AFFTC-PA-11318

Treatment Heterogeneity and Individual Qualitative Interaction

Robert S. Poulson
Gary L. Gadbury
David B. Allison

AIR FORCE FLIGHT TEST CENTER
EDWARDS AFB, CA

August 2011

AFFTC

AIR FORCE FLIGHT TEST CENTER
EDWARDS AIR FORCE BASE, CALIFORNIA
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
04-08-2011	Statistical Journal Publication	20-06-2011 to 04-08-2011

4. TITLE AND SUBTITLE

Treatment Heterogeneity and Individual Qualitative Interaction

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Robert S. Poulson, Gary L. Gadbury, David B. Allison

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)

812 TSS
Statistical Methods Group
307 E. Popson Ave.
Edwards AFB, CA 93524

8. PERFORMING ORGANIZATION REPORT NUMBER

AFFTC-PA-11318

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

812 TSS
Statistical Methods Group
307 E. Popson Ave.
Edwards AFB, CA 93524

10. SPONSOR/MONITOR'S ACRONYM(S)

N/A

11. SPONSOR/MONITOR'S REPORT NUMBER(S) N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release A: distribution is unlimited.

13. SUPPLEMENTARY NOTES

CA: Air Force Flight Test Center Edwards AFB CA CC: 012100

14. ABSTRACT

Plausibility of high variability in treatment effects across individuals has been recognized as an important consideration in clinical studies. Surprisingly, little attention has been given to evaluating this variability in design of clinical trials or analyses of resulting data. High variation in a treatment's efficacy or safety across individuals (referred to herein as treatment heterogeneity) may have important consequences because the optimal treatment choice for an individual may be different from that suggested by a study of average effects. We call this an individual qualitative interaction (IQI), borrowing terminology from earlier work - referring to a qualitative interaction (QI) being present when the optimal treatment varies across "groups" of individuals. At least three techniques have been proposed to investigate treatment heterogeneity: techniques to detect a QI, use of measures such as the density overlap of two outcome variables under different treatments, and use of cross-over designs to observe "individual effects." We elucidate underlying connections among them, their limitations and some assumptions that may be required. We do so under a potential outcomes framework that can add insights to results from usual data analyses and to study design features that improve the capability to more directly assess treatment heterogeneity.

15. SUBJECT TERMS

Causation; Crossover interaction; Individual effects; Potential outcomes; Probability of similar response; Subject-treatment interaction

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			412 TENG/EN (Tech Pubs)
Unclassified	Unclassified	Unclassified	None	30	19b. TELEPHONE NUMBER *(include area code)* 661-277-8615

Treatment Heterogeneity and Individual Qualitative Interaction

Robert S. Poulson
Statistical Methods Group
Edwards Air Force Base
Edwards, CA 93524
Robert.Poulson@edwards.af.mil

Gary L. Gadbury
Department of Statistics
Kansas State University
Manhattan, KS 66506
gadbury@ksu.edu

David B. Allison
Department of Biostatistics, Section on Statistical Genetics
University of Alabama at Birmingham
Birmingham, AL 35294
dallison@uab.edu

Author's footnote:

Robert Poulson is a PhD mathematical statistician in the statistical methods group at Edwards AFB, Edwards, CA 93524 (email: robert.poulson@edwards.af.mil). Gary Gadbury is Associate Professor in the Department of Statistics at Kansas State University, Manhattan, KS 66506 (email: gadbury@ksu.edu). David Allison is Professor and Head of the Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294 (email: dallison@uab.edu). Dr. Allison acknowledges research support from NIH R01DK078826. The authors are grateful to Kraft Foods for supplying the data for this paper.

Treatment Heterogeneity and Individual Qualitative Interaction

ABSTRACT

Plausibility of high variability in treatment effects across individuals has been recognized as an important consideration in clinical studies. Surprisingly, little attention has been given to evaluating this variability in design of clinical trials or analyses of resulting data. High variation in a treatment's efficacy or safety across individuals (referred to herein as treatment heterogeneity) may have important consequences because the optimal treatment choice for an individual may be different from that suggested by a study of average effects. We call this an individual qualitative interaction (IQI), borrowing terminology from earlier work - referring to a qualitative interaction (QI) being present when the optimal treatment varies across "groups" of individuals. At least three techniques have been proposed to investigate treatment heterogeneity: techniques to detect a QI, use of measures such as the density overlap of two outcome variables under different treatments, and use of cross-over designs to observe "individual effects." We elucidate underlying connections among them, their limitations and some assumptions that may be required. We do so under a potential outcomes framework that can add insights to results from usual data analyses and to study design features that improve the capability to more directly assess treatment heterogeneity.

KEY WORDS: Causation; Crossover interaction; Individual effects; Potential outcomes; Probability of similar response; Subject-treatment interaction.

1. INTRODUCTION

"*..,it appears that white sheep and pigs are injured by certain plants, whilst dark- coloured individuals escape*." ~ Charles Darwin

"*What is food to one to some becomes Fierce poison*" ~ Lucretius

The quotations above illustrate that individual differences in response to stimuli or 'treatments' have been the subject of interest throughout recorded history. They further illustrate two kinds of interactions. Darwin points out an interaction in which one type of animal is harmed

by a certain treatment whereas other animals are not harmed, but are not necessarily helped. In contrast, Lucretius points out a more dramatic type of interaction in which what is helpful to some is actually harmful to others. More formally, treatment heterogeneity is present when the effect of a treatment, say T, with respect to a reference treatment, R, varies across subsets or individuals in a population. At the individual level, this variability is called subject-treatment interaction (Gadbury 2010). A consequence of this heterogeneity is that different individuals or subsets may respond to treatment in opposite directions, with treatment T having higher efficacy for some and treatment R having higher efficacy for others. The term qualitative interaction (QI) has been used to describe this situation at the subset level (Peto 1982; Gail and Simon 1985), and tests have been developed to detect a QI (Gail and Simon 1985; Silvapulle 2001; Li and Chan 2006). When such tests are significant, optimal treatments may differ among subsets (Byar and Corle 1977). A "quantitative" interaction (Peto 1982) exists when the magnitudes of the effects of a treatment differ across subsets, but are in the same direction. Herein we refer to a "subset interaction" as a more general term that includes quantitative and/or qualitative interactions.

Taking the idea of subsets to its limit, we can recognize that each person is unique and can be considered their own subset. Then analogous to the QI described above, an individual qualitative interaction (IQI) is present when at least two individuals respond in the opposite direction to treatment. A fact that initially seems counter-intuitive to many clinical investigators who are used to discussing 'non-responders' in standard clinical trials (e.g., Inoue et al. 2010), is that individual effects of a treatment T with respect to R are inherently unobservable in a two treatment comparison study because only one of the two outcome variables is observable for each subject, depending on the treatment assigned to that subject. Let potential outcome variables (Rubin, 1974) to treatments T and R be given by X and Y, respectively, with an individual effect defined by the variable $D = X - Y$. Suppose, as in Gadbury and Iyer (2000),

that the potential outcome variables, $(X, Y)^T$ are modeled by a bivariate population distribution with mean $(\mu_X, \mu_Y)^T$ and covariance matrix $\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$, where the covariance $\sigma_{XY} = \sigma_X \sigma_Y \rho_{XY}$.

The distribution of *D* then has mean $\mu_D = \mu_X - \mu_Y$ and variance

$$\sigma_D^2 = Var(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_X \sigma_Y \rho_{XY}. \quad (1)$$

In a two treatment comparison study where subjects are randomly assigned to treatments T and R, the mean treatment effect, μ_D, can be estimated but the variance, σ_D^2, cannot because there is no information in observable data to estimate the correlation, ρ_{XY}.

Assume throughout that $\mu_D = E(X - Y) = \mu_X - \mu_Y > \tau$ represents a beneficial average effect of T over R where τ is some threshold particular to the treatments being compared (that is, treatment T may have costs associated with it over treatment R that require a sufficiently positive value for μ_D before it is claimed that T is a preferred treatment over R). Hereafter, for convenience, we take $\tau = 0$. Subject-treatment interaction is present in the population when $\sigma_D^2 > 0$. If there is no interaction at the individual level, $\sigma_D^2 = 0$, and there is a constant treatment effect (Holland 1986). However, as individual treatment effects become more heterogeneous, σ_D^2 gets larger, and a positive proportion of individuals in the population with a value of *D* less than zero (i.e., an IQI is present) becomes more plausible, despite $\mu_D > 0$. We denote a proportion of individuals having an unfavorable outcome to treatment T versus R as PIQI. If the bivariate distribution given above is normal, then

$$\text{PIQI} = \Phi\left(\frac{-\mu_D}{\sigma_D}\right), \quad (2)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). Normality is assumed here for convenience, but it is not required for definition of the PIQI.

It has been remarked that medicine today generally makes use of statistical information gathered about the general population (often about the "average" subject) and then applies it to the individual (cf. Marshall, 1997). Some have suggested that information about a mean treatment effect be supplemented by information about treatment effect heterogeneity (cf., Longford, 1999). This paper explores methods to determine whether σ_D^2 and/or PIQI are positive using observed data from a two treatment comparison study where treatments are randomly assigned. In particular, we discuss three approaches: tests for subset interaction, the proportion of similar responses (PSR), and cross-over designs. We review each method using the potential outcomes structure to highlight important connections and assumptions. A data example is used to illustrate the ideas. Though the focus here is a clinical setting, interest in other aspects of a treatment effect distribution besides the mean has emerged from other fields. For example, see Fan and Park, 2009, 2010, for application in econometrics.

It is recommended that new ideas for clinical trial designs and methodologies be pursued that may lead to further improvements in our ability to estimate and test aspects of individual treatment response heterogeneity. We offer potential outcomes as a useful framework for understanding individual treatment heterogeneity and its consequences. It helps to distinguish heterogeneity that is "explainable" in observed data from unexplained heterogeneity. Though subject-treatment interaction and the PIQI cannot be directly estimated in observed data without introducing additional assumptions, bounds can be estimated. Consequences of unexplained heterogeneity, reflected by estimable bounds for the PIQI, can alert investigators to the possible existence of an unobserved covariate that could be potentially predictive of individual success to a treatment application.

2. AN ILLUSTRATIVE EXAMPLE

Bookstores typically devote shelf space to a wide variety of dieting books, each book frequently containing anecdotes describing substantial weight loss and other remarkable improvements to health for particular individuals. Obesity researchers are cautious about embracing various new diets due to limited evidence that they outperform more traditional programs of weight loss. Two diets whose relative merits have been discussed are the low carbohydrate diet (sometimes called a reduced-glycemic-load diet) versus a more traditional portion controlled low-fat diet. Results from clinical studies comparing these two have been inconsistent, with some suggesting that one appears more effective for weight loss at some time points, on average, and others that show no significant difference. A March 4, 2010 article in the *Wall Street Journal* reported on an unpublished study by Stanford University researchers that suggested that individual differences in genetic predispositions contribute to substantial individual differences in the relative efficacy of one diet versus another (i.e., low fat versus low-carbohydrate) among overweight women.

Given the possibility of treatment heterogeneity in a such a study as well as an IQI, we illustrate the ideas discussed here using a data set that compared a reduced-glycemic-load diet (RGL) and a portion controlled low fat diet. The data are a subset of data analyzed and reported in Maki et al., 2007. Subjects were randomized to two treatments: T = the RGL diet (n = 43) and R = the low fat diet (n = 43). A primary outcome variable was weight change from baseline at 12 weeks, measured in kilograms. Maki et al., 2007, also report analyses of other outcome variables such as waste circumference, fat free mass, and results from laboratory tests. Several covariates were measured such as baseline values of outcome variables, age, race, and gender. Using notation described earlier, we consider the outcomes X = weight change from baseline at 12 weeks for subjects assigned to treatment T, and Y = weight change from baseline at 12 weeks for

subjects assigned to R. Positive values of *X* and *Y* are a weight "loss" (in kilograms) from baseline. We analyze data for the 69 subjects (out of 86) that remained in the study at 12 weeks (34 in treatment T and 35 in treatment R). For brevity and to focus on the topic presented herein, we do not consider issues related to compliance or drop out and initially, analyze data for only the two outcomes, *X* and *Y*.

The treatment T group had a mean weight loss of $\hat{\mu}_X = 5.39\,\text{kg}$ with a standard deviation of $\hat{\sigma}_X = 3.16\,\text{kg}$. The treatment R group had a mean weight loss of $\hat{\mu}_Y = 3.23\,\text{kg}$ with a standard deviation of $\hat{\sigma}_Y = 3.20\,\text{kg}$. An estimate of mean treatment effect is $\hat{\mu}_D = 2.16\,\text{kg}$, and a t-test of $Ho: \mu_D = 0$ against an upper tailed alternative hypothesis gives a p-value of 0.003.

Sometimes investigators will interpret unequal variance of the outcome variables in each treatment group as evidence of treatment heterogeneity. This is, in fact, partially true because the minimum bound for the subject-treatment interaction standard deviation is $\sigma_D^{min} = |\sigma_X - \sigma_Y|$, and this quantity can be large when σ_X and σ_Y are very different. Estimable bounds, $\left(\sigma_D^{\min}, \sigma_D^{\max}\right)$ are obtained by setting the correlation equal to 1 and -1, respectively (Gadbury and Iyer, 2000). The maximum bound, $\sigma_D^{max} = \sigma_X + \sigma_Y$, is not small unless both standard deviations of the potential outcome variables are small. From the example data, the estimated bounds are $\hat{\sigma}_D^{\min} = 0.04$, $\hat{\sigma}_D^{\max} = 6.36$. The estimated standard errors obtained by 2000 bootstrap samples within treatment groups (cf., Gadbury et al., 2001) indicate that the lower bound is not different from zero. The estimated standard error for $\hat{\sigma}_D^{\max}$ is 0.59. Based on only these outcome variables and these estimated bounds, there is no clear and compelling evidence in the data that subject-treatment interaction 'must' be present, but there is evidence that it 'could' be.

Assuming as before a bivariate normal distribution for weight loss outcomes *X*,*Y*, estimated bounds for the PIQI (Gadbury and Iyer, 2000) are given as,

$$PIQI^{\min} = \Phi\left(\frac{-\hat{\mu}_D}{\hat{\sigma}_D^{\min}}\right) \approx 0, \qquad PIQI^{\max} = \Phi\left(\frac{-\hat{\mu}_D}{\hat{\sigma}_D^{\max}}\right) = 0.37. \tag{3}$$

The bootstrap standard error for the upper bound is 0.049. These results suggest that the estimated proportion of the population having an effect of treatment T versus R in the opposite direction from the mean effect could be negligible or as high as 0.37. One may argue that it is more plausible that this proportion is closer to zero rather than 0.37 because the nonestimable correlation between potential outcomes should be closer to 1 rather than -1. Yet without additional information or assumptions, little more can be said about treatment heterogeneity or its consequences based on these data alone.

3. METHODS FOR ASSESSING TREATMENT HETEROGENEITY AND ITS CONSEQUENCES

Other techniques have been developed for studying population treatment heterogeneity and its consequences under different assumptions and constraints. For example, Hauck et al. (2000), in a slight variation of our current notation (X and Y are considered individual averages in a repeated measures cross-over design), proposed σ_D^2 be used to determine whether T and R were individually bioequivalent. In this section we use potential outcomes to clarify the assumptions required to estimate σ_D^2 and a PIQI under three different strategies. First we establish a connection between subset interaction and subject-treatment interaction and show how the former, with an appropriate design, is a detectable consequence of the latter. Second, we show that the proportion of similar response (PSR) or density overlap, though intuitively appealing, can be misleading when used as a proxy for treatment heterogeneity and, hence, the potential presence of IQI. Finally, we show that additional information becomes available in cross-over designs, but that direct estimation of the PIQI requires further assumptions.

3.1 Identification of Subsets

The study of subset interaction presupposes a covariate that is a "grouping variable" and some degree of homogeneity of treatment response within groups, with QI then explained by differences in treatment response across groups. If the grouping variable is continuous, then groups are subpopulations defined by values of the covariate. One reason for subset analysis then is to identify "which treatment is best for which kinds of patients," (Byar and Corle 1997, p. 455). Standard methods seek to find such subsets through an investigation of interaction effects (Byar and Corle 1977; Simon 1982) or a direct test for a qualitative interaction (Gail and Simon 1985; Silvapulle 2001; Li and Chan 2006). In each case the interaction is detectable by changes in the mean response across subsets. Using potential outcomes, the subject-treatment interaction variance σ_D^2, can be decomposed into an explainable component (i.e., a component that is estimable) and an unexplainable component (remaining subject-treatment interaction within a subset).

3.1.1 A continuous covariate

First consider, as in Gadbury et al., (2001), a continuous covariate Z (i.e., not affected by the treatment) that augments potential outcomes (X, Y). Assume the distribution of D given $Z = z_0$ is normal with conditional mean

$$\mu_{D|Z=z_0} = \mu_Y - \mu_X + (\beta_{XZ} - \beta_{YZ})(z_0 - \mu_Z) \tag{4}$$

and conditional variance,

$$\sigma_{D|Z}^2 = \sigma_{X|Z}^2 + \sigma_{Y|Z}^2 - 2\sigma_{X|Z}\sigma_{Y|Z}\rho_{XY|Z}. \tag{5}$$

β_{XZ} and β_{YZ} in equation (4) are the slope coefficients between Z and X and Z and Y, respectively, and $\rho_{XY|Z}$ in equation (5) is the partial correlation between X and Y, given Z. The conditional

variances, $\sigma^2_{X|Z}$ and $\sigma^2_{Y|Z}$, are allowed to be different across the two treatment groups but are assumed to not depend on the value of Z. Gadbury et al. (2001) showed that,

$$\sigma_D^2 = \left(\sigma_{X|Z} - \sigma_{Y|Z}\right)^2 + 2\sigma_{X|Z}\sigma_{Y|Z}\left(1 - \rho_{XY|Z}\right) + (\beta_{XZ} - \beta_{YZ})^2\sigma_Z^2$$
$$= \sigma_{D|Z}^2 + (\beta_{XZ} - \beta_{YZ})^2\sigma_Z^2. \qquad (6)$$

So σ_D^2 is comprised of two components, one that can be attributed to subset interaction (the second term in (6)) and one that can be attributed to subject-treatment interaction, within subsets (the first term in (6)). The quantity, $(\beta_{XZ} - \beta_{YZ})^2\sigma_Z^2$, can be estimated using the observed data. When $\beta_{XZ} \neq \beta_{YZ}$, then $\sigma^2_{D|Z}$ within subpopulations is smaller than the unconditional subject-treatment interaction variance, σ_D^2. The conditional proportion of IQI within the subset (or subpopulation) defined by a value of the covariate *Z*, $\text{PIQI}_{Z=z_0}$, is given by a quantity analogous to that given in equation (2), except using the conditional mean and conditional variance of *D*, given *Z*. Since, $\mu_{D|Z=z_0}$ could be greater than μ_D when $\beta_{XZ} \neq \beta_{YZ}$, and both $\sigma^2_{X|Z} \leq \sigma^2_X$ and $\sigma^2_{Y|Z} \leq \sigma^2_Y$, one could identify subsets of the population for which $\text{PIQI}^{max}_{Z=z_0} < \text{PIQI}^{max}$.

Returning to the data example, let *Z* = baseline weight. A test for a baseline-treatment interaction is significant (p-value = 0.007), $\hat{\beta}_{XZ} = 0.082$, $\hat{\beta}_{YZ} = -0.075$,, and $\hat{\sigma}_Z^2 = 165.9$ so that an estimated $\left|\hat{\beta}_{XZ} - \hat{\beta}_{YZ}\right|\hat{\sigma}_Z = 2.02$ kg of σ_D is explained by the baseline weight covariate. Figure 1 is a plot of the data showing the interaction between treatment and baseline weight. The vertical line is plotted at $\hat{\mu}_Z = \bar{z} = 89.6$ kg, where $\bar{z}$ is the mean of all 69 baseline weights. Estimable bounds for the remaining unexplained subject-treatment interaction standard deviation within subpopulations $\sigma_{D|Z}$ can be bounded by quantities $\left(\sigma^{\min}_{D|Z}, \sigma^{\max}_{D|Z}\right)$ that are estimable by setting the nonestimable partial correlation in (6) equal to 1 and -1 respectively, and estimating the

conditional variance of *X* and *Y*, given *Z*, using the mean squared error of models that regress *X* on *Z* and *Y* on *Z*. This gives estimated bounds equal to 0.1 and 6.1.

If the distribution of *D* at a given value of *Z* is normal with mean and variance given by (4) and (5), then bounds for the PIQI at a given $Z = z_0$ are given as

$$\text{PIQI}_{Z=z_0}^{min} = \Phi\left(\frac{-\mu_{D_{z_0}}}{|\sigma_{X|Z} - \sigma_{Y|Z}|}\right), \text{and } \Phi\left(\frac{-\mu_{D_{z_o}}}{\sigma_{X|Z} + \sigma_{Y|Z}}\right) = \text{PIQI}_{Z=z_0}^{max}. \quad (7)$$

The quantities in (7) can be estimated from the regression of *X* on *Z* and of *Y* on *Z*.

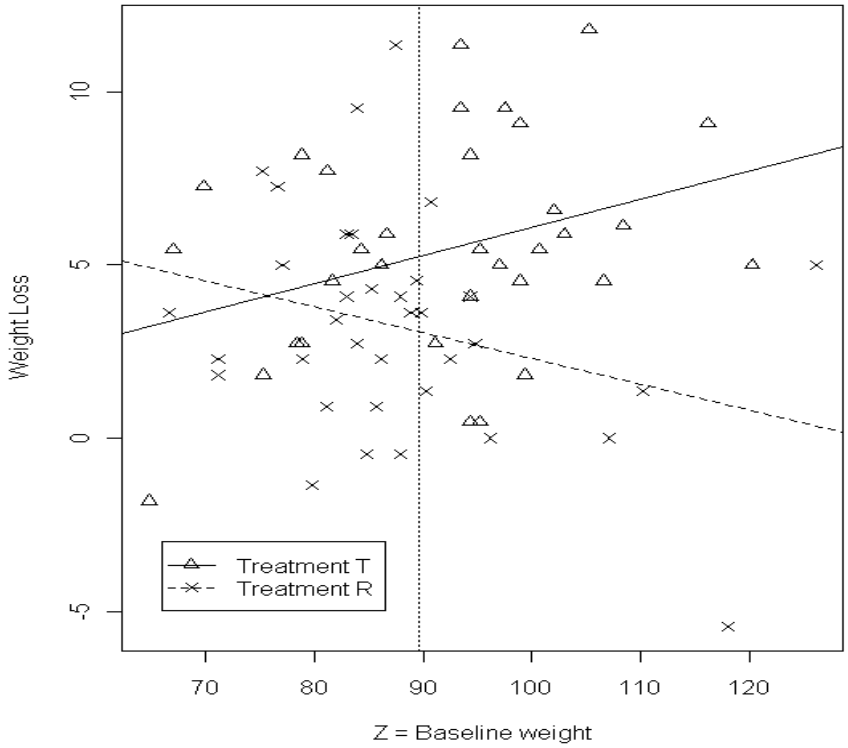

Figure 1: Plot of weight loss in kilograms (kg) at 12 weeks (positive values are a weight loss) versus baseline weight in kg. The fitted lines are from regressing *X* on *Z* and *Y* on *Z*. The vertical dashed line is at the sample mean baseline weight.

Table 1 shows the estimated mean treatment effect for 3 values of baseline weight, the mean baseline weight and one and two standard deviation(s) above baseline weight. The standard deviation of baseline weight, s_z, was computed from all 69 baseline weights. Estimates of the

two conditional standard deviations in (7) are nearly the same, so that the estimated minimum PIQI is not different from zero. The estimated maximum is shown in Table 1 along with standard errors obtained from 2000 bootstrap samples within treatment groups.

Table 1: Estimated mean treatment effect and maximum PIQI (and standard error) at three different values of baseline weight.

z_0	$\mu_{D\|Z=z_0}$ (se*)	$PIQI^{\max}_{D\|Z=z_0}$ (se*)
89.6[1]	2.16 (0.780)	0.363 (0.052)
102.5[2]	4.16 (1.256)	0.248 (0.067)
115.4[3]	6.17 (2.021)	0.156 (0.078)

Footnotes: *The standard error estimates (se) are based on 2000 bootstrap samples.
1. $z_0 = \bar{z}$ 2. $z_0 = \bar{z} + s_z$ 3. $z_0 = \bar{z} + 2s_z$

Some conclusions can be summarized from the analysis of these data using baseline weight as a covariate.

1. Data suggest some evidence of subject – treatment interaction in the population due to a significant treatment-covariate interaction.
2. Sometimes transformations are sought to remove interactions so that, on the transformed scale, more comprehensive statements about treatment effects can be made. However, if measurements are obtained on a clinically meaningful scale and subject-treatment interaction is present in the population, it cannot be removed by transformations.
3. Interactions like that shown above can highlight subpopulations that may respond differently to a treatment. Although the estimated lower bound for PIQI at the three values of baseline weight in Table 1 is not different from zero, the estimated upper bound can be quite large. However, this estimated upper bound decreases for larger values of baseline weight and, at two standard deviations above baseline, the estimate

of the maximum PIQI is only two standard errors from zero. This suggests, based on these data, that most individuals with a larger baseline weight should benefit from the RGL treatment versus a more traditional low-fat diet – at least when assessed over a 12 week period. It is less clear whether the RGL diet would perform better for 12 week weight loss than the low-fat diet for individuals with a baseline weight below the average weight of 89.6 kg.

4. Additional covariates that are predictive of weight loss, regardless of whether they interact with the treatment, can tighten the estimated bounds for PIQI by reducing the estimated conditional variances of the outcome variables, given the set of covariates.

3.1.2. A categorical covariate

Analogous results to those described above for a continuous covariate can be derived for a categorical one as well. In particular, the subject-treatment interaction variance decomposes into a covariate-treatment interaction term (the explainable component) and a within group variance (an unexplainable component).

A slightly different approach from that above helps facilitate the derivation. Suppose *Z* is a categorical covariate with g levels. A balanced design is considered here, so there are n units per group for a total of ng experimental units. Assume as before that a bivariate set of potential outcomes are randomly generated from a population model, and denote the set of potential outcomes as (X_{ij}, Y_{ij}), $i = 1, 2, \dots, g, j = 1,2, \dots, n$. From the set of potential outcomes, $D_{ij} = X_{ij} - Y_{ij}$ is a individual treatment effect, $\bar{D}_{z_i} = \frac{1}{n}\sum_{j=1}^{n} D_{ij} = \bar{X}_{i\square} - \bar{Y}_{i\square}$ is a mean treatment effect within the *i*th level of *z*, where $\bar{X}_{i\square} = (1/n)\sum_j X_{ij}$ and $\bar{Y}_{i\square} = (1/n)\sum_j Y_{ij}$. Define the variance of these individual effects as,

$$S_D^2 = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\left(D_{ij} - \bar{D}\right)^2}{ng-1} \tag{8}$$

where $\bar{D} = \sum_i \sum_j D_{ij} \Big/ ng$. The quantity in (8) can be thought of as a finite population version of σ_D^2 from the prior section. If $S_D^2 > 0$, then there is subject-treatment interaction present among the *ng* subjects in the sample. It can be shown that,

$$\frac{ng-1}{g(n-1)} S_D^2 = S_{D|Z}^2 + \frac{n\sum_{i=1}^{g}\left(\bar{D}_{z_i} - \bar{D}\right)^2}{g(n-1)}, \tag{9}$$

where $S_{D|Z}^2 = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n}\left(D_{ij} - \bar{D}_{z_i}\right)}{g(n-1)}$ represents a within group variance of individual treatment effects. Equation (9) shows that the components of S_D^2 include both subject-treatment interaction within subsets specified by $S_{D|Z}^2$ and subset treatment interaction term that is a function of $\sum_{i=1}^{n}\left(\bar{D}_{z_i} - \bar{D}\right)^2$. When $\sum_{i=1}^{n}\left(\bar{D}_{z_i} - \bar{D}\right)^2 > 0$, the mean treatment effect within subsets varies across the subsets. In the extreme case that $S_{D|Z}^2 = 0$, subject-treatment interaction in the set of *ng* subjects is completely explained by the interaction across subsets, which indicates a constant individual effect of treatment T relative to treatment R within subsets. In the other extreme that $S_{D|Z}^2 = S_D^2$, then *Z* is not useful for predicting subsets of individuals (among the *ng* individuals) who may respond successfully to one treatment over the other.

None of the quantities in equation (9) can be calculated from actual observed data post treatment assignment, because all potential outcomes are not observable. However, a post treatment assignment "estimate" for the second term in (9), $\frac{n\sum_{i=1}^{g}\left(\bar{D}_{z_i} - \bar{D}\right)^2}{g(n-1)}$, is

a scalar of the usual sum of squares for the subset-treatment interaction term in a $2 \times g$ factorial analysis of variance computation with $\frac{n}{2}$ observations for each treatment group combination. Consequently, in an ANOVA model with weight loss as a response and treatment, *Z*, and a treatment-*Z* interaction as explanatory variables, an F-test for the contribution of the interaction term may not only be used to diagnose the degree of subset treatment interaction, but also provides evidence that $S^2_{D|Z} < S^2_D$, and hence, $\sigma^2_{D|Z} < \sigma^2_D$.

$S^2_{D|Z}$, the first term in equation (9), may be used to evaluate the PIQI within groups, as before. If $D_{ij} \sim N\left(\mu_{D_{z_i}}, \sigma^2_{D|Z}\right), j = 1, \ldots, n$, then bounds for the PIQI at a given $Z = z_0$ are the same as those given in equation (7), and estimates for the parameters in these bounds, $\mu_{D_{z_o}}$, $\sigma^2_{X|Z}$, and $\sigma^2_{Y|Z}$, can be estimated from sample statistics. One could estimate $\sigma_{X|Z}$ and $\sigma_{Y|Z}$ by pooling sample variances of observed outcomes across groups or separately within each group. The latter approach is equivalent to conducting a separate analysis within each subset. It is possible that the bounds for PIQI_{z_i} vary widely across subsets, with some subsets exhibiting the plausibility of more treatment heterogeneity than others. Subsets with a positive estimated lower bound for the subject – treatment interaction variance (and/or PIQI) , or a small estimated upper bound(s) may be particularly informative.

The categorical variable available in the illustrative data set is gender. A test for a gender-treatment interaction was not significant, indicating that the effect of the RGL diet with respect to the low fat diet was not estimated to be different across genders, or that gender does not explain any treatment heterogeneity. There is another technique that has been proposed to evaluate the potential presence of treatment heterogeneity. This is the proportion of similar response (Rom and Hwang 1996; Stine and Heyse 2001), discussed next.

3.2 Proportion of Similar Response

Inman and Bradley (1989) provide a comprehensive treatment of the PSR where its calculation is defined as a measurement of overlap between two probability density functions (pdfs), given as

$$\text{PSR} = \int min\big(f_X(x), f_Y(x)\big)\, dx, \tag{10}$$

where $f_X(x)$ and $f_Y(x)$ are the pdfs of the outcome variables X and Y to treatments T and R, respectively. There has been some confusion regarding the interpretation of the PSR (Inman and Bradley 1989) and particular disagreement over its use as a measurement of treatment heterogeneity (Gastwirth 1975; Senn 1997, 2006b). The overlap of the density curves that leads to the PSR calculation provides a natural way to think about treatment heterogeneity and IQI. In support of such usage, Gastwirth (1975) noted that the maximum values for the PSR and the $P(X < Y) = P(D < 0)$ which are 1 and 0.5, respectively, have a similar interpretation. Additionally, the overlap seems to suggest that as the PSR increases, the potential for a value from $f_X(x)$ to be less than a value from $f_Y(x)$ also increases. However, in an assessment of the PSR, Senn (2006b, pp. 3944-3945) points out, "If every patient benefits by having his or her outcome improved by the same amount [under treatment T] compared to what it would have been [under treatment R], then 100 percent of the patients have benefited" (brackets added to provide context for the notation herein). Thus Senn identifies what is clear using potential outcomes, which is, if D is a constant, $\sigma_D^2 = 0$ and the PIQI $= 0$, even when the PSR > 0.

The calculation of the PSR depends on values of x such that $f_X(x) = f_Y(x)$. If the two distributions are equal, $f_X(x) = f_Y(x)$ for all x and the PSR is equal to 1. If $f_X(x) \neq f_Y(x)$ for any x, the PSR is equal to 0. For clarity of illustration, assume that X and Y follow a bivariate normal distribution throughout, and, without loss of generality, assume $\mu_X > \mu_Y$, $\sigma_X^2 = \sigma^2$, and

$\sigma_Y^2 = k^2\sigma^2$ for the remainder of this entry. When $k \neq 1$ there will be exactly two finite points of equality, x_L, x_U with $x_L < x_U$, where $f_X(x_L) = f_Y(x_L)$ and $f_X(x_U) = f_Y(x_U)$. Both x_L and x_U result from

$$\frac{(\mu_Y - k^2\mu_X) \pm \sqrt{k^2(\mu_X^2 + \mu_Y^2 - 2\mu_X\mu_Y) - k^2 2\sigma^2 \ln(k)(1-k^2)}}{(1-k^2)}. \quad (11)$$

A similar representation for the points of equality can be found in Inman and Bradley (1989). The PSR can be calculated by adding three probabilities shown in equation (12).

$$\text{PSR} = P(X \leq x_L) + P(x_L \leq Y \leq x_U) + P(X \geq x_U) \quad (12)$$

When $k = 1$, $f_X(x) = f_Y(x)$ at a single value $x_L = \frac{\mu_Y + \mu_X}{2}$. The calculation of the PSR is then simplified to

$$PSR = P(X \leq x_L) + P(Y \geq x_L) = 2 \times \Phi\left(\frac{\mu_Y - \mu_X}{2\sigma}\right). \quad (13)$$

The following proposition establishes a relationship between the PSR and the PIQI, with the details of the derivation given in the appendix.

Proposition 1 Assuming the bivariate normal distribution described earlier,

$$\text{PIQI}^{max} \geq \frac{\text{PSR}}{2}, \quad (14)$$

with equality at $k = 1$.

A similar result holds for subpopulations defined by either a continuous or a categorical covariate, *Z*. The conditional PSR is defined using the conditional distributions of *X* and *Y* given the observed covariate z_0 so that

$$\text{PSR}_{z_0} = \int min\left(f_{X|z_0}(x), f_{Y|z_0}(x)\right) dx. \quad (15)$$

As with the PSR, the relationship between PSR_{z_0} and PIQI_{z_0} depends on whether $\sigma^2_{X|Z} = \sigma^2_{Y|Z}$. Let $\sigma^2_{X|Z} = k^2_Z \sigma^2_{X|Z}$, where k^2_Z is the conditional k^2. Given conditional distributions $f_{X|z_0}$ and $f_{Y|z_0}$ at a finite value of z_0, then

$$\text{PIQI}^{max}_{z_0} \geq \frac{\text{PSR}_{z_0}}{2}. \tag{16}$$

with equality holding at $k_Z = 1$. This result follows directly from the proof of proposition 1. We are not aware of a similar result relating the PSR to the minimum bound for a PIQI.

Figure 2 illustrates the estimated PSR for the data example. The first panel shows the unconditional PSR and the other 3 show the estimated conditional PSR at the three values of z_0 given in Table 1. The estimated values for (1/2)PSR for the 4 plots in Figure 2 are very close to those reported earlier for the estimated maximum PIQI (e.g., see Table 1) because the value for k (and k_Z) are very close to 1. The estimated PSR decreases with increasing baseline weight - the result of the treatment – baseline weight interaction. Standard errors for an estimated PSR can be obtained using bootstrap samples, and in the data example were similar to those reported earlier for the estimated PIQI^{max}.

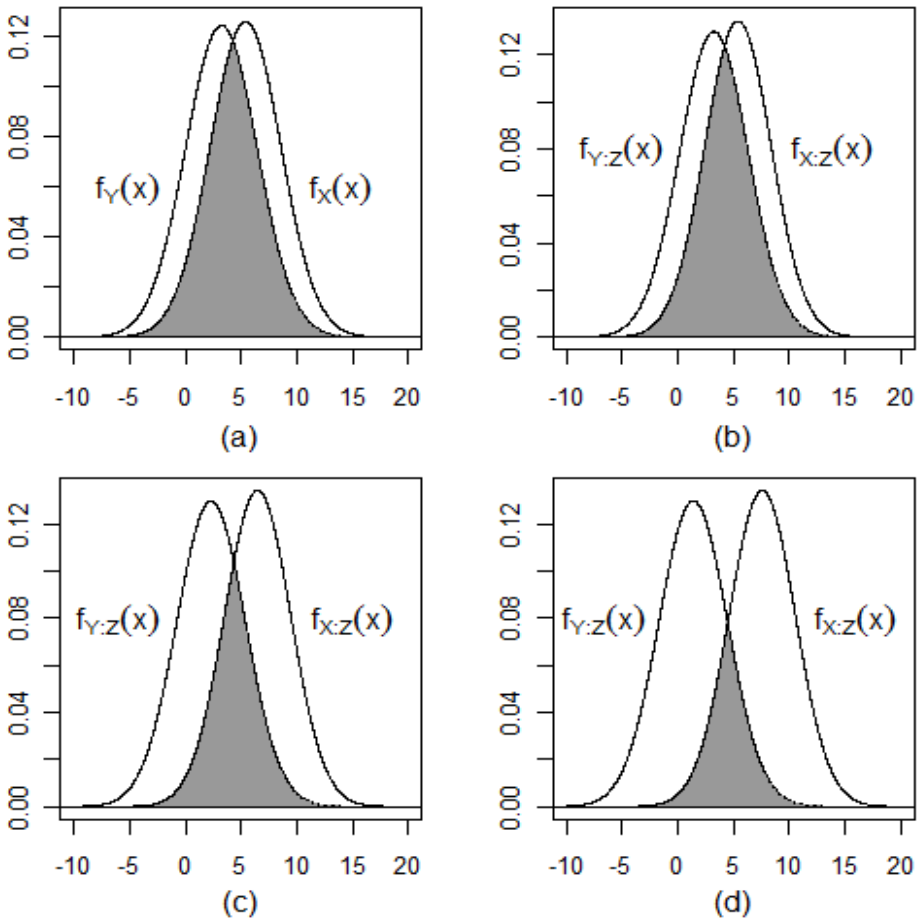

Figure 2: Illustration of the PSR using only the marginal distributions of X and Y (panel (a)) and at three values of baseline weight given in Table 1 (panels b – d).

3.3 Cross-Over Designs

Perhaps the most straightforward design for estimating σ_D^2 and the PIQI is a cross-over design. A large body of literature exists on estimating mean treatment effects, mean period effects, carry-over effects, etc. (e.g., Senn 2006a; Yang and Stufken 2008). Mixed-effects models fit to data from a cross-over design with a random subject effect may even compute what some have referred to as a "subject-treatment interaction variance" (e.g., Hauck et al. 2000; Endrenyi and Tothfalusi 1999). However, this variance computed from observed data may not equal a variance of true individual effects without certain assumptions and/or depending upon how one defines an individual effect in multiple period designs. We illustrate concepts for a two

period two treatment cross-over design, assuming no carry over effects but that period effects may vary across individuals. Potential outcomes are (X_1, Y_1) at period 1 and (X_2, Y_2) at period 2, and these are rewritten as $\left(X - t,\ Y - \tau\right)$ at period 1 and $\left(X + t,\ Y + \tau\right)$ at period 2. Now the pair $\left(X, Y\right)$ quantify a mean response over the two periods on each of the two treatments, and the pair $\left(t,\ \tau\right)$ accommodate effects from period 1 to period 2 for each outcome. There are two "true" individual treatment effects given by $D_1 = (X - Y) - (t - \tau)$ at period 1 and $D_2 = (X - Y) + (t - \tau)$ at period 2. In some applications it may be D_1 that is the effect of most interest. Another effect may be defined as the average over the two time periods, denoted as $D = (D_1 + D_2)/2$. The true individual effect is constant across the two time periods if $t - \tau = 0$, an assumption that may be reasonable with no carry-over effects.

Since each individual is crossed over from one treatment to another after a washout period, an individual treatment effect may seem to be observable. Assume that n_1 subjects are randomly assigned to the sequence, *TR*, where *TR* implies treatment *T* at time 1 and treatment *R* at time 2, and n_2 subjects to the reverse sequence of treatments, *RT*, with $n_1 + n_2 = n$. The observed differences are d_j, for $j = 1,2, \ldots, n$ and can be written as $d_j = (X_j - Y_j) - (t_j + \tau_j)$ if the *j*th subject was assigned to sequence *TR*, and $d_j = (X_j - Y_j) + (t_j + \tau_j)$ if assigned to sequence *RT*.

A straightforward naïve estimate of the PIQI may be obtained using equation (2), with $\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j$ to estimate μ_D and $s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(d_j - \bar{d}\right)^2$ to estimate σ_D^2. Following results analogous to Gadbury (2001) it can be shown that s_d^2 is positively biased for VAR$(D) = \sigma_D^2$ when an individual effect is defined by $D = (D_1 + D_2)/2$, and the bias term is a function of $(t_j + \tau_j),\ \ j = 1, \ldots, n$. If $(t_j + \tau_j)$ is assumed to be constant across individuals (i.e., a constant

sequence effect), then the bias can be estimated from observed data. If this is not assumed, but it is assumed that the true individual effect is constant across periods, meaning that $t_j = \tau_j$ for $j = 1,\ldots,n$, then the bias can be estimated if t_j is constant across individuals, i.e., a constant period effect. If period effects vary across individuals, but $t_j = \tau_j$ for all *j*, then the bias can be estimated with an extension to the design such as the *Balaam* (Balaam 1968) design, where some subjects remain on the same treatment over the two periods (i.e., *TR*, *RT*, *TT*, and *RR* sequences).

Required assumptions for direct estimation of σ_D^2 may be more plausible in certain applications than assumptions that are required without the multiple period feature of the design. Even with no additional assumptions, estimated bounds for σ_D^2 (or PIQI) may be tighter than those obtained from single period designs. Repeated measures cross-over designs have advantages over single period designs for estimating subject-treatment interaction and its consequences (e.g., Senn 2001). More methodological development is needed to define the required assumptions and resulting estimators from different types of cross-over designs, and potential outcomes may be the best structure to use when doing this.

Cross-over designs, however, are not always practical to implement in many applications (cf., Brown 1980; Senn 2001). For instance, in applications like the data example used herein, there may be limitations in using cross-over designs when the primary outcome variable is weight loss. The true individual effect of a treatment at time 1 may be substantially larger than at time 2 because people tend to lose weight more rapidly at first, and substantial carry-over effects may be likely as well.

4. DISCUSSION AND CONCLUSIONS

In 1892 Sir William Osler stated, “If it were not for the great variability among individuals medicine might as well be a science and not an art” (extracted from Roses 2000, page

857). Thus the topics discussed here have been long recognized as important considerations when selecting a "best" treatment for an individual. Individual treatment heterogeneity and its consequences should be an important consideration when designing clinical trials and interpreting treatment efficacy and safety for a target population. The quantities discussed herein may also inform the pursuit of pharmacogenetic research which seeks to identify genomic predictors of response to treatments (e.g., Hu et al. 2006). It is often poorly understood how much heterogeneity might be present in the first place and whether a search for such gene-treatment interactions that explain this heterogeneity will be fruitful (Senn 2001). Perhaps we should invest most readily in finding genetic factors influencing variability in treatment response for those treatments for which we have actually demonstrated, rather than merely presumed, large variability in response.

Evaluating the plausible variance in treatment response, and even more so the proportion of a population with an IQI, has other applications as well. The Latin enjoinder *primum non nocere* (above all, do no harm) frequently (mis)attributed to Hippocrates (Smith 2005) remains a mainstay of medical thinking today. Thus, regulatory agencies such as the US Food and Drug Administration may wish to know not only the average effect of a drug compared to placebo, but the probability that it will have a poorer effect than no drug. Similarly, when faced with the possibility of approving a new drug that is no more efficacious on average than an existing drug which has been widely used and which has already survived the baptism of fire that is widespread clinical use, it is tempting to ask "why do we need to approve this new drug if it is no more efficacious than the old drug we know and trust?" A typical response is one voiced by the director of the UK's National Institute for Health and Clinical Excellence's health technology evaluation centre, who recently stated, "Different people respond in different ways to treatment, and the committee heard from clinical experts and patients about the importance of having

multiple options available" (Mayor 2010). That is, it is often presumed that although *drug A* may be no better than *drug B* on average, for some persons, *drug A* works better than does *drug B* and vice versa for other persons. If this is the case, then having multiple drugs on the market may be important even if there is no difference in their effects on average and they cannot be used in combination. However, rather than accepting the premise as true a priori, the results shown herein may help lead to new ideas for the evaluation of the plausibility and frequency of such IQIs.

The ideas may also be useful in evaluating advertising claims. Consider the context of claims for weight loss products which can often be quite extravagant. The US Federal Trade Commission (FTC) states that "No [weight loss] product will work for everyone," and therefore claims implying that a "product causes substantial weight loss for all users" is a likely sign of fraud (FTC statement). Is there evidence a company could provide to FTC to show that in their randomized clinical trial (RCT) showing a positive mean effect, the plausible proportion of people who will have an effect less than a threshold τ is negligible? Alternatively, is there evidence that FTC could muster to show a company that their claim of a universal positive effect is almost certainly untrue despite their being a positive mean effect? Again, the results described herein may help clarify the issues involved when answering these questions.

Finally, one can imagine applications in legal settings (see, for example, Marchant 2001, 2010). Imagine that a plaintiff (e.g., a consumer) sues a defendant (e.g., a distributor of a drug, food, or pharmaceutical) claiming that use of defendant's product caused a stroke secondary to markedly elevated blood pressure (BP) as a result of using the product. Imagine further that defense experts present evidence that well-designed RCTs show an *average* effect of the product on BP to be less than or equal to zero. Plaintiff's experts reply that there is great interindividual variability in response and even though the average response is less than or equal to zero, some

people will be hypersensitive hyper-responders with extreme BP increases. What evidence can the court bring to bear on the question of how probable it is that plaintiff was such a hyper-responder? The first question which must precede this is what evidence is there that hyper-responders in the opposite direction even exist and with what frequency? The techniques herein may provide a plausible range of answers.

Potential outcomes are a natural way to define individual treatment effects and metrics that quantify treatment heterogeneity as well as the risk of a qualitative interaction across individuals or groups of individuals. Existing techniques that seek to evaluate treatment heterogeneity have limitations, and these limitations are made clear by potential outcomes. The potential outcomes framework can delineate heterogeneity that is observable from that which is not, and unobservable heterogeneity can often be bounded by quantities that can be estimated in observed data. Thus, the potential outcomes framework is a useful complement to existing techniques to evaluate treatment heterogeneity and qualitative interactions, and they should be used as such when analyzing data from randomized trials. Their use may also suggest new directions in the design of randomized trials – directions that do not compromise estimation of mean effects but also allow for more direct evaluation of treatment heterogeneity. Eventually, perhaps, reporting of treatment heterogeneity and risks of qualitative interactions (at either the individual or group level), in addition to summary measures such as mean effects, will be a more standard practice, and a response to a perceived need that has been recognized by others in recent years (cf., Longford 1999).

APPENDIX: *Proof of Proposition 1*:

The equality at $k = 1$ is straightforward. Let $k > 1$. When $\rho_{XY} = -1$, the (x, y) pairs are constrained to the line $y = \mu_Y + k\mu_X - kx$ with probability one. If we let x and y be equal and set to the common value $x_{-1} = \frac{\mu_Y + k\mu_X}{1+k}$, then it can be shown that

$$\text{PIQI}^{max} = P(X \leq x_{-1}) = P(Y \geq x_{-1}). \quad (17)$$

Therefore,

$$\begin{aligned}
2 \times \text{PIQI}^{max} &= P(X \leq x_{-1}) + P(Y \geq x_{-1}) \\
&= P(X \leq x_L) + P(x_L \leq X < x_{-1}) + P(x_L \leq Y < x_U) - P(x_L \leq Y < x_{-1}) \\
&\quad + P(Y \geq x_U) + P(X \geq x_U) - P(X \geq x_U) \\
&= P(X \leq x_L) + P(x_L \leq Y < x_U) + P(X \geq x_U) + P(x_L \leq X < x_{-1}) \\
&\quad - P(x_L \leq Y < x_{-1}) + P(Y \geq x_U) - P(X \geq x_U) \\
&= \text{PSR} + P(x_L \leq X < x_{-1}) - P(x_L \leq Y < x_{-1}) + P(Y \geq x_U) - P(X \geq x_U)
\end{aligned}$$

Thus, $\text{PIQI}^{max} \geq \frac{\text{PSR}}{2}$, because $P(x_L \leq X < x_{-1}) \geq P(x_L \leq Y < x_{-1})$ and $P(Y \geq x_U) \geq P(X \geq x_U)$ when $k \geq 1$. Proof for $k < 1$ is similar. ■

REFERENCES

Balaam, L. N. (1968), "A Two-period Design with *t*2 Experimental Units," *Biometrics*, 24, 61–73.

Brown, B. W. (1980), "The Crossover Experiment for Clinical Trials," *Biometrics*, 36, 69–79.

Byar, D. P., and Corle, D. K. (1977), "Selecting Optimal Treatment in Clinical Trials Using Covariate Information," *Journal of Chronic Diseases*, 30, 445–459.

Darwin, C. (1871), *On the Origin of Species* (5th ed.), New York: D. Appleton and Company, p. 26.

Endrenyi, L., and Tothfalusi, L. (1999), "Subject-by-Formulation Interaction in Determinations of Individual Bioequivalence: Bias and Prevalence," *Pharmaceutical Research*, 16, 186–190.

Fan, Y., and Park, SS. (2009), "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets," in *Nonparametric Econometric Methods (Advances in Econometrics, Volume 25)*. Edited by Qi Li and Jeffrey Racine. Emerald Group Publishing Limited, 3 – 70.

Fan, Y., and Park, SS. (2010), "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory*, 26, 931 – 951.

FTC Statement. Retrieved on Sept. 10, 2010 from
http://www.ftc.gov/bcp/edu/pubs/business/adv/bus60.pdf.

Gadbury, G. L., and Iyer, H. K. (2000), "Unit-Treatment Interaction and its Practical Consequences," *Biometrics*, 56, 882–885.

Gadbury, G. (2001), "Randomization Inference and Bias of Standard Errors," *Journal of the American Statistical Association*, 55, 1–4.

Gadbury, G. L., Iyer, H. K., and Allison, D. B. (2001), "Evaluating Subject-Treatment Interaction when Comparing Two Treatments," *Journal of Biopharmaceutical Statistics*, 11, 313–333.

Gadbury, G.L. (2010), Subject-Treatment Interaction. In *Encyclopedia of Biopharmaceutical Statistics, Third Edition, Revised and Expanded*. Edited by Shein-Chung Chow. Informa Healthcare, London. p. 1316 – 1321.

Gail, M., and Simon, R. (1985), "Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets," *Biometrics*, 41, 361–372.

Gastwirth, J. L. (1975), "Statistical Measures of Earnings Differentials," *The American Statistician*, 29, 32–35.

Hauck, W. W., Hyslop, T., Mei-Ling, C., Patnaik, R., and Williams, R. L. (2000), "Subject-by-Formulation Interaction in Bioequivalence: Conceptual and Statistical Issues," *Pharmaceutical Research*, 17, 375–380.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Hu, J., Redden, D.T., Berrettini, W.H., Shields, P.G., Restine, S.L., Pinto, A., Lerman, C., Allison, D.B. (2006), 'No Evidence for a Major Role of Polymorphisms During Bupropion Treatment," *Obesity (Silver Spring)*, 14, 1863–1867.

Inman, H. F., and Bradley, E. L. (1989), "The Overlapping Coefficient as a Measure of Agreement Between Probability Distributions and Point Estimation of the Overlap of Two Normal Densities," *Communications in Statistics: Theory and Methods*, 18, 3851–3874.

Inoue, J., Hoshino, R., Nojima, H., Ishida, W., Okamoto, N. (2010), "Investigation of Responders and Non-responders to Long-Term Donepezil Treatment," *Psychogeriatrics*, 10(2), 53–61.

Li, J., Chan, I. S. F. (2006), "Detecting Qualitative Interactions in Clinical Trials: An Extension of Range Test," *Journal of Biopharmaceutical statistics*, 16, 831–841.

Longford, N. T. (1999), "Selection Bias and Treatment Heterogeneity in Clinical Trials," *Statistics in Medicine*, 18, 1467–1474.

Lucretius. Retrieved Sept. 9, 2010 from http://classics.mit.edu/Carus/nature_things.4.iv.html

Maki, K.C., Rains, T.M., Kaden, V.N., Raneri, K.R., Davidson, M.H. (2007), "Effects of a reduced-glycemic-load diet on body weight, body composition, and cardiovascular disease risk markers in overweight and obese adults," *American Journal of Clinical Nutrition*, 85, 724 – 734.

Marchant, G.E. (2001), "Genetics and Toxic Torts," *Seton Hall Law Review*, 31, 949.

Marchant, G.E. (2010). Retrieved on Sept. 11, 2010 from http://www.law.asu.edu/files/Programs/Sci-Tech/Commentaries/Marchant%20Formatted.rev.doc

Marshall, A. (2007), "Laying the foundations for personalized medicines", *Nature Biotechnology*, 15, 954 – 957.

Mayor, S. (2010), "NICE Recommends Widening Choice of Biological Drugs for Patients with Rheumatoid Arthritis," *BMJ*, 340, c3477.

Peto, R. (1982), *Statistical Aspects of Cancer Trials*, Chapman and Hall, 867–871.

Rom D. M., and Hwang, E. (1996), "Testing for Individual and Population Equivalence Based on the Proportion of Similar Responses," *Statistics in Medicine*, 15, 1489–1505.

Roses, A. D. (2000), "Pharmacokinetics and the Practice of Medicine," *Nature*, 405, 857–865.

Rubin, D. B. (1974), "Estimating Causal Effects for Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Senn, S. (1997), Letter to the editor on "Testing for Individual and Population Equivalence Based on the Proportion of Similar Responses," *Statistics in Medicine*, 16, 1301–1306.

Senn, S. (2001), "Individual Therapy: New Dawn or False Dawn?" *Drug Information Journal*, 35, 1479–1494.

Senn, S. (2006a), "Cross-over Trials in *Statistics in Medicine*: The First '25' Years," *Statistics in Medicine*, 25, 3430–3442.

Senn, S. (2006b), Letter to the editor on "Probability Index: An Intuitive Non-parametric Approach to Measuring the Size of Treatment Effects," *Statistics in Medicine*, 25, 3944–3948.

Simon, R., (1982), "Patient Subsets and Variation in Therapeutic Efficacy," *British Journal of Pharmacology*, 14, 473–482.

Smith, C.M. (2005), "Origin and Uses of Primum Non Nocere--Above All, Do No Harm!," Journal *of Clinical Pharmacology*, 45, 371–377.

Stine, R. A., and Heyse, J. F. (2001), "Non-parametric Estimates of Overlap," *Statistics in Medicine*, 20, 215–236.

Silvapulle, M. J. (2001), "Tests for Qualitative Interaction: Exact Critical Values and Robust Tests," *Biometrics*, 57, 1157–1165.

Yang, M., and Stufken, J. (2008), "Optimal and Efficient Crossover Designs for Comparing Test Treatments to a Control Treatment Under Various Models," *Journal of Statistical Planning and Inference*, 138, 278–285.