

A Longitudinal Model and Graphic for Benefit-risk Analysis, With Case Study

Jonathan D. Norton, PhD
 Division of Biometrics II,
 Office of Biostatistics,
 Center for Drug Evaluation
 and Research, Food and
 Drug Administration,
 Silver Spring, Maryland

A novel method for simultaneously visualizing benefit and risk over time is presented. The underlying model represents a subject's benefit-risk state at a given time as one of five discrete clinical states, one being premature study withdrawal. The new graphic uses colors to represent each subject's changing state over the course of the clinical trial. The user can quickly

grasp how a treatment affects subjects in aggregate, then further examine how individuals are affected. It is possible to tell whether the beneficial and harmful outcomes are correlated. The method is particularly appropriate for treatments that provide only symptomatic relief. An approved drug for chronic pain is presented as a worked example.

Key Words

Benefit-risk; Graphics;
 Longitudinal; Safety;
 Missing data

Correspondence Address

Jonathan D. Norton,
 10903 New Hampshire
 Ave., WO 22, Rm. 3562,
 Silver Spring, MD, 20904
 (email: Jonathan.Norton@
 fda.hhs.gov).

The opinions and conclusions expressed in this article are solely the views of the author and do not necessarily reflect those of the Food and Drug Administration.

INTRODUCTION

Both FDA and external stakeholders acknowledge a need for improved benefit-risk (B-R) assessment (1). Unfortunately, there is little consensus about how best to approach this problem in a quantitative manner. I posit that the main reason for this lack of consensus is that the problem is an inherently difficult one. The core difficulty with B-R assessment is that the benefits and risks are typically measured on different scales. The intended benefit of an analgesic, for example, is to reduce a subject's pain level, often reported on a scale from 0 to 10. On the other hand, the potential harms from analgesics, like other drugs, come in many forms, as is apparent from a casual perusal of the package insert. This being the case, it is not clear how to come up with a single number that summarizes every patient's experience in a clinical trial. A second challenge to B-R assessment, and in fact evaluation of efficacy alone, is the prevalence of what is commonly (and not always correctly) referred to as missing data. A report by the National Research Council (2) describes the breadth and implications of this problem, which is also discussed more within. A third challenge to B-R assessment, one that I believe has been underrated, is that it is often appropriate to consider the conceptual B-R response in a trial to be longitudinal. This is particularly true for studies of treatments for chronic diseases. If the disease process is known to wax and wane, then the B-R

assessment from a trial should account for this fact.

Having acknowledged the difficulty of the general problem, this article focuses primarily on medical treatments that are intended to provide only symptomatic relief. This focus simplifies handling of subjects who withdraw prematurely from a trial and cease to provide data. While the resulting empty entries in the study database are commonly referred to as "missing data," it might be more appropriate to call them *counterfactual data*, as they are placeholders for numbers that may have meaning only under a state of affairs that did not actually occur, namely the subject staying on study medication for the duration of the trial. In the case of a drug that provides only symptomatic relief, one should arguably treat these cases as treatment failures. The logic for this classification is that if the patient actually felt better on the assigned treatment (possibly a placebo), then he or she would have in all likelihood stayed in the trial, barring a non-treatment-related obstacle to further participation.

An aside: My use of the term *counterfactual* may confuse readers who are accustomed to the term being used to refer to a state of affairs in which the subject was randomized to a different treatment, that is, in relation to the Neyman-Rubin (3) model of causal inference. That is a hypothetical situation that always has a coherent meaning, as one can always imagine the possibility of the same subject being random-

ized differently. Notably, Rubin (4) himself rejects the term counterfactual in that context. In contrast, when one ponders the hypothetical case of a subject who prematurely withdrew *instead* staying in the study, it becomes rather messy. Are we considering the same person hypothetically having a more favorable clinical response? Having the same response but being more stoic? Wanting to withdraw due to a poor response, but being forced to stay in the trial involuntarily (and illegally)? An advantage of treating withdrawal itself as an outcome is that such questions need not be pondered. See the National Research Council report (2) for a different viewpoint on these issues.

The author has developed a longitudinal benefit-risk model that is particularly appropriate for treatments that provide symptomatic relief for chronic conditions. It is based on a model introduced by Chuang-Stein et al. (5) that assigns one of five categorical outcomes to each patient at the end of a trial: benefit without an adverse event (AE; “serious side effects” in original reference), benefit with an AE, neither, AE without benefit, and early withdrawal due to unacceptable side effects. Chuang-Stein expanded the model (6), but the work in this article builds on the earlier publication (5). Due to the difficulty of ascertaining the true reason for withdrawal, I propose treating all premature withdrawals equally. I will simply refer to AEs, but it should be assumed that AEs of minor severity will not be counted. Which AEs are sufficiently serious or unpleasant to detract from efficacy is a judgment that must be made, ideally using a prespecified algorithm. The five outcomes are roughly ordered by decreasing favorability, but the under the original Chuang-Stein et al. model the actual relative value is determined by assigning a positive or negative weight to each outcome.

The utility of looking at the interaction between benefit and harm, rather than considering only their marginal distributions, may not be immediately apparent. Consider a hypothetical example from Thomas Permutt (personal communication): suppose that drug A has efficacy in all women and causes harm in all men,

whereas drug B causes both benefit and harm in all women and has no effect in men. Also assume that the harm is large enough to outweigh the benefit. While both drugs have the same marginal distributions of benefit and risk, only drug A has the potential to be useful. The same principle applies even if there is no easily ascertained predictive trait like sex.

My extension to the Chuang-Stein et al. model is to allow a patient’s benefit-risk category to change over the course of a study. For example, a patient could be in the benefit-only state for a week, then experience an AE in addition to the benefit, then have only the AE, then withdraw from the study. While I do not assume that the process is memoryless, it may be conceptually helpful to think of it as a Markov chain.

Different longitudinal patterns of benefit and harm could have different clinical implications. Suppose, for example, that many patients had unpleasant AEs in the early weeks of the trial that eventually resolved, leaving the patient in the efficacy-only state. This would suggest that a practitioner should encourage her patients to stay on treatment even if it was at first difficult to tolerate. Suppose, on the other hand, that early AEs were predictive of either premature withdrawal or a poor clinical state at the end of the study. This would suggest completely different advice. Recall that the assumption is that the drug provides only symptomatic relief; a drug that halts the progression of a serious illness, while perhaps making the patient feel worse, might be evaluated differently.

Ultimately, a model is as useful as the methods that are built on it. At present, I have developed a graphic called the Individual Response Profile that displays the longitudinal benefit-risk profile for each subject in the study. Inferential methods are briefly addressed in the discussion section.

METHOD

Table 1 shows the possible benefit-risk states, as adapted from Chuang-Stein et al. (5), with my suggested plotting colors when full color is available, as well as the tones used in this publication.

TABLE 1

Benefit-risk Categories, With Colors			
Category	Description	Full Color	Figure 1
1	Benefit without AE	Green	White
2	Benefit with AE	Yellow	Light Blue
3	Neither	Gray	Gray
4	AE without Benefit	Red	Dark Blue
5	Withdrew	Black	Black

Production of the graphic requires a listing of AEs from the trial, ideally with start and stop dates, and as well as dated efficacy responses. The graphic is produced as follows:

1. Define meaningful time windows relative to the beginning of the study. It is desirable to use time windows that were prospectively defined, for example, according to planned study visits.
2. Assign a binary efficacy response for each patient and time point, as long as the patient was in the study. A prespecified method should be selected for imputing intermittently missing efficacy data.
3. Determine which AEs detract from the efficacy response, ideally using a prespecified algorithm. Impute missing start and end dates as needed.
4. For each time window and subject, compute which of the five states the subject was in during that period.
5. Sort the subjects.
6. Plot the resulting subject-state matrix as a color image.

There is no particular limitation on how the subjects can be sorted. If the last period of the trial is of most interest, then the subjects can be sorted primarily according to their response on the last period, then by their response on the previous period, and so on. To give a simple example, suppose that a study had three periods. Subject A was in categories 3, 2, and 1 for the three respective periods; subject B was in categories 1, 3, and 1. They had the same response in the final period, but subject A had a better (lower-numbered) response in the second period. Hence subject A is ranked higher. The results from the first period are not used in

this case, because there is no tie left to break. This sorting order, which is used in the worked example in the next section, allows the user to see the history that preceded a given final outcome. Conversely, one could sort from the earliest period forward. Another possibility is to compute each subject's average response over time and use that as the primary sort key. However, I find that temporal sorting produces a cleaner-looking figure. If the number of subjects is relatively low, it may be helpful to include individual identifiers for each row.

One implementation decision that must be made is how to represent the results of the final period that the subject was on drug in the case of an early withdrawal. The question is whether to represent the period as state 5 (black) or to assign one of the other states based on the available efficacy and safety data. I have taken the latter approach because it provides the most information to the user of the graph. Note that the periods following early withdrawal are always represented as black.

Ana Szarfman and her coauthors at the Food and Drug Administration (7) developed graphics that are visually somewhat similar to what I propose here. The chief difference between this older work and the new method is that Szarfman et al. focused exclusively on safety, not efficacy. They gave this type of longitudinal safety graphic the nickname Napoleon's March, in honor of a well-known graphical map by Charles Minard showing Napoleon's disastrous invasion of Russia, as seen in Tufte (8). For consistency of terminology, I have adopted the name Sherman's March for the new graphic, referring to a

pivotal campaign in the US Civil War. The logic of this name is that while both marches had a sizable human cost, Sherman's March was also an effective military campaign. At the time that this article was being prepared, I learned that Ceesay and Entsuah (9) proposed a model with some similar elements at the Joint Statistical Meetings, about 6 weeks after I presented the model in this article at the International Chinese Statistical Association's Applied Statistics Symposium (10).

CASE STUDY AND RESULTS

Exalgo is an extended-release hydromorphone product that was approved in March 2010 for management of moderate to severe pain in opioid-tolerant patients requiring continuous, around-the-clock opioid analgesia for an extended period of time. I was the primary statistical reviewer for the New Drug Application, and chose this example primarily out of convenience. The reader should not infer that this product is either representative of or atypical among drugs in its class.

The sponsor conducted an adequate and well-controlled study of Exalgo in opioid-experienced subjects. After screening and enrollment, subjects entered a titration phase of 2–4 weeks. Subjects were randomized only after they were titrated to an effective and tolerable dose. Of the 459 subjects who entered the titration phase, 268 (58%) were randomized in equal numbers to either their titrated dosing or a matching placebo. The double-blind treatment phase lasted 12 weeks, and included a 2-week taper for those subjects who were randomized to placebo. The taper is needed to mitigate the effects of opioid withdrawal. The periods were defined following the planned visits in the original study, with the modification that the minimum period length was set to be a week.

Note: In the course of reviewing this application, FDA held an advisory committee meeting on the potential risk of abuse of this product. While one could argue that this hazard should be included in risk-benefit calculations, it has predominantly been assessed using observational data. Moreover, the at-risk population for

abuse extends beyond the patients receiving treatment. The scope of this article includes only AEs that are observed during a clinical trial.

For this analysis, a subject was deemed to have analgesic benefit during a given period if the average daily pain score was at least 30% lower than that reported at screening. The 30% criterion comes from John Farrar and coworkers (11), who determined that this percentage of reduction in pain constitutes a "clinically important improvement."

As noted earlier, a particular challenge of applying this method is to decide which AEs can be considered to counterbalance an efficacy response. A particularly strict approach would be to only count serious adverse events (SAEs), a category that is defined by regulation and includes events with outcomes such as death or disability. While this tight cutoff may be appropriate for some drugs, it seems unreasonable for an analgesic. A complaint of persistent nausea, for example, while not meeting any of the criteria for an SAE, could certainly be unpleasant enough to counteract a patient's improvement in well-being from a moderate decrease in pain. On the other hand, not all reported AEs will rise to the level of detracting from efficacy.

For the present study, the benefit-risk categories are based on AEs that the clinical investigator rated as moderate or severe. The study protocol defines a moderate event as one that "may be of sufficient severity to make patient uncomfortable; performance of daily activities may be influenced; intervention may be needed." While this definition is somewhat equivocal due to its use of the word *may*, it seems consistent with the type of event that would at least partially negate the benefit from an analgesic.

Figures 1A and 1B show the individual response profiles for subjects randomized to the hydromorphone and placebo arms, respectively. There were 134 subjects in each arm. The figures are sorted starting from the last period. The thickness of each row is proportional to the number of subjects who had that temporal profile. For example, starting from bottom of Figure 1A, 30 subjects were in the benefit-only state (white) for the entire period shown in the fig-

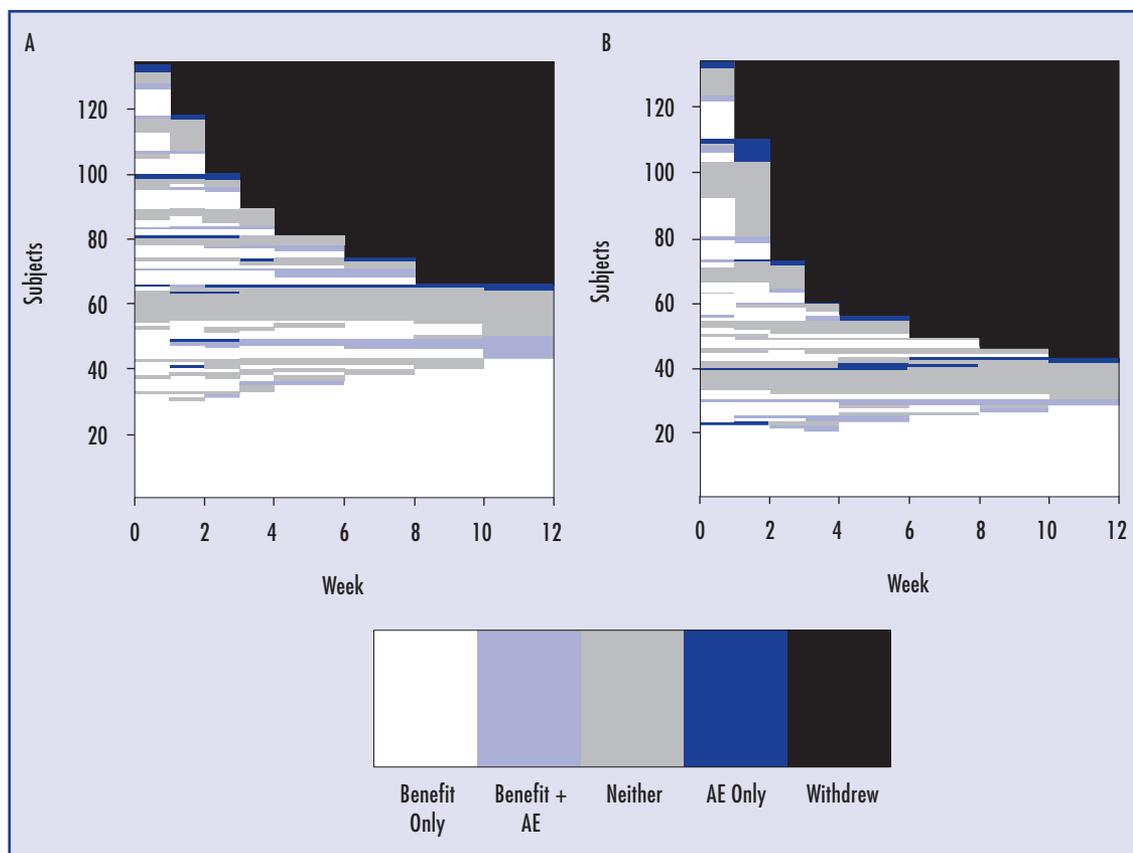


FIGURE 1

Individual response profiles for patients randomized to hydromorphone (A) or placebo (B). N = 134 in each arm.

ure. Going further up, there was one subject who was in the benefit-only state for week 1, then in the neutral (gray) state for week 2, and then in the benefit-only state for the remainder of the study.

Comparing the two figures, it is immediately apparent that the hydromorphone arm had more AE-free benefit (white) and fewer drop-outs (black) at the end of the study, that is, in the last vertical slice. (Chi-square tests confirm these impressions, with $P < 0.05$ in each case.) It also appears that a subject's outcome is largely determined by the end of week 8. In regard to the AEs (blue states), no particular pattern appears that would differentiate hydromorphone from placebo. However, this may simply reflect the fact that few moderate-to-severe AEs were actually observed.

DISCUSSION

In the original article, Chuang-Stein et al. proposed three summary measures of benefit-risk, or scores. These were functions of the sample

proportions falling in each of the five categories, with signed coefficients used to determine the relative weight of different categories. They developed a Wald-type test to compare treatments, using asymptotic variances computed via the delta method. In a personal communication, Dr Chuang-Stein suggested that a stochastic ordering test could be used in lieu of assigning explicit weights to each category.

Adding the dimension of time greatly complicates analysis. Temporal correlation is expected, and in fact guaranteed following an early withdrawal. One possible approach to deal with the multiple time points would be to compute a weighted sum of test statistics from individual time points. One choice of test statistic (to be computed for each time point) would be the Cochran-Mantel-Haenszel (CMH) test for a contingency table with an ordered classification (12). Since there is only one stratum at each time point, this would actually be a weighted sum of squared correlation coefficients. For a drug for a chronic condition, it would be ap-

appropriate to assign more weight to later time points, as they would be more relevant to the long-term efficacy of the drug. The null distribution of this weighted sum could be determined by resampling patient profiles. A downside to this test is that it is not specifically powered to detect an alternative in which the same treatment arm consistently has a better response over time. Another approach, which does not share this flaw, would be to compute a modified CMH test in which the strata are time points, and the strata are given prespecified weights according to their desired influence on the outcome. In contrast to the previously described test statistic, this one is a sum of directional effects and hence is sensitive to a consistent treatment effect. Note, however, that the usual distribution theory for the CMH test does not apply in this case of dependent observations, but a resampling approach could again be used to determine an appropriate cutoff.

In the course of showing figures like 1A and 1B to numerous people, I have received comments suggesting that the figure would look clearer or more orderly if there were more contiguous blocks of the same color. Having given these comments due consideration, I believe that the issue is largely one of perceptual psychology. Given the constraint that each row represents the temporal profile of a single subject, I believe that the sorting used in the figures make them as subjectively clean-looking as possible. However, there are certainly many possible ways to sort the rows, which emphasize different features of the data. Alternatively, if one is not interested in preserving individual subjects, then one can sort each time period (vertical slice) separately, showing the multinomial distributions. This was the approach used by Ceesay and Entsuaah (9).

One might question whether it is appropriate to treat all study withdrawals equally. As noted earlier, I believe that this approach is most appropriate for treatments that provide only symptomatic relief. The figure could certainly be modified to use different colors or shades to discriminate between types of withdrawal, using

categories such as those in Gould (13). The downside to this approach is that it increases the complexity of the figure, making it more difficult to compare treatments at a glance. Nevertheless, finer gradations would be appropriate if many subjects withdrew due to SAEs; one would certainly want to draw attention to on-study deaths, which did not occur in the present study. Conversely, in some settings a subject may withdraw from a study because the treatment provided a cure (or he experienced a spontaneous remission). If this could be accurately ascertained, and there was no irreversible harm from treatment, then some withdrawals could be counted as positive outcomes.

The graphic described here is intended to provide a limited view of the data. A comprehensive benefit-risk analysis of a drug, such as would be appropriate for regulatory purposes, should consider the similarity of the AEs observed in different subjects, whether they were reversible, whether they are likely to be attributable to treatment, and a host of other questions. Such questions are not addressed in the present study, because the drug in question is simply used for the purpose of exposition. An analysis intended to support an approval decision or to address a known safety concern would necessarily be much more thorough.

Along these lines, I take this opportunity to emphasize the importance of standard data formats and tools for benefit-risk analysis. Without adoption of such standards, an analysis of the type shown here will not become common practice; adapting the analysis code to the quirks of nonstandardized data sets will simply require too much labor. Even if one does not agree with the approach presented here, the need for more sophisticated and timely risk-benefit analysis is difficult to deny.

Acknowledgments—The author gratefully acknowledges the value of the clinical data submitted by Covidien Pharmaceuticals. He also thanks Thomas Permutt, Frank Pucino, Ana Szarfman, and two anonymous reviewers for their helpful comments.

REFERENCES

1. O'Neill RT. A perspective on characterizing benefits and risks derived from clinical trials: can we do more? *Drug Inf J*. 2008;42:235–245.
2. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press; 2010.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701.
4. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100:322–331.
5. Chuang-Stein C, Mohberg NR, Sinkula MS. Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Stat Med*. 1991;10:1349–1359.
6. Chuang-Stein C. A new proposal for benefit-less-risk analysis in clinical trials. *Control Clin Trials*. 1994;15:30–43.
7. Szarfman A, Talarico L, Levine JG. Analysis and risk assessment of hematological data from clinical trials. In IG Sipes, CA McQueen, AJ Gandolfi, eds. *Comprehensive Toxicology 4*. Amsterdam: Elsevier Science; 1997.
8. Tufte ER. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press; 1983.
9. Ceesay P, Entsuah R. Benefit risk assessment incorporating time component. Paper presented at Joint Statistical Meetings, August 4, 2010, Vancouver.
10. Norton JD. A longitudinal model for medical benefit-risk analysis, with case study. Paper presented at International Chinese Statistical Association 2010 Applied Statistics Symposium, June 21, 2010, Indianapolis.
11. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001;94:149–158.
12. Mantel N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc*. 1963;58:690–700.
13. Gould AL. A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics*. 1980;36(4):721–727.

The author reports no relevant relationships to disclose.

