STATISTICAL DIFFERENCES BETWEEN THE TEST AND REFERENCE PRODUCTS WITHOUT ADDITIONAL SCIENTIFIC JUSTIFICATION

22. Throughout my review, I found numerous instances in which the agency appears to have disregarded the scientific conclusion of its "default equivalence" approach, by overlooking measured differences with sufficiently low p values (less than 0.05) as to be considered "statistically significant."

23. The application of this “default equivalence” approach is illustrated by ANDA 75-824 for butorphanol tartrate 1mg nasal spray, metered, sponsored by Roxane. The applicant reported comparative measures (expressed as arithmetic and geometric means) in various tests of the proposed generic product versus the approved pioneer, along with the applicable p values. *See* FDA bioequivalence review at 7 (attached hereto as Tab 1). The null hypothesis was one of equivalence. Remarkably, for spray comparison No. 8, as shown in the FDA review document, the agency reached the conclusion of bioequivalence even when the p value associated with measured differences was less than 0.05, by dismissing the relative difference as being only 1.7%. Similar data are reported for spray pattern (SP), plume geometry (PG), droplet size distribution (DSD), cascade impaction, and priming studies on pages 7-13 of the review document. *See id.*

24. Another example is ANDA 75-759 for butorphanol tartrate 1mg spray, metered, sponsored by Mylan. The review documentation indicates that p values associated with some measured differences were less than 0.05. *See* FDA bioequivalence review at 9-15, 20-35 (attached hereto as Tab 2). These statistically significant differences were not addressed by the sponsor, nor did these differences prevent a regulatory determination of equivalence. While the review documents note that the observed ratio between the values for test and reference products fell within a plus/minus 10% range, they offer no explanation why this statistically significant difference did not preclude a finding of bioequivalence. *See id.*

25. Similarly, data submitted in support of ANDA 74-800 for a cromolyn sodium 5.2mg inhaler product, sponsored by A.L. Pharma, noted low sample sizes (n=10) for certain studies, as well as measured differences between test and reference products great enough to have low p values (less than 0.05). FDA appeared to disregard these low p values in making an overall finding of bioequivalence. *See* FDA bioequivalence review at 2-12 (attached hereto as Tab 3).

26. Data in an ANDA for a second cromolyn sodium product, ANDA 75-702 sponsored by Bausch & Lomb, appear to be even weaker than the data in the A.L. Pharma ANDA. First, no actual measurements from the equivalence tests were reported; the data were limited to the summary statistics of the ratios of the means (arithmetic means) for the test and reference products, and the associated p values. This summary report was deemed acceptable by the agency, even though a mix of p values – some less and some greater than 0.05 – were reported. *See* FDA bioequivalence review at 6-11 (attached hereto as Tab 4).

27. Finally, ANDA 75-156 for an ipratropium bromide nasal spray product, sponsored by Apotex, included a remarkable array of p values. Again, many were substantially less than 0.05, indicating a statistically significant measured difference between the products. The agency nonetheless stated that “most” of the differences were insignificant and that the product performances were “similar” or “comparable.” FDA bioequivalence review at 5-8, 24 (attached hereto as Tab 5).

The review documentation, however, fails to include any other information that might explain FDA's dismissal of the statistically different results. *See id.*

III. FDA'S ACCEPTANCE OF OBSERVED DIFFERENCES AGAINST A PREDEFINED LIMIT OF AN ACCEPTABLE DIFFERENCE IS NOT COMPLETE, AND APT TO YIELD INCORRECT CONCLUSIONS, WITHOUT ADDITIONAL STATISTICAL ANALYSIS

28. Some of the ANDA review documentation reported that the observed difference between test and reference products fell within a pre-defined limit of acceptable difference. However, the mere reporting of an observed difference, based on one set of data, is not adequate to prove true equivalence between products. A finding of true equivalence must be based on a statistical analysis that incorporates both the observed sample difference as well as the variability in the data. This ensures that the variability in the data does not mask a "true" difference that exceeds the predefined limits.

29. Put another way, rigorous statistics would require having a high degree of confidence that the "true" difference lies within the predefined limits. For this purpose, statisticians employ the concept of a confidence interval. A confidence interval can be viewed as the converse (or "flip side") of the p value. In the correct approach to equivalence testing, the null hypothesis of *inequivalence* is rejected if the measured difference falls within the predefined limits, and there is a low likelihood (p value) that a measurement in this range would have occurred if the "true" difference actually exceeded the predefined limits. A confidence interval

expresses the same concept conversely: It states a range of values, always centered around the difference actually measured in the test, that is expected to contain the “true difference” with X% confidence (typically 95%). In other words, it gives a range of values for the “true difference” that could have produced, with high probability, the observed difference. The idea is that, if repeat samples could be taken, an interval associated with 95% confidence would have only a 5% chance of not containing the true difference.

30. In the correct approach to bioequivalence testing, FDA requires that the coverage probability be 95%. In other words, even if the sample difference in the test lies within the predefined limits, FDA will not accept the result as proving equivalence (*i.e.*, as *disproving* the null hypothesis of *inequivalence*) unless statistical analysis shows that the range of values expected to contain the true difference with 95% confidence is completely contained within the pre-defined limits. This additional analysis of the confidence interval assessment is essential for proving equivalence. Without this component, a regulatory reviewer has no way of knowing if the observed results between products could have been produced from inequivalent products with large variability. With the confidence interval assessment, the reviewer knows that there is less than a 5% chance that these results would have been observed from truly inequivalent products (*i.e.*, the reviewer can be 95% confident that the products are indeed equivalent). Because of the nature of the null hypothesis, this confidence interval analysis is conducted based on the intersection of two one-sided 95% confidence bounds.

31. While some of the ANDA review documentation that I analyzed reports that the *observed* difference for an *in vitro* test fell within a predefined limit of acceptable difference, the nasal spray ANDAs did not report any confidence intervals for the measured parameters. If they had, and if a confidence interval analysis had been carried out, FDA's review decisions might have been very different.

32. For example, the review of ANDA 75-759, which shows p values both above and below 0.05 for certain parameters, reports observed ratios of the average measurement for test and reference products falling within a ratio of 0.90 to 1.11, but gives no indication of the relevant confidence interval. *See* Tab 2 at 11. The agency's willingness, within the "default equivalence" paradigm, to disregard measured average differences great enough to have p values below 0.05 cannot be justified on the basis that the measured average difference fell within predefined limits. The average difference can fall within the pre-defined limits, yet still have a confidence interval that does not fall within these limits, meaning that true product equivalence *cannot* be claimed. It is inconsistent with good statistical practice, and long-standing FDA practice, to rely on the fact that the observed average difference fell within predefined limits, without regard to whether the associated confidence interval, in its entirety, also fell within these limits.

33. Another example is the ipratropium bromide product in ANDA 75-156. Measured differences (expressed as ratios of the values for the test and reference

products) that fell within limits of 0.90 to 1.11 were deemed acceptable even though there is no indication in the review documentation that confidence intervals were developed or considered. *See* Tab 5 at 5-8. In fact, the p values shown in the review documents appear particularly poor; many of them are less than 0.05, indicating that even in the “default equivalence” paradigm, there was compelling evidence of *inequality*. The agency concluded, nonetheless, that “most” of the differences were insignificant, without further explanation. *See id.*

34. Given that use of confidence intervals is essential for proving equivalence, I endeavored, based on the summary statistics presented in tables, to derive the confidence intervals for several of the key *in vitro* tests carried out by nasal spray ANDA applicants. My purpose was to assess the extent to which the results would have “passed” if conventional and complete statistical criteria had been fully applied. For example, I used the reported raw data to calculate certain confidence intervals for ANDA 75-824. The outcome of my confidence interval analysis for this ANDA – some failing results as well as some passing – is representative of the outcomes of similar calculations I made for the other ANDAs.

35. Based on the data presented in Table 2 of the bioequivalence review for ANDA 75-824 (*see* Tab 1 at 7), the 90% confidence interval for the true ratio of means for content uniformity for beginning (sprays #8 and #9) and end (spray #23) fell entirely within the predefined limits of 0.90 to 1.11; thus supporting a conclusion of equivalence. On the other hand, I reached a conclusion of *not*

equivalent for spray pattern data for the Dmin and Ovality ratios presented in Table 4 of the bioequivalence review (*see id.* at 8), for the beginning stage at the distance of 5cm, based on the same limits of 0.9 to 1.11. The Dmin confidence interval extended from 1.10 to 1.21, while the ovality ratio confidence interval extended from 0.89 to 1.04, with corresponding p values of 0 and 0.22806. For example, for the Dmin measurement, this means that the true mean result from the generic product could be as much as 21% greater than the true mean result from the pioneer product. Also, in droplet size distribution testing, the beginning and middle stage data for “D50” would not have passed a confidence interval analysis, again based on limits of 0.90 to 1.11 (D50 represents a diameter size percentile, such that 50% of droplets in an emitted dose would have fallen beneath this threshold in diameter size). Even though these results failed to prove *in vitro* equivalence, the proposed generic product was determined to be bioequivalent.

36. Overall, in my confidence interval analyses, I found that most of the content uniformity data across the ANDA packages would have supported a conclusion of equivalence, but a good deal of data around spray pattern and droplet size distribution would not. These are among the key *in vitro* tests that FDA’s draft guidance has recommended to establish bioequivalence among intranasal products.

IV. CONCLUSIONS

37. In the absence of a final guidance document for locally acting nasal spray products, the agency has been applying statistical methods that lack the rigor characteristic of FDA's standard approach, and in any event are not clear or consistent from application to application.

38. FDA has applied a statistical analysis in which the generic product is presumed to be equivalent to the pioneer unless compelling data to the contrary are presented by the generic sponsor. As Dale Conner, Director of the Division of Bioequivalence in FDA's Office of Generic Drugs, explained in a recent public presentation on bioequivalence testing, "[w]hen you're unable to show that they're different, it doesn't mean that they're the same. It isn't a proof of sameness. It's simply that you failed to show they're different." See "The FDA Process for Approving Generic Drugs," Online Training Seminar (Dec. 2004) at www.fda.gov/cder/index.html. This is why FDA developed its standard approach to bioequivalence, which starts from the premise of *inequivalence*, and rejects this premise only in the face of statistically significant data (using a "confidence interval" analysis) to the contrary.

39. It is also troubling that even within the "default equivalence" approach, as it has been applied to recent ANDA approvals for generic nasal spray products, the agency appears to have granted approvals in the face of differences that have

statistically significant p values (less than 0.05), without additional justification.

The hallmark of good statistical science is to define criteria for success or failure in advance, and then apply them fully and faithfully to avoid subjectivity or post-hoc evaluation of the data.

40. Finally, my review supports the need for clarity in this area. The lack of statistical regularity in the approvals to date is troubling. To avoid continuing along this path, the agency and the industry would be well served were FDA to complete the guidance process it began more than five years ago, and publish a clear and validated methodology for establishing bioequivalence for locally acting nasal spray products. As GSK has repeatedly observed, this requires publication – and adequate opportunity for public comment – of complete proposals for satisfactory statistical methodologies and acceptance criteria.

Beth Morgan
Jane E ("Beth") Morgan

16-Jun-2005
Date