Available online at www.sciencedirect.com
ScienceDirect
Regulatory Toxicology and Pharmacology 48 (2007) 51–58
Regulatory Toxicology and Pharmacology
www.elsevier.com/locate/yrtph

The simultaneous analysis of discrete and continuous outcomes in a dose–response study: Using desirability functions ☆

Todd Coffey [a,1], Chris Gennings [a,*], Virginia C. Moser [b]

[a] Department of Biostatistics, Virginia Commonwealth University, PO Box 980032, Richmond, VA 23298, USA
[b] Neurotoxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, RTP, NC 27711, USA

Received 10 October 2006
Available online 12 January 2007

Abstract

Multiple types of outcomes are sometimes measured on each animal in toxicology dose–response experiments. The potential false-positive rate from statistical tests on each endpoint may be inflated. We introduce a method of deriving a composite score that combines information from discrete and continuous outcomes through the use of desirability functions. These functions transform observed responses of any type to a 0-to-1 unitless scale. The geometric mean is used to combine the scores and then a statistical model is fit to the dose–response curve of the overall score. The overall desirability score is more sensitive to toxicity evident in only a few endpoints than other composite scores that are based on sums of components. We analyze the overall score using a nonlinear exponential model with a threshold parameter. In this example, the threshold parameter was statistically significant and its estimate was less than the lowest dose. Compared to the vehicle control, the lower overall scores at this dose group were due to lower levels of brain and blood cholinesterase (90% and 82% of control, respectively) whereas other endpoints were not altered, thus demonstrating the sensitivity of the desirability function to detect low levels of toxicity in a small number of outcomes.

Keywords: Composite score; Data analysis; Multiple outcomes; Nonlinear model; Scoring scheme

1. Introduction

Some types of toxicological evaluations include multiple outcomes on each experimental subject. The many endpoints are often used to make a comprehensive assessment of the effect of the chemical(s) under study. The measurement of multiple outcomes is common in studies to evaluate neurobehavioral toxicity, reproductive toxicity, clinical chemistry, organ weights, or pathology (e.g., Chapin et al., 1997; Crowell et al., 2004; Moser, 2000; Reed et al., 2004). Due to the large amount of data collected on multiple endpoints, the statistical analysis and interpretation are challenging. The multiple statistical tests that may be performed can greatly inflate Type I error rates, and multiple comparison adjustments are often overly conservative, e.g., Bonferroni correction (see Neter et al., 1996, p. 154), leading to reduced power to detect effects of interest. In addition, other characteristics of these experiments may provide additional complications to the analysis. For example, the endpoints may be a combination of binary, ordinal, count, and continuous data. In addition, the dose–response curves of different endpoints may be best described with nonlinear models or have different shapes or regions of activity (e.g., increasing and decreasing functions, different slopes or plateaus, etc.).

Several approaches to the analysis of such data have been proposed. Some authors (Dunson, 2003; Coffey and Gennings, 2007) have developed methodology to simultaneously analyze multiple types of outcomes by incorporating

☆ This manuscript has been subjected to review by the National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents necessarily reflect the views of the Agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

* Corresponding author. Fax: +1 804 828 8900.
E-mail address: gennings@vcu.edu (C. Gennings).
[1] Present address: Amylin Pharmaceuticals, Inc., San Diego, CA 92121, USA.

0273-2300/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.yrtph.2006.12.004

the correlation between outcomes. While these approaches control the Type I error rate and may result in improved precision, they require interpretation of each outcome and do not provide an overall estimate or interpretation of toxicity. In addition, the estimation of the correlation matrix with these methodologies may become unstable for a large number of outcomes. Another approach is to derive one overall score that uses information from each outcome. Moser (1991) developed a scheme for a composite severity score based on deviations from the concurrent control group. This approach has been implemented in the literature (McDaniel and Moser, 1993; Moser et al., 1995, 1997). The advantage of a composite score is one of dimension reduction: many responses are converted into a single score, which is then analyzed using standard statistical procedures. This method preserves the Type I error rate and results in an interpretation of the overall toxicity from the many outcomes, but questions remain as to the optimal approach for assignment of severity scores or for combing the scores into domains.

In this paper we describe the development of an overall score based on desirability functions for the many types of outcomes measured in toxicology experiments. Desirability functions were first proposed by Harrington (1965) for use in optimizing the quality of a manufactured product. Product quality is often measured by multiple endpoints, and engineers are faced with the challenge of finding levels of process factors that result in acceptable quality for all endpoints. Harrington's approach is used to find the levels of the factors that optimize the overall quality of the many endpoints. This approach has been widely adopted in the manufacturing industry and is regarded by engineers involved in product optimization as the most popular method for simultaneously analyzing many outcomes (Wu, 2005). Recently this methodology has also been applied to the titration of multiple drug regimens in medical research (Shih et al., 2003). To our knowledge, this methodology has not been applied to toxicology data.

The central idea of desirability scores is to create a function for each outcome that transforms the observed response to a unitless score (0-to-1) based on the appropriateness (or desirability) of the response. The individual scores are then combined into a single composite score through the geometric mean, and a standard statistical analysis can be performed. This flexible approach can handle the multiple types of response variables measured in toxicology experiments, can use different desirability functions for each outcome, and can incorporate weights to rank the importance of each endpoint. Of course, the desirability functions must be specified and so there is a degree of subjectivity involved in the choosing of the functions or weights. However, subjectivity in specifying weights or scoring schemes is a criticism of other composite scores, and its influence can be minimized by using consensus expert opinion.

We propose the use of desirability functions for toxicology research because of two properties that are potentially valuable. First, the composite score is calculated using the geometric mean—a function of the product of each of the individual desirability scores. Due to the mathematical nature of the product of scores between 0 and 1, response values assigned a low desirability score decrease the overall score more rapidly than other functions (such as the arithmetic mean). This important property is one of the main reasons the geometric mean was proposed (Harrington, 1965; Derringer and Suich, 1980). As Harrington (1965) explained, because the overall quality of a product measured from many outcomes is often based on the least desirable attribute, a method to produce a composite score should substantially penalize an unacceptable response for one outcome that otherwise has acceptable responses for other endpoints. Thus, desirability functions were designed to detect unacceptable responses in a small number of outcomes.

In the case of toxicology dose–response experiments, toxicity is often simultaneously manifested in many endpoints. However, there may be important cases when the toxicity manifests itself only in the most sensitive endpoint(s). The approach we introduce based on desirability functions is more adept at detecting this toxicity than other composite scores. The consequence of this sensitivity to toxicity in as little as one endpoint is that the estimate of the threshold level of toxicity will be affected and will likely be smaller than for other composite scores. This sensitivity in the threshold may be useful to risk assessors who are trying to estimate lower doses producing toxicity.

To illustrate the advantage of the geometric mean, consider a hypothetical scenario depicted in Fig. 1 in which ten outcomes are measured on a subject; eight of the responses are assigned a desirability score of 1, and two of the responses correspond to scores less than one (for simplicity, both take the same value). For incremental decreases of 0.1 in the two smaller scores, Fig. 1 demonstrates how the geometric mean decreases more rapidly than the arithmetic

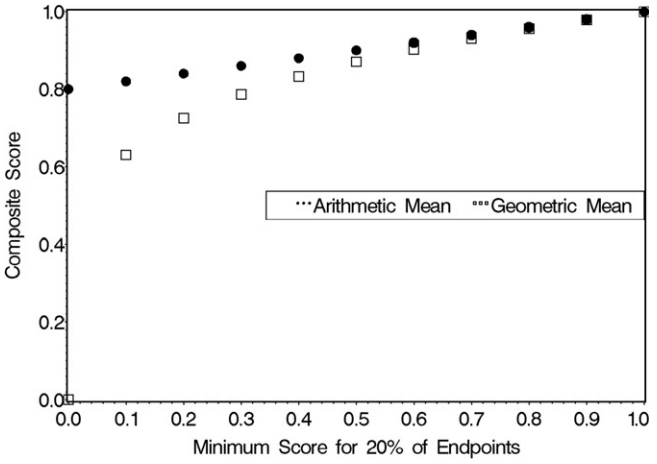

Fig. 1. Comparison of geometric and arithmetic mean for hypothetical scenario: 80% of outcomes have a score of 1.0, 20% have a score given by the horizontal axis.