Research Article

Statistics in Medicine

Received 27 October 2009, Accepted 2 March 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3923

Generalized pairwise comparisons of prioritized outcomes in the two-sample problem

Marc Buyse[a,b*†]

This paper extends the idea behind the U-statistic of the Wilcoxon–Mann–Whitney test to perform generalized pairwise comparisons between two groups of observations. The observations are outcomes captured by a single variable, possibly repeatedly measured, or by several variables of any type (e.g. discrete, continuous, time to event). When several outcomes are considered, they must be prioritized. We show that generalized pairwise comparisons extend well-known non-parametric tests, and illustrate their interest using data from two randomized clinical trials. We also show that they lead to a general measure of the difference between the groups called the 'proportion in favor of treatment', denoted Δ, which is related to traditional measures of treatment effect for a single variable. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: generalized pairwise comparisons; prioritized outcomes; measure of treatment effect; randomization test

1. Introduction

Parametric and non-parametric tests are available to compare two groups of observations in terms of a single variable, possibly repeatedly measured, or of several variables in multivariate analysis. These tests, however, are not designed to handle the fully general situation in which two groups of individuals must be compared in terms of several variables of different types measured on different occasions. In this paper, we extend the idea behind the U-statistic of the Wilcoxon–Mann–Whitney test to perform generalized pairwise comparisons between the observations of the two groups [1, 2]. This method allows us to compare two groups regardless of the number and types of variables considered (e.g. discrete, continuous, time to event). We show that generalized pairwise comparisons extend well-known non-parametric tests, and illustrate their interest using data from two randomized clinical trials. We also show that they lead to a general measure of the difference between the groups called the 'proportion in favor of treatment', denoted Δ, which is related to traditional measures of treatment effect for a single variable.

The paper is organized as follows. In Section 2, we introduce two motivating case studies, one in ophthalmology, the other in oncology. Generalized pairwise comparisons are described for a single variable in Section 3, and extended to the multivariate case in Section 4. Sections 5 and 6 cover, respectively, the estimation and testing of the general measure of treatment effect Δ. In Section 7, pairwise comparisons are shown to unify non-parametric tests for binary, continuous, and time-to-event variables, while the link between the proportion in favor of treatment and other measures of treatment effect is discussed in Section 8. Section 9 illustrates these concepts on actual data from the two case studies, and Section 10 discusses some of the benefits and limitations of the proposed approach.

2. Presentation of case studies

2.1. A randomized trial in advanced colorectal cancer

We will first illustrate generalized pairwise comparisons using data from a randomized trial of 420 patients with advanced colorectal cancer [3]. Patients were randomized to either a standard regimen of 5-fluorouracil and leucovorin ('LV5FU2'), or to the same regimen plus oxaliplatin. Two time-to-events variables were of interest in this trial: progression-free survival, defined as the time from randomization to objective disease progression or death, whichever came first, and

[a] *International Drug Development Institute, 30 avenue provinciale, 1340 Louvain-la-Neuve, Belgium*
[b] *Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Belgium*
[*] *Correspondence to: Marc Buyse, International Drug Development Institute, 30 avenue provinciale, 1340 Louvain-la-Neuve, Belgium.*
[†] *E-mail: marc.buyse@iddi.com*

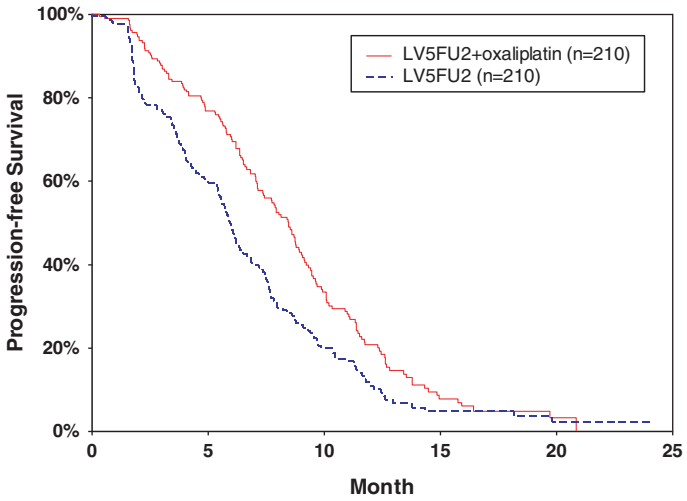

Figure 1. Progression-free survival for patients with advanced colorectal cancer randomized to standard treatment LV5FU2 with or without oxaliplatin.

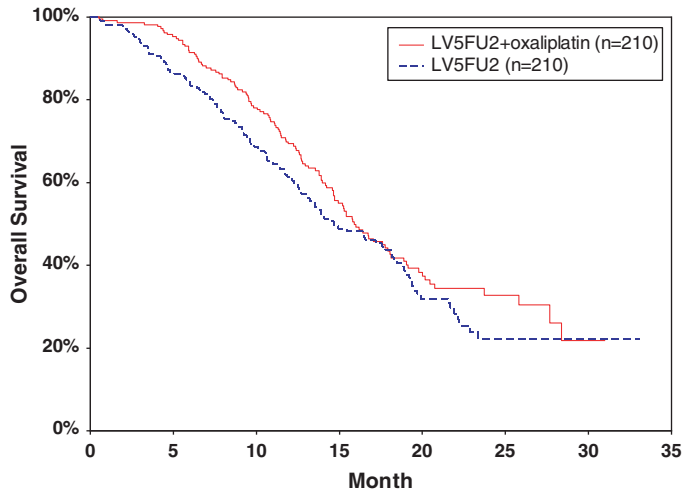

Figure 2. Survival for patients with advanced colorectal cancer randomized to standard treatment LV5FU2 with or without oxaliplatin.

survival. The trial showed a highly statistically significant benefit of the addition of oxaliplatin on progression-free survival (hazard ratio $=0.66$, logrank test $P=0.0001$, Figure 1), but it failed to reach conventional statistical significance for the benefit of oxaliplatin on survival (hazard ratio $=0.83$, logrank test $P=0.135$, Figure 2).

These results led to controversies in interpretation. The new drug oxaliplatin was in fact approved for marketing several years earlier in Europe by the European Medicines Evaluation Agency (EMEA) than in the Unites States by the Food and Drug Administration (FDA). The method proposed in this paper sheds a different light on these data by considering both times (to death and to progression) to test for, and estimate, an overall treatment benefit that has some intuitive appeal and is not directly related to the survival or progression-free survival hazard ratios.

2.2. Two randomized trials in age-related macular degeneration

We will further discuss the potential of generalized pairwise comparisons using data from two double-blind randomized clinical trials in 1186 patients with neovascular age-related macular degeneration [4]. Patients were randomized to receive intraocular injections of pegaptanib, an antivascular endothelial growth factor therapy, or sham injections with a syringe applied on the surface of the eye to simulate the pressure of an injection, every 6 weeks over a period of one year. The treatment efficacy was assessed 6 weeks after each injection through the visual acuity, defined as the number of letters correctly read on a standardized visual acuity chart. This visual acuity chart comprises lines of 5 letters each, with the size of the letters decreasing from the top line to the bottom line. The chart must be read with one eye masked from a fixed distance.

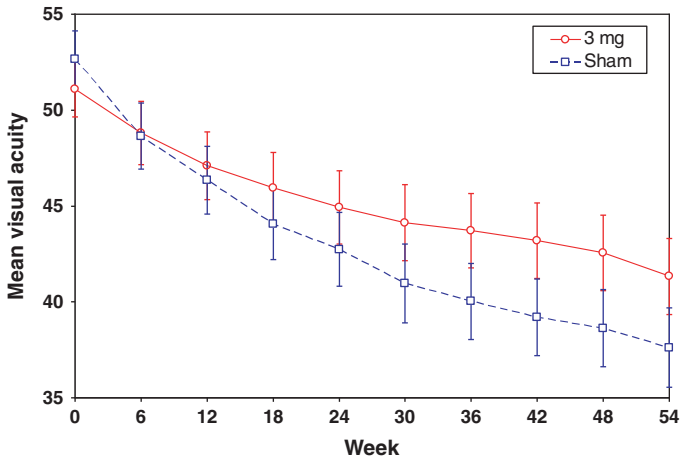

Figure 3. Mean visual acuity as a function of time for patients with age-related macular degeneration randomized to sham or to 3 mg of pegaptanib.

Visual acuity ranges from 0 (complete blindness) to 100 (perfect vision). Figure 3 illustrates the drop in mean visual acuity over time for two groups of patients: those randomized to receive a dose of 3 mg of pegaptanib (which had less favorable results than the approved dose of 0.3 mg and is used here for illustrative purposes) and those randomized to receive sham injections.

The primary 'endpoint' of the trial, as defined by the EMEA and the FDA, was the proportion of patients losing at least 15 letters of visual acuity one year (54 weeks) after starting therapy. The last observation of visual acuity was carried forward for patients in whom the measurement at 54 weeks was unavailable. With this data imputation, the proportion of patients losing at least 15 letters of visual acuity was 35 per cent for patients receiving a dose of 3 mg of pegaptanib vs 45 per cent for patients receiving sham injections (an absolute benefit of 10 per cent, χ^2 test $P=0.015$) [4]. This analysis, although very suggestive of a treatment benefit, suffered from several drawbacks: it only used the observations made at baseline and at one year, did not consider gains in vision, and required data imputation. The method proposed here is subject to none of these problems, yet it complies with the regulatory requirement that differences in visual acuity must equal or exceed 15 letters to reflect a meaningful treatment benefit.

3. Generalized pairwise comparisons

We are interested in the general situation of two groups of individuals (whom we call 'patients' in the clinical trial examples) to be compared in terms of one or more outcome measures (or 'endpoints') observed at one or more occasions for each individual. Formally, the outcome measures of interest are captured by random variables, the values of which are the individual outcomes. We assume that one group of n individuals is exposed to an intervention or treatment (labeled 'T'), while the other group of m individuals serves as a control (labeled 'C'). Such a situation is typical of comparative trials where patients are allocated to treatment or control through a random mechanism, as in the two case studies described in Section 2 and analyzed in Section 9. As for other two-sample tests, there is no requirement that the two groups be formed by random allocation: they can also be formed by independent random sampling from two populations, or by any other sampling scheme under a condition of exchangeability of individuals under the null hypothesis that will be further discussed below [5].

Pairwise comparisons require consideration of pairs of individuals, one taken from group T and the other taken from group C. The outcomes of these two individuals are compared and the pair is said to be 'favorable' if the outcome of the individual in group T is better than the outcome of the individual in group C, 'unfavorable' if the outcome of the individual in group T is worse than the outcome of the individual in group C, 'neutral' if there is no difference between the outcomes of the two individuals, or 'uninformative' if it cannot be determined which of the two individuals has a better outcome (e.g. if the outcome is missing for at least one of the two individuals). For a pair to be considered favorable, unfavorable or neutral, a 'better outcome' must be defined for every possible pair of values of the variable of interest. In most cases, and in the examples described below, the definition of a better outcome is self-evident, but it may be made as complex as required by, or meaningful for, the situation at hand. In the following sections, a better outcome is defined for common types of variables.

Table I. Pairwise comparisons for a binary variable.

Pairwise comparison	Pair is
$X_i=1,\ Y_j=0$	favorable
$X_i=1,\ Y_j=1$	neutral
$X_i=0,\ Y_j=0$	neutral
$X_i=0,\ Y_j=1$	unfavorable

Table II. Generalized pairwise comparisons for a continuous variable.

Pairwise comparison	Pair is
$X_i-Y_j>\tau$	favorable
$\lvert X_i-Y_j\rvert\leqslant\tau$	neutral
$X_i-Y_j<-\tau$	unfavorable

3.1. Binary variable

Assume that the outcome measure of interest is binary in nature. For reasons that will become clear later in this paper, it is convenient to denote this binary variable X in the treatment group and Y in the control group, with $X=1$ (or $Y=1$) indicating success, and $X=0$ (or $Y=0$) indicating failure. Table I displays the possible situations that can arise in the comparison of X_i, the outcome of the ith individual ($i=1,\ldots,n$) in group T with Y_j, the outcome of the jth individual ($j=1,\ldots,m$) in group C.

In Section 7.1, a link will be established between pairwise comparisons and Fisher's exact test.

3.2. Continuous variable

Assume now that the outcome measure of interest is captured by continuous variable X in the treatment group and Y in the control group. Assume further, without loss of generality, that larger values of X (and Y) are preferable to smaller values of X (and Y). In some applied settings, the difference between the values of these two variables may have to exceed a pre-specified threshold, denoted τ, to be considered meaningful. The threshold can be a function of the precision with which X (and Y) is measured. In clinical trials, the threshold can also reflect a difference regarded as clinically relevant. Table II displays generalized pairwise comparisons of continuous variables with a threshold τ.

In Section 7.2, pairwise comparisons will be shown to be equivalent to the Wilcoxon rank-sum test in the special case where $\tau=0$.

3.3. Time-to-event variable

The variables X and Y can be right censored. We call these variables 'times to event', though right censoring can occur in situations in which the variables X and Y do not represent time. Let variables ε and η denote the censoring of variables X and Y, respectively, with $\varepsilon_i=1$ (or $\eta_j=1$) indicating that X_i (or Y_j) is a complete, or 'uncensored', observation. Table III displays the possible outcomes of pairwise comparisons. Censored observations, which may cause pairs to be uninformative, are denoted X_i' and Y_j' to distinguish them from complete observations.

In Section 7.3, generalized pairwise comparisons will be shown to be equivalent to Gehan's generalized Wilcoxon rank-sum test in the special case where $\tau=0$.

3.4. Other variable types

Generalized pairwise comparisons can accommodate variables of any type as long as there is a unique way of defining a better outcome for each possible pair of values of the variable considered. As such, this approach is extremely general: for instance, there need not be a transitive relation between the values of the variable, as in the 'rock–paper–scissors' game where rock beats scissors, scissors beat paper, and paper beats rock. A variable taking these three values would lend itself to pairwise comparisons even though no parameterization of this problem is possible to develop a two-sample test.

3.5. Pairwise indicator for a single outcome

We introduce a pairwise indicator that reflects whether T, C, or neither is favored in a pair of individuals when a single outcome measure is of interest. In many experimental designs, stratification is used to match individuals for important characteristics such as gender, age, stage of disease, etc. In such designs, pairwise comparisons are restricted to pairs

Table III. Generalized pairwise comparisons for a time-to-event variable.

Censoring of X and Y	Pairwise comparison	Pair is
$\varepsilon_i=1,\ \eta_j=1$	$X_i-Y_j>\tau$	favorable
	$\lvert X_i-Y_j\rvert\leqslant\tau$	neutral
	$X_i-Y_j<-\tau$	unfavorable
$\varepsilon_i=0,\ \eta_j=1$	$X'_i-Y_j>\tau$	favorable
	$\lvert X'_i-Y_j\rvert\leqslant\tau$	uninformative
	$X'_i-Y_j<-\tau$	uninformative
$\varepsilon_i=1,\ \eta_j=0$	$X_i-Y'_j>\tau$	uninformative
	$\lvert X_i-Y'_j\rvert\leqslant\tau$	uninformative
	$X_i-Y'_j<-\tau$	unfavorable
$\varepsilon_i=0,\ \eta_j=0$	$X'_i-Y'_j>\tau$	uninformative
	$\lvert X'_i-Y'_j\rvert\leqslant\tau$	uninformative
	$X'_i-Y'_j<-\tau$	uninformative

of individuals within the same stratum. In order to account for such stratification, we define a pairwise indicator for the pair formed by the ith individual ($i=1,\ldots,n_k$) in group T and the jth individual ($j=1,\ldots,m_k$) in group C in the kth stratum ($k=1,\ldots,K$):

$$p_{ijk}=\begin{cases}1 & \text{if the pair is favorable}\\ -1 & \text{if the pair is unfavorable}\\ 0 & \text{if the pair is neutral}\end{cases}$$

4. Prioritized outcomes

Generalized pairwise comparisons can be extended to several outcomes arising from successive thresholds of a single outcome measure (Section 4.1), from repeated observations of a single outcome measure (Section 4.2), or from several outcome measures (Section 4.3). We will consider the extension to several outcome measures when an ordering of the multivariate space can be defined by prioritizing the variables. Wittkowski *et al.* [6] have developed a related approach based on a partial ordering of the multivariate space. Other extensions are possible but will not be pursued here.

4.1. Multiple thresholds

In Tables II and III, generalized pairwise comparisons could use a threshold to reflect meaningful differences in outcomes captured by a continuous or time-to-event variable. A natural extension consists of considering successive thresholds of that variable. If we assume, as above, that larger values of the variables are preferable to smaller values, then larger thresholds will take priority over smaller thresholds.

For instance, in macular degeneration (Section 2.2), a better outcome in pairwise comparisons can be defined as a difference in vision of at least l letters, with l varying from more than 30 letters (a remarkable difference that reflects a highly relevant difference in vision) to less than 5 letters (a minor difference that could easily be due to measurement error, changes in reading conditions, or chance variation). In the regulatory setting of a new drug approval for this condition, a difference in visual acuity of at least 15 letters is considered clinically relevant. In this example, thresholds representing decreasing differences in the numbers of letters could be chosen, say, as $l=30, 25, 20, 15, 10, 5, 0$.

4.2. Repeated observations

Generalized pairwise comparisons can easily be extended to repeated observations of a variable capturing the outcome measure of interest if the different occasions at which the variable is potentially measured are prioritized. For instance, when the variable is measured repeatedly over time (longitudinal data), a later difference between the groups may be more relevant than an earlier one in so far as it reflects a sustained effect of the intervention or treatment over time. In this case, a later difference will take priority over an earlier difference in pairwise comparisons. The clinical trial in macular generation (Section 2.2) again provides an example of such a situation in which up to 10 longitudinal measurements of visual acuity, taken 6 weeks apart, are available for each patient.

Table IV. Generalized pairwise comparisons for repeated observations.

Occasion with higher priority	Occasion with lower priority	Pair is
favorable	ignored	favorable
unfavorable	ignored	unfavorable
neutral	ignored	neutral
uninformative	favorable	favorable
uninformative	unfavorable	unfavorable
uninformative	neutral	neutral
uninformative	uninformative	uninformative

A better outcome will be defined for generalized pairwise comparisons by using the repeated observations of the variable in descending level of priority of the occasions, as shown in Table IV for two occasions. The generalization to more than two occasions is immediate using an iterative argument.

Missing values: Table IV suggests a natural way of handling missing values in generalized pairwise comparisons: whenever a pairwise comparison is uninformative at an occasion of higher priority, the occasions of lower priority are examined. This feature is useful, for example, to analyze longitudinal data in the presence of attrition, i.e. when some individuals are observed until a certain time but not beyond that time. When there is attrition, the observations of one group are only compared with the observations made at the same time in the other group. The method suggested by Gould [7] to handle informative withdrawals (patients leaving a trial for efficacy or lack thereof) can also be built in generalized comparisons if desired.

Breaking of neutral pairs: The various occasions can also serve to break neutral pairs whenever this is deemed relevant. Hence in Table IV, if the pair was neutral for the occasion with higher priority, the occasion with lower priority could be used to decide whether the pair is favorable, unfavorable, or neutral.

4.3. Several outcome measures

Generalized pairwise comparisons can also be extended to several outcome measures by prioritizing the variables that capture them in order to define a better outcome, just as the occasions were prioritized in the case of repeated observations of a single outcome measure. A better outcome is defined for each of these variables, and a better outcome overall is then defined as a better outcome for the variable with the highest priority, as in Section 4.2. The case study in advanced cancer (Section 2.1) provides a situation in which the outcomes of interest are captured by two times to event (time to disease progression and time to death). In order of priority, time to death unquestionably comes first. Hence, in generalized pairwise comparisons, if the two patients of a pair have died, the better outcome corresponds to the patient with the longer survival time. If the two survival times cannot be compared as a result of the two patients being alive, or one patient being alive with a shorter survival time than the patient who died, then the times to disease progression are considered.

The prioritized variables can be of different types. In advanced cancer, for instance, in addition to time to death and time to disease progression, the achievement of a 'tumor response' (defined for solid tumors as greater than 50 per cent shrinkage of the tumor surface area) may sometimes be a relevant indicator of treatment benefit, though the time to achieve such a response is generally unimportant since most responses are obtained soon after starting therapy. In generalized pairwise comparisons, tumor response could be a binary outcome with lowest priority, used only for those pairwise comparisons that are uninformative both for time to death and for time to disease progression. In the same spirit, Moyé *et al.* [8] proposed a generalized Gehan–Wilcoxon test to combine time to death and time to a given decrease in ejection fraction in a post-myocardial infarction trial. Finkelstein and Schoenfeld [9] further proposed a test that combines a time-to-event variable with a longitudinally measured variable. They showed the versatility of their test in clinical trials for patients with acquired immune deficiency syndrome, in which various types of longitudinal data are relevant, such as repeated episodes of pneumonia, measurements of the head circumference in children, or quality of life assessments in adults. Generalized pairwise comparisons encompass such tests and therefore permit the comparison of two groups of individuals in terms of complex outcomes.

4.4. Pairwise indicators for prioritized outcomes

We now define two pairwise indicators for the lth outcome ($l=1,\ldots,L$), with outcomes numbered from 1 (highest priority level) to L (lowest priority level) in the pair formed by the ith individual ($i=1,\ldots,n_k$) in group T and the jth individual ($j=1,\ldots,m_k$) in group C in the kth stratum ($k=1,\ldots,K$). The first indicator is set as follows:

$$u_{ijk}(l)=\begin{cases}1 & \text{if the } l\text{th outcome is uninformative}\\ 0 & \text{otherwise}\end{cases}$$

The second indicator may be set as follows:

$$p_{ijk}(l)=\left\{\begin{array}{ll}1 & \text{if the } l\text{th outcome is favorable}\\ -1 & \text{if the } l\text{th outcome is unfavorable}\\ 0 & \text{if the } l\text{th outcome is neutral}\end{array}\right\} \quad \text{and} \quad u_{ijk}(h)=1 \quad \forall\, h<l$$

5. Estimation

5.1. Proportion in favor of treatment

Generalized pairwise comparisons can be carried out on all pairs of individuals formed with one individual taken from group T and one individual from group C. If there are n individuals in group T and m individuals in group C, there are $n\cdot m$ such pairs. In the presence of stratification, assuming there are K strata (indexed by $k=1,\ldots,K$), $n=\sum_{k=1}^{K} n_k$, $m=\sum_{k=1}^{K} m_k$, and the number of pairs is equal to $\sum_{k=1}^{K} n_k\cdot m_k \leqslant n\cdot m$.

The 'proportion in favor of treatment' (denoted by the letter Δ) is the net difference between the number of favorable pairs and the number of unfavorable pairs divided by the total number of pairs. When interest focuses on a single outcome, Δ can be expressed in terms of the pairwise indicator defined in Section 3.5:

$$\Delta=\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{m_k} p_{ijk}}{\sum_{k=1}^{K} n_k\cdot m_k}. \tag{1}$$

Δ is a signed proportion. It is equal to 1 if the treatment group is uniformly better than the control group, i.e. if all individuals in group T have a better outcome than those in group C (in which case all informative pairs are favorable), to -1 if the control group is uniformly better than the treatment group, and to 0 if there is no net difference between the groups.

5.2. Cumulative proportions in favor of treatment

When prioritized outcomes are considered, the proportion in favor of treatment for the lth outcome ($l=1,\ldots,L$) is given by

$$\delta(l)=\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{m_k} p_{ijk}(l)}{\sum_{k=1}^{K} n_k\cdot m_k}. \tag{2}$$

The cumulative proportion in favor of treatment for the lth outcome or an outcome of higher priority is given by $\Delta(l)=\sum_{i\leqslant l}\delta(i)$. Conventionally, the cumulative proportion in favor of treatment for the outcome of lowest priority, $\Delta(L)$, is simply denoted Δ.

6. Testing

6.1. Randomization test

A randomization test can be used to test the null hypothesis $H_0:\Delta=0$ and to calculate a confidence interval for the observed proportion in favor of treatment, Δ_{obs}. The randomization test requires the simulation of a large number of experiments (say, S) identical to the actual experiment being analyzed, i.e. with all individual data kept unchanged, except their treatment group (treatment or control), which is re-allocated at random [10, 11]. For data arising from a randomized experiment, the re-allocation should use the same algorithm as in the original experiment (e.g. simple randomization, permuted blocks within strata, etc.) [12]. For data not coming from a randomized experiment, the condition required by the randomization test is that the individuals be exchangeable between groups [5]. The re-allocation should therefore keep all experimental design features other than treatment group unchanged. Since individuals are re-allocated at random to either group, the proportion in favor of treatment in the ith simulated experiment, denoted Δ_i, should differ from zero only through the play of chance.

6.2. Confidence interval

The randomization test-based $(1-\alpha)$ per cent confidence interval for Δ_{obs} is calculated as follows using the empirical distribution of Δ under H_0: let $\Delta_{\alpha/2}$ be the value of $\Delta_i (i=1,\ldots,S)$ that leaves at most $\alpha/2$ per cent values of Δ_i to its

left (i.e. $\Delta_i \leqslant \Delta_{\alpha/2}$), and $\Delta_{1-\alpha/2}$ the value that leaves at most $\alpha/2$ per cent values of Δ_i to its right (i.e. $\Delta_{1-\alpha/2} \leqslant \Delta_i$). The $(1-\alpha)$ per cent confidence interval for Δ_{obs} is given by $[\Delta_{\text{obs}}+\Delta_{\alpha/2}; \Delta_{\text{obs}}+\Delta_{1-\alpha/2}]$.

6.3. *Significance probability*

The significance probability (P-value) of the randomization test can be calculated directly from the empirical distribution of Δ. Let s_1 be the number of Δ_i $(i=1,\ldots,S)$ for which $\Delta_i \geqslant \Delta_{\text{obs}}$, and s_2, the number of Δ_i for which $|\Delta_i| \geqslant |\Delta_{\text{obs}}|$. The P-value is equal to s_1/S for a one-sided test, and to s_2/S for a two-sided test.

For very small P-values, the number of simulated experiments becomes prohibitive to obtain a reasonably accurate value with this direct approach. In such cases, it is preferable to base the calculation of the P-value on the asymptotic normality of the empirical distribution of Δ under H_0. The standard deviation of the empirical distribution of Δ is $\sigma = \frac{1}{2}\cdot(\Delta_{1-\alpha/2}-\Delta_{\alpha/2})/Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the $(1-\alpha)$ per cent standardized normal deviate. The P-value is equal to $\Phi(-\Delta_{\text{obs}}/\sigma)$ for a one-sided test, and to $2\cdot\Phi(-\Delta_{\text{obs}}/\sigma)$ for a two-sided test, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

6.4. *Multiple testing of prioritized outcomes*

When prioritized outcomes are considered, cumulative proportions in favor of treatment ($\Delta(l)$, $l=1,\ldots,L$) can be estimated and tested in the same manner as Δ. No assumption of independence is required for the variables capturing the outcomes of interest. In order to preserve the overall significance level of a study, inferences about $\Delta(l)$ should take multiple testing into account. Since $\Delta(l)$ includes pairwise comparisons of the lth outcome and of all higher priority outcomes ($h=1,\ldots,l-1$), successive significance testing of prioritized outcomes is akin to repeated significance testing in group sequential designs [13]. The methods developed for group sequential testing can therefore be used for generalized pairwise comparisons, including flexible α-spending functions [14].

6.5. *Number of simulated experiments*

The number of simulated experiments can be chosen to ensure that the P-value is estimated with a given precision. Specifically, one may wish to ensure that the P-value be correct to the second significant digit (this convention will be used in the analysis of the examples in Section 9). If the true randomization test P-value satisfies the condition $10^{-(k+1)} \leqslant P < 10^{-k}$, the second significant digit of P is the $(k+2)$nd decimal. The minimum number of simulated experiments required to be $(1-\alpha)$ per cent confident that the second significant digit of P differs from its true value by at most 1 unit, is given by $4\cdot(Z_{1-\alpha})^2\cdot 10^{2(k+2)}\cdot P\cdot(1-P)$. Thus, for P close to 0.05, about 730 000 simulated experiments have to be simulated to be 95 per cent confident that the true randomization P-value is above or below 0.05, the level conventionally associated with statistical significance.

7. Relationship with two-sample non-parametric tests

7.1. *Binary variable*

If the variable of interest is binary, the randomization test for Δ is a Monte Carlo approximation to Fisher's exact test if (and only if) the re-allocation to either treatment or control keeps the total number of individuals in each group fixed.

7.2. *Continuous variable*

If the variable of interest is continuous, all pairwise comparisons are a straightforward extension of the Wilcoxon rank-sum test. Recall first how the Wilcoxon rank-sum statistic is constructed [15]. Let us assume for simplicity that there is no stratification. As above, let n and m represent, respectively, the number of non-missing observations in groups T and C. Let $S_1 < S_2 < \cdots < S_n$ be the ordered ranks of the observations in group T. The Wilcoxon rank-sum statistic is

$$W_S = \sum_{i=1}^{n} S_i$$

and the Mann–Whitney form of the statistic is

$$W_{MW} = W_S - n\cdot(n+1)/2.$$

Tables of critical values of W_{MW} can be used for small sample sizes ($n \leqslant 10$ and $m \leqslant 10$), while for large sample sizes, a normal approximation can be used.

Now the Mann–Whitney statistic W_{MW} can also be obtained from all pairwise comparisons. Denote X and Y the variables of interest, taking values $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ in groups T and C, respectively, and consider all possible pairs (X_i, Y_j) consisting of one observation from group T and one from group C. It can be shown (see, e.g. [16]) that the Mann–Whitney statistic W_{MW} is equal to the number of pairs (X_i, Y_j) with $X_i > Y_j$ plus half the number of pairs with $X_i = Y_j$. It follows that W_{MW} can be expressed as

$$W_{MW} = \tfrac{1}{2} \cdot n \cdot m \cdot (1 - \Delta).$$

Hence Δ is a linear transformation of W_{MW}. Δ is in fact the U-statistic for the Wilcoxon test

$$U = \Delta = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}, \tag{3}$$

where

$$U_{ij} = \begin{cases} +1 & \text{if } X_i > Y_j \\ -1 & \text{if } X_i < Y_j \\ 0 & \text{otherwise.} \end{cases}$$

The randomization test for Δ is a Monte Carlo approximation to the exact Wilcoxon test.

7.3. *Time-to-event variable*

Gehan [17] extended the Wilcoxon test to account for censoring of time-to-event variables. The U-statistic for Gehan's generalized Wilcoxon test is given by (3) with

$$U_{ij} = \begin{cases} +1 & \text{if } X_i > Y_j \text{ or } X'_i > Y_j \\ -1 & \text{if } X_i < Y_j \text{ or } X_i < Y'_j \\ 0 & \text{otherwise} \end{cases}$$

with X'_i and Y'_j representing censored observations, as in Section 3.3.

Efron [18] pointed out that the expectation of Gehan's U-statistic under the alternative hypothesis depended on the censoring distribution, and that the test was only valid under the assumption that the censoring distributions were equal in the treatment and the control group [19]. This assumption of equal censoring is subsumed by the exchangeability of individuals between the treatment and control groups, a condition required for the validity of the randomization test for Δ, which is a Monte Carlo approximation to the exact Gehan's generalized Wilcoxon test. This test has fallen out of favor in many applications because it is less efficient than the logrank test and Cox's regression model under the assumption of proportional hazards [20, 21].

8. Relationship with measures of treatment effect

8.1. *Binary variable*

For a binary variable, the most straightforward measure of treatment effect is the absolute risk difference, $p_T - p_C$, where p_T and p_C are the probabilities of success in the treatment and control groups, respectively. Letting $P\{\cdot\}$ denote the probability,

$$\begin{aligned} \Delta &= P\{X=1, Y=0\} - P\{X=0, Y=1\} \\ &= p_T \cdot (1 - p_C) - (1 - p_T) \cdot p_C = p_T - p_C. \end{aligned} \tag{4}$$

8.2. *Continuous variable*

For a continuous variable, a general measure of treatment effect is the standardized mean difference, also called Cohen's d or 'effect size' [22]. When the variable is normally distributed with common variance in the two groups [23, 24],

$$\Delta = 2 \cdot \Phi(d/\sqrt{2}) - 1. \tag{5}$$

8.3. Time-to-event variable

For a time-to-event variable, the most commonly used measure of treatment effect is the hazard ratio (λ). Δ can be expressed as a function of λ and the proportion of informative pairs (f) [25, 26]:

$$\Delta = f \cdot \frac{1-\lambda}{1+\lambda}. \tag{6}$$

8.4. Δ as a general measure of treatment effect

A large body of literature has been recently devoted to measures of treatment effect that do not depend on the type of variable considered [24, 27]. Δ is one such measure, and is closely connected to the 'probabilistic index', denoted $P(X>Y)$, defined as the probability that an individual taken randomly from the treatment group has a better outcome than an individual taken randomly from the control group [28]. Δ is a linear transformation of $P(X>Y)$:

$$\Delta = 2 \cdot P(X>Y) - 1 \tag{7}$$

and these two measures of treatment effect are therefore strictly equivalent.

One advantage of Δ over $P(X>Y)$ may be its easier interpretation. For instance, $P(X, Y)=0.5$ would be interpreted as meaning that the experiment provides no evidence that T differs from C in either direction. This situation would correspond to $\Delta=0$, which is a more direct and intuitively obvious way of expressing the (lack of) treatment benefit. Moreover, the cumulative proportions in favor of treatment for various thresholds, times of measurement, or other prioritized outcomes can help interpret any differences between the groups being compared, as will become evident in the analysis of the case studies in the next Section.

9. Analysis of case studies

9.1. Advanced colorectal cancer

Recall first that the logrank test failed to show a significant survival benefit of the addition of oxaliplatin to LV5FU2 (Figure 2). The authors of the paper also reported that Gehan's generalized Wilcoxon test just reached significance ($P=0.05$) [3]. This result is confirmed through pairwise comparisons ($\Delta=10.1$ per cent, $P=0.05$, top panel of Table V), which additionally show that pairwise differences in times to death exceed one year in 4.4 per cent of pairs ($P=0.043$), and 6 months in 8.3 per cent of pairs ($P=0.038$). Although these cumulative differences would not reach statistical significance after proper adjustment for multiplicity (see Section 6.4), they provide useful information not given by standard two-sample non-parametric statistics.

The logrank test did show a significant benefit of the addition of oxaliplatin to LV5FU2 on progression-free survival (Figure 1). The proportion of pairs with a longer time to progression is also highly significant ($\Delta=24.2$ per cent, $P<0.0001$, second panel of Table V). The pairwise difference, however, exceeds one year in only 1.5 per cent of pairs ($P=0.09$). Again, these statistics shed additional light on the magnitude of the treatment benefit on the classical endpoint of progression-free survival.

Table V. Generalized pairwise comparisons in the advanced colorectal cancer trial (44 100 pairs without stratification).

Differences in	Oxaliplatin better (a) (per cent)	Standard better (b) (per cent)	$\delta(l)$ (a–b) (per cent)	$\Delta(l)=\sum_{i\leqslant l}\delta(i)$ (per cent)	95 per cent C.I. for $\Delta(l)$* (per cent)	Test for $\Delta(l)$*
times to death:						
at least 12 months	10.8	6.5	4.4	4.4	(0.1; 8.5)	$P=0.043$
between 6 and 12 months	14.7	10.8	3.9	8.3	(0.4; 16.1)	$P=0.038$
less than 6 months	17.0	15.2	1.8	10.1	(0.0; 20.1)	$P=0.050$
times to progression:						
at least 12 months	2.6	1.1	1.5	1.5	(−0.2; 3.3)	$P=0.090$
between 6 and 12 months	15.5	5.4	10.1	11.6	(6.4; 16.9)	$P<0.0001$
less than 6 months	35.5	22.9	12.6	24.2	(14.1; 34.4)	$P<0.0001$
prioritized outcomes:						
time to death	42.6	32.5	10.1	10.1	(0.0; 20.1)	$P=0.050$
time to progression	9.1	4.4	4.7	14.8	(4.4; 25.2)	$P=0.0054$

*Unadjusted for multiplicity.

Table VI. Generalized pairwise comparisons in the age-related macular degeneration trial (12 907 pairs with stratification for region, baseline visual acuity, type and size of macular lesion, and prior photodynamic therapy).

Differences in	Pegaptanib better (a) (per cent)	Sham better (b) (per cent)	$\delta(l)$ (a–b) (per cent)	$\Delta(l)=\sum_{i\leqslant l}\delta(i)$ (per cent)	95 per cent C.I. for $\Delta(l)^*$ (per cent)	Test for $\Delta(l)^*$
loss of $\geqslant$15 *letters at week 54*:						
proportion	29.6	18.5	11.1	11.1	(2.7; 19.4)	$P=0.0095$
change in vision from baseline to week 54:						
$\geqslant$30 letters	11.4	4.9	6.5	6.5	(2.5; 10.5)	$P=0.0013$
$\geqslant$25 letters	4.4	2.6	1.8	8.3	(3.3; 13.3)	$P=0.0011$
$\geqslant$20 letters	5.2	3.1	2.1	10.4	(4.4; 16.4)	$P=0.0007$
$\geqslant$15 letters	6.2	4.0	2.2	12.6	(5.6; 19.6)	$P=0.0005$
$\geqslant$10 letters	7.9	5.6	2.3	14.9	(6.7; 23.0)	$P=0.0003$
$\geqslant$5 letters	8.6	7.2	1.4	16.3	(7.1; 25.4)	$P=0.0005$
$<$5 letters	9.5	8.7	0.8	17.1	(7.3; 26.9)	$P=0.0007$
any change in vision from baseline to last available visit:						
at week 54	48.0	35.0	13.0	13.0	(4.2; 21.7)	$P=0.0036$
at week 48	3.1	1.0	2.1	15.1	(6.1; 24.0)	$P=0.0010$
at week 42	1.1	0.8	0.2	15.3	(6.2; 24.3)	$P=0.0009$
at week 36	2.5	1.0	1.5	16.8	(7.7; 25.9)	$P=0.0003$
at week 30	0.8	1.00	−0.2	16.6	(7.4; 25.8)	$P=0.0004$
at week 24	0.9	0.6	0.3	16.9	(7.7; 26.1)	$P=0.0003$
at week 18	0.8	0.3	0.4	17.3	(7.9; 26.8)	$P=0.0003$
at week 12	1.2	0.4	0.8	18.1	(8.5; 27.7)	$P=0.0002$
at week 6	0.3	0.1	0.2	18.3	(8.8; 28.3)	$P=0.0002$

*Unadjusted for multiplicity.

Using generalized pairwise comparisons on the prioritized outcomes of time to death and time to progression, Δ is 14.8 per cent in favor of oxaliplatin ($P=0.0054$), 10.1 per cent in terms of survival, and another 4.7 per cent in terms of time to progression (third panel of Table V). Generalized pairwise comparisons of times to death or progression are arguably most relevant, since time to progression is used for patients who have not yet died, and as such this analysis makes use of all the available information on important time-related events.

9.2. Age-related macular degeneration

Recall that the proportion of patients losing at least 15 letters of visual acuity was 35 per cent for patients receiving pegaptanib vs 45 per cent for patients receiving sham injections [4]. This result is confirmed through pairwise comparisons without stratification (87 616 pairs, $\Delta=9.8$ per cent, $P=0.015$), using the binary endpoint of loss of $\geqslant$15 letters of visual acuity at week 54. Pairwise comparisons can also be stratified for the factors used in the published analyses (12 907 pairs, $\Delta=11.1$ per cent, $P=0.0095$, top panel of Table VI).

When vision changes are considered a continuous endpoint at week 54, the proportion in favor of treatment, considering a difference of at least 15 letters of visual acuity to be relevant, is 12.6 per cent (second panel of Table VI), which is slightly larger than the difference between the proportions of patients losing at least 15 letters of visual acuity at week 54 (11.1 per cent), but far more significant ($P=0.0005$). Table VI shows cumulative proportions in favor of treatment for other thresholds that might be of interest, and the corresponding P-values. These are extreme enough to reach significance for all thresholds, even after allowance for multiple testing (Section 6.4).

This analysis at week 54 uses the same convention as the published analysis, i.e. the last observation is carried forward for any missing measurement at week 54. Using pairwise comparisons, such data imputation is not required (third panel of Table VI). The proportion in favor of treatment is equal to 18.3 per cent overall, with about two-thirds of this effect (13.0 per cent) based on observations at week 54.

10. Discussion

Generalized pairwise comparisons offer an alternative approach to standard non-parametric tests for the two-sample problem. With this method, there is no need to create artificial 'endpoints' that lack either relevance or statistical efficiency. In the case study in ophthalmology, for instance, the loss of 15 letters of visual acuity, which is the 'endpoint' required for regulatory approval of new drugs for macular degeneration, suffers from several drawbacks, including poor efficiency, misclassification of the outcome, and potential for a ceiling or floor effect [29]. Mean changes in visual

acuity, evaluated on a continuous scale, would have better statistical sensitivity, but might reach statistical significance for changes in visual acuity that do not reflect worthwhile clinical differences. The traditional way of avoiding this drawback is to impose that the *mean* difference between the randomized groups exceeds a certain threshold (see, e.g. example 1.3 in [30]), but this may be putting the bar too high if some patients do not derive much benefit from treatment, while others do. Using generalized pairwise comparisons, it is easy to define a threshold for differences in vision that clearly defines a better outcome (such as 15 letters of visual acuity), and estimate the proportion in favor of treatment beyond that threshold (Table VI).

Generalized pairwise comparisons do not require data imputation for patients with missing data. Longitudinal models could also be used to analyze these data without imputation [31]. Either approach would be preferable to the naive imputation that was used by carrying the last observation forward [4]. Although such methods of data imputation are well known to result in misleading inferences, they are still commonly used on account of their simplicity [32, 33]. The impact of different types of missing data on the validity of generalized pairwise comparisons is a topic for further research [34].

Generalized pairwise comparisons may prove especially valuable when several variables need to be simultaneously considered, as illustrated in the case study in advanced cancer. The traditional methods of analysis either consider each variable separately, or define a composite 'endpoint' that captures all relevant events in a single variable. In advanced cancer, for example, progression-free survival is defined as the time to disease progression or death, whichever occurs first. A clear drawback of this variable is that it ignores the time to death after disease progression, and thus might favor a treatment that prolonged time to progression but shortened time to death. In contrast, generalized pairwise comparisons use either time to progression or time to death, depending on the variable of highest priority available for each pairwise comparison (hence different variables may be used for the same individual in different pairwise comparisons). In this way, the method makes full use of the information available for all clinically important events, thus leading to a potentially more informative evaluation of new drugs [35]. The respective priorities of the different variables depend on the situation at hand, and may be a matter of debate. Prospectively deciding on the respective priorities of different variables capturing outcomes of interest is not only a major advantage, but also a potential difficulty, of generalized pairwise comparisons. Cumulative proportions in favor of treatment allow the contribution of each variable to be estimated separately, using their pre-defined priority level. Not all contributing variables need to be of the same type [9]. Hence, time-to-event variables such as time to disease worsening can be combined with binary variables such as toxicities, or continuous variables such as quality of life scores. Traditional statistical tests based on exact or asymptotic probability distributions quickly become analytically intractable for several variables of different types, or in high dimensional spaces. Morales *et al.* [36] use pairwise comparisons to define multivariate U-statistics (which they call μ-scores) to analyze the relationship between high dimensional phenotypic data and genetic mutations in patients with Fanconi anemia.

The proportion in favor of treatment, Δ, is a general measure of difference between the groups being compared. When a single outcome measure is considered, Δ is related to traditional measures of effect associated with each type of variable considered (equations (4) to (6)). For a binary variable, Δ is equal to the absolute risk difference between the two groups. For a continuous variable, Δ is related to the effect size through the standard normal cumulative distribution function. For a time-to-event variable, Δ is related to the hazard ratio and the proportion of censored observations [26].

Regardless of the type of variable(s) considered, Δ is a linear transformation of the probabilistic index (equation (7)), which has been advocated as a useful measure of treatment effect [24, 25]. Senn [23, 37] has drawn attention to some difficulties with general measures of treatment effect such as Δ, the probabilistic index, or the 'proportion of similar response', a closely related measure of treatment similarity proposed in the context of equivalence testing [38, 39]. All these measures depend on location *and* scale parameters, and as a result their interpretation may not be straightforward. For a binary variable, Δ can be interpreted directly as the proportion of individuals improved by treatment (Section 8.1). For other types of outcome measures (e.g. continuous or time-to-event variable), this simple interpretation is not generally valid.

Confidence intervals and tests of significance for Δ have been obtained here through randomization tests [11]. These tests are simple to implement regardless of the complexity of the outcomes considered, but they can be computer-intensive if significance probabilities and confidence intervals are to be calculated with reasonable precision. With efficient re-randomization algorithms in standard statistical packages, generalized pairwise comparisons may open unprecedented possibilities of assessing prioritized outcomes in the two-sample problem.

Acknowledgements

This paper is dedicated to the memory of Stephen W. Lagakos (1946–2009) who provided thoughtful advice on an early draft he had generously accepted to review. The paper has also greatly benefited from the comments of Christian Gluud, Joseph Heyse, Geert Molenberghs, Steven Senn, Kristian Thorlund, Jørn Wetterslev, and Knut Wittkowski. The author is grateful to Eyetech Inc.,

Pfizer Inc., and sanofi~aventis for permission to re-analyze data from clinical trials testing oxaliplatin and pegaptanib. François Torche of Innovative Minds has developed efficient SAS macros (available upon request) to implement the methods proposed in this paper.

References

1. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 1948; **19**:293–325.
2. Lee AJ. *U-Statistics*. Marcel Dekker Inc.: New York, 1990.
3. de Gramont A, Figer A, Seymour M, Homerin M, Hmissi A, Cassidy J, Boni C, Cortes-Funes H, Cervantes A, Freyer G, Papamichael D, Le Bail N, Louvet C, Hendler D, de Braud F, Wilson C, Morvan F, Bonetti A. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *Journal of Clinical Oncology* 2000; **18**:2938–2947.
4. Gragoudas ES, Adamis AP, Cunningham ET, Feinsod M, Guyer DR, The VEGF Inhibition Study in Ocular Neovascularization Clinical Trial Group. Pegaptanib for neovascular age-related macular degeneration. *New England Journal of Medicine* 2004; **351**:2805–2816.
5. Good P. Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods* 2002; **1**:243–247.
6. Wittkowski KM, Lee E, Nussbaum R, Chamian FN, Krueger JG. Combining several ordinal measures in clinical studies. *Statistics in Medicine* 2004; **23**:1579–1592.
7. Gould L. A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics* 1980; **36**:721–727.
8. Moyé LA, Davis BR, Hawkins CM. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Statistics in Medicine* 1992; **11**:1705–1717.
9. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine* 1999; **18**:1341–1354.
10. Good P. *Resampling Methods*: *A Practical Guide to Data Analysis* (3rd edn). Birkhauser: New York, 2006.
11. Edgington ES, Onghena P. *Randomization Tests*. Chapman & Hall/CRC: New York, 2007.
12. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials*: *Theory and Practice*. Wiley: New York, 2002.
13. Jennison V, Turnbull B. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: New York, 2000.
14. Lan KK, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
15. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945; **1**:80–83.
16. Lehmann EL. *Nonparametrics*: *Statistical Methods Based on Ranks* (Revised edn). Springer: New York, 2006.
17. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965; **52**:203–223.
18. Efron B. The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium*, vol. 4. University of California Press: Berkeley, CA, 1965; 831–853.
19. Desu MM, Raghavarao D. *Nonparametric Statistical Methods for Complete and Censored Data*. Chapman & Hall/CRC: New York, 2004.
20. Peto R, Peto J. Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society*, *Series A* 1972; **135**:185–207.
21. Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
22. Glass GV. Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; **5**:3–8.
23. Senn SJ. Testing for individual and population equivalence based on the proportion of similar responses (Letter and Authors' Reply). *Statistics in Medicine* 1997; **16**:1303–1306.
24. Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* 2006; **25**:591–602.
25. Moser BK, McCann MH. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clinical Trials* 2008; **5**:248–252.
26. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators (Letter). *Clinical Trials* 2008; **5**:641–642.
27. Hauck WW, Hyslop T, Anderson S. Generalized treatment effects for clinical trials. *Statistics in Medicine* 2000; **19**:887–899.
28. Grissom R. Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology* 1994; **79**:314–316.
29. Beck RW, Maguire MG, Bressler NM, Glassman AR, Lindblad AS, Ferris FL. Visual acuity as an outcome measure in clinical trials of retinal diseases. *Ophthalmology* 2007; **114**:1804–1809.
30. Sprent P, Smeeton NC. *Applied Nonparametric Statistical Methods* (4th edn). Chapman & Hall/CRC: New York, 2007.
31. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Wiley: Hoboken, NJ, 2007.
32. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**:445–464.
33. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials* 2004; **1**:368–376.
34. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
35. Schilsky RL. Endpoints in cancer clinical trials and the drug approval process. *Clinical Cancer Research* 2002; **8**:935–938.
36. Morales JF, Song TT, Auerbach AD, Wittkowski KM. Phenotyping genetic diseases using an extension of μ-scores for multivariate data. *Statistical Applications in Genetics and Molecular Biology* 2008; **7**:19.
37. Senn SJ. Probabilistic index: an intuitive non-parametric approach to measuring the size of the treatment effects (Letter and Authors' Reply). *Statistics in Medicine* 2006; **25**:3944–3948.
38. Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine* 1996; **15**:1489–1505.
39. Stine RA, Heyse JF. Non-parametric estimates of overlap. *Statistics in Medicine* 2001; **20**:215–236.