UNIVERSITY OF CALIFORNIA SAN FRANCISCO

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

STANTON A. GLANTZ, PhD
Professor of Medicine (Cardiology)
American Legacy Foundation Distinguished Professor of Tobacco Control
Director, Center for Tobacco Control Research and Education

530 Parnassus Suite 366
San Francisco, CA 94143-1390
Phone: (415) 476-3893
Fax: (415) 514-9345
glantz@medicine.ucsf.edu

**Comment Regarding Power and Effect Size in**
**Guidance for Industry and Food and Drug Administration Staff; Section 905(j) Reports: Demonstrating Substantial Equivalence for Tobacco Products; Guidance**
**FDA-2010-D-0635**

January 9, 2013

I have reviewed this Guidance Document and was surprised not to see any mention of statistical power or effect size, which are crucial issues in evaluating "negative" studies that are presented to show that some change to a tobacco product does not affect its risk profile.

A key question that underlies all statistical analysis of data is the question of how sensitive a study needs to convincingly accept the null hypothesis of no effect (i.e., substantial equivalence). In other words, how confident can one be in interpreting a "negative" finding in the sense of not finding a statistically significant difference before one can conclude "equivalence." This is a particular problem because most statistical hypothesis testing (which gives rise to P values and "statistical significance") implicitly assumes that the investigator *wants* to find a difference and so is focused on estimating the probability that a reported difference is a change random finding rather than a real effect (i.e., a false positive conclusion).

The issue at hand in conclusions of "substantial equivalence" is the other side of the coin, namely controlling the *risk of a false negative conclusion* (i.,e., concluding that there is no difference when one actually exists but is obscured by random noise in the data). This is what is called statistical *power* and is used to decide how big studies need to be. Since in most cases (such as when trying to demonstrate that a drug positively affects disease outcomes) people are hoping to find an effect and because obtaining high powers often requires very large sample sizes (the power goes up more-or-less in proportion to the square root of the sample size), the conventionally desired power is 80%, meaning that people are willing to accept a 20% risk of a false negative. By comparison, people usually are only willing to accept a 5% risk of a false positive for conventional statistical significance when they are seeking a positive finding.

Thus, if one were symmetrical in setting the standards for accepting a negative conclusion (i.e., no difference or substantial equivalence) as for positive conclusions (statistical significance), the FDA should require 95% power in any study claiming to demonstrate a negative (null) finding substantial equivalence.

Another complexity in thinking about this issue is that to compute power you need to specify *how big an effect is worth detecting*. In statistical significance testing you are always testing against the null hypothesis of *no effect*. For power, by definition, the effect size is not zero, but it is up to the investigator (or regulator) to decide how big an effect is *worth* detecting. This is another very important issue because for something like smoking which exposes millions of people to a potential toxin even small changes in risk (a few percent) would mean that a large number of people were affected.

Thus, someone -- presumably the FDA -- needs to specify how big a population level effect they considered "substantial." A 5% increase in risk? 1%? 10%? With smoking killing over 400,000 people a year a 1% risk increase would be about 4000 affected people.

Thus, in any determinant of "substantial equivalence" the applicants at the very least need to specify how large a risk increase they would deem worth detecting and submit data from studies that were big enough to adequate power of actually detecting that risk.

If, for example, the FDA were to require that studies cited in substantial equivalence claims have a 95% power to detect a 1% increase in risk, there would still be a 5% chance that there was a 1% increase in risk (or more) despite a study coming up "negative."

At the very least, the FDA should require that applicants clearly specify (1) the minimum detectable effect and (2) the power to detect that effect for *every* study submitted in an application for substantial equivalence.

The FDA, in acting on applications, should then explicitly accept or reject the proposed minimum detectable effect (which would the practical embodiment of how different the proposed product could be from the current product to be deemed “substantial equivalent”) and the power of the cited studies to actually detect that difference (which measures the risk of a false negative finding, i.e., accepting the conclusion of substantial equivalence when the new product is not actually substantially equivalent).

This is the central statistical issue in substantial equivalence and something that the FDA must address directly. As would be covered in any discussion of these issues in an introductory statistics course, simply reporting that differences between two products failed to reach “statistical significance” is not enough evidence to conclude that the products are the same.

Stanton A. Glantz, PhD
Professor of Medicine