LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Monika Jelizarow, Alarcos Cieza, Ulrich Mansmann

Global permutation tests for multivariate ordinal data: alternatives, test statistics, and the null dilemma

Technical Report Number 142, 2013
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Global permutation tests for multivariate ordinal data: alternatives, test statistics, and the null dilemma

Monika Jelizarow[1], Alarcos Cieza[1,3], Ulrich Mansmann[1,2]

[1] *Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians University, Munich, Germany*

[2] *Department of Statistics, Ludwig-Maximilians University, Munich, Germany*

[3] *Faculty of Social and Human Sciences, School of Psychology, University of Southampton, UK*

Corresponding author: Monika Jelizarow, e-mail: jelizarow@ibe.med.uni-muenchen.de

Abstract. We discuss two-sample global permutation tests for sets of multivariate ordinal data in possibly high-dimensional setups, motivated by the analysis of data collected by means of the World Health Organisation's International Classification of Functioning, Disability and Health. The tests do not require any modelling of the multivariate dependence structure. Specifically, we consider testing for marginal inhomogeneity and direction-independent marginal order. Max-T test statistics are known to lead to good power against alternatives with few strong individual effects. We propose test statistics that can be seen as their counterparts for alternatives with many weak individual effects. Permutation tests are valid only if the two multivariate distributions are identical under the null hypothesis. By means of simulations, we examine the practical impact of violations of this exchangeability condition. Our simulations suggest that theoretically invalid permutation tests can still be 'practically valid'. In particular, they suggest that the degree of the permutation procedure's failure may be considered as a function of the difference in group-specific covariance matrices, the proportion between group sizes, the number of variables in the set, the test statistic used, and the number of levels per variable.

Keywords: Global test; Hotelling-type statistic; ICF; Marginal inhomogeneity; Marginal order; Multivariate ordinal data; Non-exchangeability; Permutation test; Sum statistic

1. Introduction

Two-group comparisons of multivariate ordinal data constitute an important problem in statistical practice. The present work has primarily been motivated by the need for methodology that adequately addresses such problems with data collected by means of the International Classification of Functioning, Disability and Health (ICF) (World Health Organisation, 2001; Ustün et al., 2003). The ICF was endorsed by the 54th World Health Assembly in 2001 with the aim of providing a unified classification framework that allows for the description of functioning and disability both across health conditions and for specific health conditions such as depression, obesity, and stroke. Going beyond a purely biomedical approach, it takes into account individual, social, and environmental aspects of functioning and disability. The ICF comprises over 1400 health-related ordinally scaled items called ICF categories. ICF core sets are health condition-specific selections from the overall pool of ICF categories, defined by health experts (e.g. physicians and physiotherapists) at international ICF consensus conferences. ICF core sets thus facilitate the

implementation of the ICF in clinical practice and research (Stucki and Grimby, 2004). Each ICF category belongs to one of the four sets below, the so-called ICF components:

(b) body functions (coded with b, e.g. b144: 'memory functions'),
(s) body structures (coded with s, e.g. s410: 'structure of cardiovascular system'),
(d) activities and participation (coded with d, e.g. d640: 'doing housework'),
(e) environmental factors (coded with e, e.g. e310: 'immediate family').

The coding scheme for ICF categories in b, s, and d is 0 (no impairment), 1 (mild impairment), 2 (moderate impairment), 3 (severe impairment), and 4 (complete impairment). The coding scheme for ICF categories in e is -4 (complete barrier), -3 (severe barrier), -2 (moderate barrier), -1 (mild barrier), 0 (neither barrier nor facilitator), 1 (mild facilitator), 2 (moderate facilitator), 3 (severe facilitator), and 4 (complete facilitator), yet this scale is usually coarsened to just five levels in practice. ICF categories in the sets b, s, d, and e can in turn be divided further into more specific sets called ICF chapters (e.g. the ICF category 'memory functions' (b144) belongs to the ICF chapter 'mental functions' (b1)), resulting in a tree structure with different levels of detail.

We consider data from a multicentre study based on the ICF core set for stroke which comprises $p = 130$ ICF categories (Geyh et al., 2004). (For the complete list and information on the ICF chapters involved see online supplement A.) Fig. 1 sketches its three-level tree structure; the most detailed fourth level considering an individual ICF category as a set with cardinality 1 is omitted for ease of visualisation. The dataset includes $n = 104$ patients after first stroke of which $n_1 = 46$ underwent rehabilitation in Asian countries and $n_2 = 58$ in European countries. (For more information on the dataset see Section 6.) The question of interest is whether and, most notably, *where* stroke patients from Asian versus European countries differ in their 130-dimensional ICF pattern. Besides that, two-group comparisons of ICF patterns have been the major objective of numerous other ICF studies conducted worldwide (Holper et al., 2010; Herrmann et al., 2011; Tschiesner et al., 2011). As all 191 member states of the World Health Organisation (WHO) have agreed to use the ICF in clinical practice and research and many have already started, it is expected by the WHO that the number of such ICF studies will rapidly increase over the years to come. Since, from the statistical viewpoint, they pose comparable challenges, our ICF stroke study shall be considered as an example.

A typical way to tackle such two-group problems is to conduct a univariate test for each ICF category (e.g. Pearson chi-squared test) and then adjust the univariate p-values for multiplicity such that the Familywise Error Rate (FWER) is controlled at the prespecified level α (e.g. using the Bonferroni procedure or the less conservative procedures of Holm (1979), Hochberg (1988), or Hommel (1988)). While it is simple to use, this approach has potentially low power in the complex data situation we consider, both because the multiplicity penalty becomes rather severe when p is large and because it ignores the obvious dependence between many ICF categories (e.g. 'memory functions' (b144) and 'attention functions' (b140)). Power and interpretability could be enhanced if we were able to exploit the prior information on the datas' structure inferentially. For instance, it may be worthwhile and meaningful to perform the statistical analysis at the broader level of ICF chapters or components. In our ICF example, the Bonferroni penalty would thereby decrease from 130 to 24 and 4, respectively. Alternatively, recent advances in simultaneous inference have made it possible to use the entire tree structure inferentially (Meinshausen, 2008; Goeman and Solari, 2010; Goeman and Finos, 2012). In both cases, the multiplicity adjustment procedure rests upon the availability of a suitable test that provides set-specific p-values. The construction of such global tests is intricate in itself

and becomes particularly challenging when the data are multivariate ordinal. For illustration, let us consider the ICF component 'body functions' (b). Provided that all 41 ICF categories included can take five distinct values, the two 41-way contingency tables which cross-classify the $n_1 = 46$ and $n_2 = 58$ multivariate observations have $5^{41} \approx 4.55 \times 10^{28}$ cells; they are thus very sparse, which does not allow us to consider the full multivariate structure. This shows that test statistics based on the maximum likelihood will be impossible to compute because the maximum likelihood relies on the joint distribution. The situation does not substantially improve if the multivariate ordinal data are dichotomised, aside from the fact that dichotomisation usually results in a loss of information. Another way to reduce the number of parameters involved may be to treat the multivariate ordinal data as multivariate normal. However, normality seems rather questionable in the scenarios we consider, and even if it is assumed, test statistics that take into account the covariances between variables will still not be computable when the data are high-dimensional, such as Hotelling's T^2 which requires the $(p \times p)$ sample covariance matrix to be inverted. For this reason, in the case of multivariate normal data, simple test statistics that dispense with the covariances between variables have become popular (Chung and Fraser, 1958; Ackermann and Strimmer, 2009). The case of multivariate non-normal data has so far received only little attention in the literature. In this paper we use previous results of Agresti and Klingenberg (2005) and Klingenberg et al. (2009) to construct tests based on such simple test statistics for the case of multivariate ordinal data, with the aim to make multiplicity adjustment procedures for structured hypotheses applicable to ICF-based problems.

The paper is structured as follows. In Section 2 we define and discuss the global hypotheses of interest: inhomogeneity and, as a special case, direction-independent stochastic order between the ordinal variables' marginal distributions. Joint distributions are left unspecified. In Section 3 we propose simple test statistics that are sensitive towards the alternative hypotheses from Section 2. In this context we will see that, under working independence between variables, the test statistic of Klingenberg et al. (2009) reduces to the sum of univariate Cochran-Armitage trend test statistics, providing important insight into its power properties. For inference, we focus on the popular permutation procedure. The latter is known to be valid only if the two multivariate distributions are identical under the null hypothesis, which is not the case under the null hypothesis we consider. We address this issue in Section 4 and discuss the 'null dilemma' that arises when no superior inference method is available. In Section 5 we examine, by means of simulations, the permutation procedure's robustness properties under theoretically unfavourable conditions. In Section 6 we analyse the ICF core set data for stroke patients and illustrate the practical benefits of the proposed methodology.

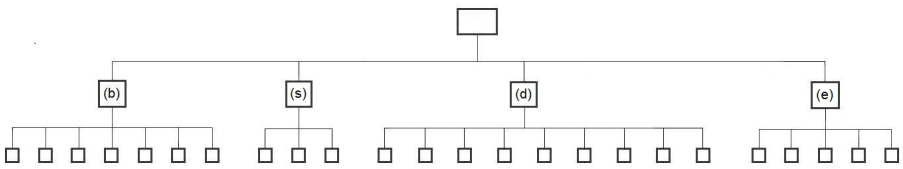

Figure 1. Tree-structure of the ICF core set for stroke. Sets on the same level do not overlap.

2. Global hypotheses about marginal distributions

2.1. Marginal Inhomogeneity (MI)

We address the situation in which two independent groups of sizes n_1 and n_2, $n_1 + n_2 = n$, shall be compared based on p-dimensional ordinal data vectors, with all p ordinal variables having the same number $c \geq 2$ of levels. Suppose that the n_i multivariate observations in group i form an i.i.d. sample of a $(p \times 1)$ random vector $\boldsymbol{X}_i = \left(X_{i1}, ..., X_{ip}\right)^\top$ which has a multivariate multinomial distribution $\boldsymbol{\Pi}_i$ with unknown dependence structure, $i = 1,2$. Let $\pi_i(v_1, ..., v_p)$ denote the joint probability $P(X_{i1} = v_1, ..., X_{ip} = v_p)$ for an entire pattern in group i, where $v_k \in \{1, ..., c\}$, $k = 1, ..., p$. Unless further specified when the two groups are considered different, it seems natural to test the null hypothesis $H_0 : \boldsymbol{X}_1 \overset{d}{=} \boldsymbol{X}_2$ against the alternative $H_1 : \boldsymbol{X}_1 \overset{d}{\neq} \boldsymbol{X}_2$, where '$\overset{d}{=}$' means equality in distribution. H_0 (i.e. $\pi_1(v_1, ..., v_p) = \pi_2(v_1, ..., v_p)$ for all c^p possible sequences $(v_1, ..., v_p) \in \{1, ..., c\}^p$) is referred to as Identical Joint Distribution (IJD), and H_1 (i.e. $\pi_1(v_1, \ldots, v_p) \neq \pi_2(v_1, ..., v_p)$ for at least one $(v_1, ..., v_p) \in \{1, ..., c\}^p$) as Non-Identical Joint Distribution (NJD). However, because confirmation of NJD is little informative as to *why* it has been confirmed, in most ICF studies it seems preferable to test the one-way multinomial distributions $\Pi_{ik} = \{\pi_{ik}(v)\}_{v=1}^c$ of the random variables X_{ik}, with $\pi_{ik}(v)$ denoting the marginal probability $P(X_{ik} = v)$, $v \in \{1, ..., c\}$. The associated hypotheses are

$$H_0^m : \bigcap_{k=1}^{p} H_{0k} = \bigcap_{k=1}^{p} \{X_{1k} \overset{d}{=} X_{2k}\}, \tag{1}$$

$$H_1^m : \bigcup_{k=1}^{p} H_{1k} = \bigcup_{k=1}^{p} \{X_{1k} \overset{d}{\neq} X_{2k}\}, \tag{2}$$

where the intersection null hypothesis H_0^m in Eq. 1 (i.e. $\{\pi_{1k}(v)\}_{v=1}^c = \{\pi_{2k}(v)\}_{v=1}^c$ simultaneously for all k) is referred to as Simultaneous Marginal Homogeneity (SMH), and H_1^m in Eq. 2 (i.e. $\{\pi_{1k}(v)\}_{v=1}^c \neq \{\pi_{2k}(v)\}_{v=1}^c$ for at least one k) as Marginal Inhomogeneity (MI). For $c = 2$, this problem was tackled by Agresti and Klingenberg (2005). Evidently, IJD $\Rightarrow$ SMH (i.e. IJD is more restrictive than SMH). For $p = 1$, IJD and SMH are equivalent. We come back to the distinction between both null hypotheses and its importance in permutation-based inference in Section 4.

2.2. Marginal Order (MO)

As in our ICF stroke study, it is the primary aim in many other ICF studies to detect MI. In some instances, however, the information provided under MI may be too unspecific and the research question may focus on special cases of MI. In the ICF context, the most important special case of MI is marginal stochastic order. The random variables X_{1k} and X_{2k} are stochastically ordered if either (i) $P(X_{1k} \leq v) \geq P(X_{2k} \leq v)$, written $X_{1k} \preceq X_{2k}$, or (ii) $P(X_{1k} \leq v) \leq P(X_{2k} \leq v)$, written $X_{1k} \succeq X_{2k}$, for all $v \in \{1, ..., c\}$. Without loss of generality, if the inequality in (i) is strict for at least one v, X_{1k} and X_{2k} are said to be stochastically strictly ordered, written $X_{1k} \prec X_{2k}$. Let the narrower alternative be

$$\widetilde{H}_1^m : \bigcup_{k=1}^{p} \widetilde{H}_{1k} = \bigcup_{k=1}^{p} \{\{X_{1k} \prec X_{2k}\} \cup \{X_{1k} \succ X_{2k}\}\}, \tag{3}$$

where $\{X_{1k} \prec X_{2k}\}$ and $\{X_{1k} \succ X_{2k}\}$ are mutually exclusive for all k. Under $\widetilde{H}_1^m$ in Eq. 3, we thus have either $P(X_{1k} \leq v) > P(X_{2k} \leq v)$ or $P(X_{1k} \leq v) < P(X_{2k} \leq v)$ for at least one

k and v, and we shall refer to this two-sided alternative as Marginal Order (MO), noting that MO $\Rightarrow$ MI. The one-sided counterpart (i.e. $\bigcup_{k=1}^{p}\{X_{1k} \prec X_{2k}\}$) was tackled by Klingenberg et al. (2009), motivated by the statistical analysis of ordinally scaled adverse effects data from toxicity studies. Here it is plausible to assume that, for all adverse effects, there is equal or greater chance of observing severe effects (i.e. high levels) in the treatment group (group 2) than in the placebo group (group 1). For ICF studies, a similar assumption will be rarely plausible. In our particular ICF example, for instance, some body functions may be more severely impaired among Asian patients than among European patients, while the opposite holds for other functions. Because we are usually equally interested in '$X_{1k} \prec X_{2k}$' and '$X_{1k} \succ X_{2k}$' contributions to set effects, it is sensible to consider the direction-independent stochastic order alternative MO. Compared to MI, it is the more appropriate choice if we wish to explicitly take into account the natural ordering of the c levels. If this is not essential in the application at hand, it seems reasonable to choose MI which is broader in the sense that it includes but is not restricted to stochastically ordered one-way multinomial distributions. Given that the problems 'SMH against MI' and 'SMH against MO' are closely related and similarly widespread in ICF-based applications, both are discussed in the present paper, and the former is exemplified in Section 6.

3. Global test statistics

3.1. Testing for MI

To test for MI in the case $c = 2$, Agresti and Klingenberg (2005) proposed a test statistic that is a quadratic form in the vector of differences in sample means. We shall see below that their test statistic can easily be generalised to the case $c \geq 2$, even though in most practical situations it will not be computable without additional assumptions on the covariance structure between variables. Let $n_{ik}(v)$ be the number of subjects with observed level v of the kth variable in group i, with respective sample proportion $\widehat{\pi}_{ik}(v) = \frac{n_{ik}(v)}{n_i}$. As $\widehat{\pi}_{ik}(c) = 1 - \sum_{v=1}^{c-1} \widehat{\pi}_{ik}(v)$, the truncated $((c-1)p \times 1)$ vector of marginal sample proportions (i.e. sample means) for group i is $\widehat{\boldsymbol{\pi}}_i = \left(\widehat{\pi}_{i1}(1), ..., \widehat{\pi}_{i1}(c-1), ..., \widehat{\pi}_{ip}(1), ..., \widehat{\pi}_{ip}(c-1)\right)^\top$. Let $\boldsymbol{d} = \widehat{\boldsymbol{\pi}}_2 - \widehat{\boldsymbol{\pi}}_1$ denote the vector of differences in marginal sample proportions with entries $d_k(v) = \widehat{\pi}_{2k}(v) - \widehat{\pi}_{1k}(v)$. From basic multinomial theory it is known that $\mathbb{E}(\boldsymbol{d}) = \boldsymbol{\pi}_2 - \boldsymbol{\pi}_1$, and that the $((c-1)p \times (c-1)p)$ covariance matrix $\mathrm{Cov}(\boldsymbol{d}) = \boldsymbol{\Sigma}$ has the entries

$$\mathrm{Var}(d_k(v)) = \sum_{i=1}^{2} \frac{\pi_{ik}(v)\,(1 - \pi_{ik}(v))}{n_i}, \tag{4}$$

$$\mathrm{Cov}(d_k(v), d_k(\tilde{v}))_{v \neq \tilde{v}} = -\sum_{i=1}^{2} \frac{\pi_{ik}(v)\pi_{ik}(\tilde{v})}{n_i}, \tag{5}$$

$$\mathrm{Cov}(d_k(v), d_{\tilde{k}}(\tilde{v}))_{k \neq \tilde{k}} = \sum_{i=1}^{2} \frac{\pi_{ik\tilde{k}}(v, \tilde{v}) - \pi_{ik}(v)\pi_{i\tilde{k}}(\tilde{v})}{n_i}. \tag{6}$$

A test statistic sensitive towards MI can now be constructed as the simple quadratic form $\boldsymbol{d}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{d}$, with $\widehat{\boldsymbol{\Sigma}}$ being the sample version of $\boldsymbol{\Sigma}$. As becomes apparent from Eq. 4 and 5, the variances and covariances within variables can easily be estimated from the sample proportions $\widehat{\pi}_{ik}(v)$. Under the null hypothesis SMH, we can pool the data to obtain the more efficient pooled estimator $\widehat{\boldsymbol{\pi}}_0$ with entries $\widehat{\pi}_{0k}(v) = \frac{n_{1k}(v) + n_{2k}(v)}{n_1 + n_2}$. The covariances between variables from Eq. 6, however, depend on the two-way multinomial distributions $\Pi_{ik\tilde{k}} = \left\{\pi_{ik\tilde{k}}(v, \tilde{v})\right\}_{v,\tilde{v}=1}^{c}$, where $\pi_{ik\tilde{k}}(v, \tilde{v}) = P(X_{ik} = v, X_{i\tilde{k}} = \tilde{v})$, $k \neq \tilde{k}$. Their estimation proves to be problematic. Firstly, when we pool the data for this purpose, we additionally

assume that the two groups have the same $\binom{p}{2}$ two-way multinomial distributions under the null hypothesis, which is more restrictive than SMH. This assumption was made by Agresti and Klingenberg (2005), rendering their test statistic an analogue of Hotelling's T^2 for multivariate binary data. Secondly, even pooled data usually lead to sparse two-way contingency tables unless n is very large and/or c is small. A Hotelling-type approach along the lines of Agresti and Klingenberg (2005) is therefore bound to fail in most ICF-based applications. For instance, provided that $c = 5$ for all ICF categories in our ICF stroke study, 8318 of the $\binom{130}{2} = 8385$ (5×5) tables have one or more empty cells, rendering numerous $\pi_{0k\tilde{k}}(v, \tilde{v})$s inestimable. As a result, we may obtain an estimate of $\boldsymbol{\Sigma}$ that is not positive definite. To prevent this, it seems inevitable to considerably simplify the covariance structure between variables. We prefer to assume independence, which results in an estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_0$ that is block-diagonal. The kth null-estimated $((c-1) \times (c-1))$ block and its inverse are given by $\widehat{\boldsymbol{\Sigma}}_{0k} = \frac{n_1+n_2}{n_1 n_2}\left(\text{diag}(\widehat{\boldsymbol{\pi}}_{0k}) - \widehat{\boldsymbol{\pi}}_{0k}\widehat{\boldsymbol{\pi}}_{0k}^\top\right)$ and $\widehat{\boldsymbol{\Sigma}}_{0k}^{-1} = \frac{n_1 n_2}{n_1+n_2}\left(\text{diag}(\widehat{\boldsymbol{\pi}}_{0k})^{-1} + \widehat{\pi}_{0k}(c)^{-1}\mathbf{1}\mathbf{1}^\top\right)$, respectively, where $\mathbf{1}$ is a $((c-1) \times 1)$ vector of ones. Then, the quadratic form can be written as

$$S = \sum_{k=1}^{p} \boldsymbol{d}_k^\top \widehat{\boldsymbol{\Sigma}}_{0k}^{-1} \boldsymbol{d}_k, \tag{7}$$

which is the sum of variable-specific test statistics. It can readily be verified that the p summands are equivalent to marginal Pearson chi-squared test statistics (of homogeneity), each with an asymptotic chi-squared null distribution with df $= c - 1$, so we shall refer to S as chi-squared sum statistic. Under independence between variables, this chi-squared sum statistic has an asymptotic chi-squared null distribution with df $= p(c-1)$. However, independence rarely holds and is particularly questionable in our context where variables from the same set describe more similar aspects than variables from different sets. As mentioned earlier, we will therefore turn our attention to null distributions derived via the permutation procedure which accounts for the dependence between variables by resampling entire multivariate observations. See Section 4 for further details.

3.2. Testing for MO

To construct a test statistic that is sensitive towards MO, we can exploit the results from Section 3.1. Let $\widehat{\boldsymbol{\pi}}_i' = (\widehat{\pi}_{i1}(1), \ldots, \widehat{\pi}_{i1}(c), \ldots\ldots, \widehat{\pi}_{ip}(1), \ldots, \widehat{\pi}_{ip}(c))^\top$ denote the non-truncated $(cp \times 1)$ vector of marginal sample proportions for group i, and be $\boldsymbol{d}' = \widehat{\boldsymbol{\pi}}_2' - \widehat{\boldsymbol{\pi}}_1'$. In order to take into account the variables' ordinal nature, we can multiply $\boldsymbol{d}'$ with a $(p \times cp)$ matrix $\boldsymbol{U} = \text{diag}\left(\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_p^\top\right)$, where $\boldsymbol{u}_k^\top = (u_k(1), \ldots, u_k(c))$ contains monotonically increasing scores alloted to the c levels of the kth variable. This results in $\boldsymbol{s} = \boldsymbol{U}\boldsymbol{d}'$, which is the $(p \times 1)$ vector of mean score differences with covariance matrix $\text{Cov}(\boldsymbol{s}) = \boldsymbol{\Delta} = \boldsymbol{U}\text{Cov}(\boldsymbol{d}')\boldsymbol{U}^\top$. It is sensible to estimate $\text{Cov}(\boldsymbol{d}')$ under SMH based on the pooled $\widehat{\boldsymbol{\pi}}_0'$ and, eventually for the same reasons outlined in Section 3.1, the assumption of independence between variables. Then, the estimated $(p \times p)$ covariance matrix $\widehat{\boldsymbol{\Delta}}_0$ is block-diagonal, and the kth null-estimated block is given by the scalar $\widehat{\delta}_{0k} = \boldsymbol{u}_k^\top \widehat{\text{Cov}}(\boldsymbol{d}_k')\boldsymbol{u}_k$. To test for the one-sided counterpart of MO (i.e. $\bigcup_{k=1}^{p}\{X_{1k} \prec X_{2k}\}$) in multivariate ordinal data, Klingenberg et al. (2009) employed the test statistic $S' = p^{-1}\mathbf{1}^\top \widehat{\boldsymbol{\Delta}}_0^{-\frac{1}{2}} \boldsymbol{s} = p^{-1}\sum_{k=1}^{p} \widehat{\delta}_{0k}^{-\frac{1}{2}} s_k$, which is equivalent to the sum of variable-specific standardised mean score differences (up to the factor p^{-1}). Hence, in order to test for MO where stochastic order but not its direction is relevant, we

propose to use the direction-independent test statistic

$$\widetilde{S}' = \sum_{k=1}^{p} \widehat{\delta}_{0k}^{-1} s_k^2, \tag{8}$$

which is the sum of squared variable-specific standardised mean score differences. As with the chi-squared sum statistic S from Eq. 7, the p summands that form S' and $\widetilde{S}'$, respectively, turn out to be well-known in the literature: a closer look at Klingenberg's S' reveals that, up to the factor p^{-1}, it is equivalent to the sum of marginal Cochran-Armitage (CA) trend test statistics (Cochran, 1954; Armitage, 1955), for any choice of scores. The proof is given in Appendix A. Thus, our test statistic $\widetilde{S}'$ is equivalent to the sum of squared marginal CA test statistics, and we shall therefore refer to $\widetilde{S}'$ as CA sum statistic. As the marginal CA test statistic has an asymptotic standard normal null distribution, under independence, the CA sum statistic has a chi-squared null distribution with df = p.

The link between S' and $\widetilde{S}'$, respectively, and the CA test statistic deserves special attention because it provides important information on which inferences may or may not be drawn from a test result. The crux is that the CA test statistic is intended to test for some suspected trend in the binomial proportions across the c ordered levels. Which particular trend the test statistic will be sensitive towards is determined by scores which are in one-to-one correspondence with the scores $u_k(v)$ from above. For the CA sum statistic $\widetilde{S}'$, we suppose that the scores are uniform over all k (i.e. $u_k(v) = u(v)$) and that they increase or decrease monotonically. Note that uniform scores are not compulsory, but they are a convenient choice in most applications. It is now easily verfied that MO is fulfilled if there is some monotonic trend, that is, if either $\frac{n_{2k}(1)}{n_{\cdot k}(1)} \leq \frac{n_{2k}(2)}{n_{\cdot k}(2)} \leq \cdots \leq \frac{n_{2k}(c)}{n_{\cdot k}(c)}$ with $\frac{n_{2k}(1)}{n_{\cdot k}(1)} < \frac{n_{2k}(c)}{n_{\cdot k}(c)}$ or $\frac{n_{2k}(1)}{n_{\cdot k}(1)} \geq \frac{n_{2k}(2)}{n_{\cdot k}(2)} \geq \cdots \geq \frac{n_{2k}(c)}{n_{\cdot k}(c)}$ with $\frac{n_{2k}(1)}{n_{\cdot k}(1)} > \frac{n_{2k}(c)}{n_{\cdot k}(c)}$ for at least one k, where $n_{\cdot k}(v) = n_{1k}(v) + n_{2k}(v)$. The reverse, however, is not true. A monotonic trend in the binomial proportions thus implies an alternative that is narrower than MO. Consequently, because tests that rest upon the CA sum statistic $\widetilde{S}'$ are essentially designed to detect such monotonic trends, they may have low power to detect MO if there is no such trend. This should be kept in mind whenever MO, perhaps unexpectedly, could not be confirmed.

Compared to the chi-squared sum statistic, the CA sum statistic will result in more power when the suspected trend or its inverse is correct for all k for which H_{1k} in Eq. 2 is fulfilled, but it is likely to result in considerably less power otherwise. In the case $c = 2$, the two sum statistics S and $\widetilde{S}'$ are equivalent for any choice of scores and will therefore result in equally powerful tests. In the case $c > 2$, S and $\widetilde{S}'$ are equivalent only if we use the data-driven scores of Zheng et al. (2009) which, however, do not necessarily increase or decrease monotonically.

3.3. Adopting the marginal perspective: sum and max-T statistics

We have presented the test statistics S and $\widetilde{S}'$ as special cases of multivariate quadratic forms, under the assumption that the variables be independent. This multivariate perspective is beneficial, particularly because it immediately clarifies why the independence assumption will be difficult to circumvent in real-life applications where n is typically small to moderate, p is moderate to large, and $c > 2$. Nevertheless, the fact that both S and $\widetilde{S}'$ have turned out to be composed of well-known traditional univariate test statistics provokes to directly look at them from the less sophisticated yet popular marginal perspective. This is in the spirit of Pesarin's permutation-based Non-Parametric Combination (NPC) method (Pesarin, 2001) which combines marginal p-values through some

well-chosen combination function into one test statistic for the entire set. Direct sum statistics of the form $\sum_{k=1}^{p} T_k$ (e.g. S) and $\sum_{k=1}^{p} T_k^2$ (e.g. $\widetilde{S}'$), where T_k is the kth marginal test statistic, likewise fall into the NPC framework. We treat all T_ks on the same footing, but different weights can in principle be incorporated when the variables are of different importance. (The ICF category 'heart functions' (b410) might be considered more important than the ICF category 'voice functions' (b310), for example.) A prominent counterconcept to sum statistics are max-T statistics where the maximum over all (possibly transformed) univariate test statistics in a set is assumed to adequately reflect the whole set's effect. The max-T enables a shortcut of the FWER-controlling closure test principle of Marcus et al. (1976), rendering it useful when multiple tests are to be conducted at the individual level (Westfall and Young, 1993). For the assessment of *set effects*, however, we consider sum statistics more suitable, for two reasons. Firstly, they can be interpreted conveniently as the accumulated effect of variables over a whole set. Secondly, they generally lead to more powerful tests in the presence of many weak or moderate individual effects. To be able to perform such tests based on the proposed sum statistics, we still need their distributions under the null hypothesis. For inferences to be valid, the latter should take the multivariate dependence structure in the data into account, even if the sum statistics do not so. Permutation-based null distributions can accomplish this, but only at the price of an assumption that is rarely justified in practice. We address this issue in Section 4.

4. Permutation-based global inference about marginal distributions: the null dilemma

In high-dimensional multivariate scenarios, permutation null distributions of test statistics have become popular since, apart from being easy to calculate, they automatically preserve the dependence structure in the data and yield exact level-α tests. The price to pay in order for these appealing properties to be provided is that the multivariate observations must be exchangeable within and between groups under the null hypothesis (i.e. the observations' joint distribution must be invariant to group label permutation). In our context, this condition is fulfilled under IJD, but not under SMH. Permutation tests for MI or MO will thus not be valid unless the null hypothesis is IJD, where validity refers to whether the type I error rate tends to the prespecified level α. In practice, however, IJD is unrealistic or at least questionable. Perhaps the only scenario where it appears realistic is that encountered in randomised studies, but most ICF studies are non-randomised. In our ICF stroke study, for example, the dependence structure between the ICF categories in the ICF chapter 'attitudes' (e4) is expected to be different for Asian and European patients, rendering IJD untenable. Whether we test SMH against MI or against MO, this inevitably leads to what we call here the 'null dilemma': we can either use the permutation null distribution despite its deficiency under SMH, but then the test result must be interpreted carefully because it may be conservative or anticonservative, or we can attempt to derive an alternative bootstrap null distribution, but bootstrap tests are only asymptotic level-α tests (Efron and Tibshirani, 1993) and usually come with their own problems, especially when $n < p$ (Troendle et al., 2004). Note that further options may exist in specific situations, yet the two mentioned are most common in statistical practice. Because the permutation procedure is preferred whenever it appears applicable, it is desirable to understand its robustness properties under SMH. Several authors have established conditions under which permutation tests remain valid even under non-exchangeability, at least in an asymptotic sense (Romano, 1990; Pollard and van der Laan, 2004; Huang et al., 2006; Kaizar et al., 2011). For test statistics that rely on differences in sample mean vec-

tors, Huang et al. (2006) compared the permutation distribution and true distribution in terms of cumulants. Unless the cumulants are equal in the two multivariate distributions to be compared, it turned out that the odd-order cumulants of the test statistic's permutation and true distribution will be different in all possible situations, while the even-order cumulants will be asymptotically equal if $n_1 = n_2$. In the multivariate normal case where merely the first two cumulants (i.e. mean vector and covariance matrix) are non-zero, the permutation and true distribution thus coincide asymptotically if $n_1 = n_2$, rendering the permutation procedure asymptotically valid. In the multivariate ordinal case, however, there may be infinitely many non-zero cumulants. Hence, even if $n_1 = n_2$, here the permutation procedure is invalid.

While the validity constraints of permutation tests have been well studied on the theoretical side, it is unclear yet which impact they have on the practical side. In the simulation experiments of Klingenberg et al. (2009), the permutation procedure remained applicable under SMH, even for $n_1 \neq n_2$. Kaizar et al. (2011), on the other hand, found scenarios under SMH in which the max-T permutation test based on Fisher test statistics fails. More systematic simulation experiments on this issue are presented in Section 5.

5. Robustness properties of the permutation procedure: a simulation study

5.1. Simulation setup

We conducted an extensive simulation study with the aim to better understand, for small to moderate sample sizes, the behaviour of permutation tests under SMH, that is, in case of violations of exchangeability. In particular, we considered tests based on the sum statistics S and $\widetilde{S}'$ (with equally spaced scores $u(v) = v$) as well as their max-T counterparts (i.e. the maximum univariate chi-squared and squared CA test statistic). Systematic power comparisons under MI without MO and/or MO were outside the scope of this study. Multivariate ordinal data were generated using the 'mean mapping method' from the `R` package `orddata` (Kaiser and Leisch, 2010), which is based on cutting multivariate normal distributions at quantiles defined by the ordinal variables' marginal distributions. (One needs to specify p vectors of c marginal probabilities adding up to 1 and a positive semi-definite $(p \times p)$ correlation matrix.) As a result of this technique, it was not possible to examine the effect of non-exchangeability in cumulants of order higher than two.

We considered the set sizes $p = \{20, 100\}$ with $c = 4$ and the overall sample sizes $n = \{20, 40, 60, 80\}$ which were split into $(n_1, n_2) = \{(10, 10), (20, 20), (30, 30), (40, 40)\}$ (balanced groups), $(n_1, n_2) = \{(8, 12), (16, 24), (26, 34), (32, 48)\}$ (unbalanced groups), and $(n_1, n_2) = \{(5, 15), (12, 28), (18, 42), (24, 56)\}$ (very unbalanced groups). In order to reflect SMH, we set the marginal probabilities to $(0.25, 0.25, 0.25, 0.25)$ for all variables in both groups. We generated (non-)exchangeability (in the second cumulant) by means of 14 pairs of uniform correlation matrices: $(\rho_1, \rho_2) = \{(0, 0), (0.25, 0.25), (0.5, 0.5), (0.75, 0.75), (0, 0.25), (0.25, 0.5), (0, 0.5), (0.25, 0.75), (0, 0.75), (0.25, 0), (0.5, 0.25), (0.5, 0), (0.75, 0.25), (0.75, 0)\}$, with ρ_i denoting the correlation parameter in group i. Thus, the number of different combinations of set sizes, group sizes, and correlation parameters was $2 \times 3 \times 4 \times 14 = 336$. (Note that for equal group sizes there is no difference between, for example, $(\rho_1, \rho_2) = (0, 0.25)$ and $(\rho_1, \rho_2) = (0.25, 0)$. Such scenarios were not generated individually.) For each such parameter constellation, the type I error rate was estimated from 1000 datasets as the average rejection rate of true null hypotheses. The test statistics' permutation null distributions were approximated based on 5000 permutation resamples, and the desired significance level was $\alpha = 0.05$. It is important to note that because the margins of the p one-way tables are invariant to group label permutation, the respective type I error rates are to be

interpreted *conditional* upon the observed table margins. Furthermore, note that we used mid-p-values (Lancester, 1961) to adjust for discreteness. Mid-p-values are calculated as the proportion of resampled test statistics more extreme than the observed one plus half (instead of all) of the proportion of resampled test statistics equal to the observed one. While this approach does not guarantee type I error rate control, various numerical evaluations have shown that null mid-p-values tend to be more uniformly distributed than ordinary null p-values (Hirji, 1991; Agresti, 2001; Klingenberg et al., 2009).

5.2. Simulation results

All simulation results are reported in detail in online supplement B. For the 336 parameter constellations, the heat maps in Fig. 2 illustrate the deviations of the actual type I error rate from the nominal type I error rate ($\alpha = 0.05$) with the permutation null distribution of the sum statistic S. Values < 0 indicate conservative behaviour (shown in violet) and values > 0 anticonservative behaviour (shown in red). To spot possible biases (i.e. systematic fluctuations around the ideal value 0 (shown in white)) more easily, values outside the simulation margin of error of approximately $\pm 1.38\%$ are additionally highlighted. For $p = 20$ (Fig. 2A), the actual type I error rate is close to the nominal one in the scenarios with balanced group sizes, regardless of whether under exchangeability (i.e. when $\rho_1 = \rho_2$) or non-exchangeability (i.e. when $\rho_1 \neq \rho_2$). For unbalanced and very unbalanced group sizes, this applies only under exchangeability. Under non-exchangeability, it seems crucial to distinguish which group the higher correlation is combined with: higher correlation in the larger group (i.e. $\rho_1 < \rho_2$) entails conservative behaviour (the actual type I error rate ranges from 0.025 to 0.054 for unbalanced and from 0.011 to 0.040 for very unbalanced group sizes), whereas higher correlation in the smaller group (i.e. $\rho_1 > \rho_2$) entails overly anticonservative behaviour (the actual type I error rate ranges from 0.051 to 0.081 for unbalanced and from 0.048 to 0.122 for very unbalanced group sizes). Perhaps unexpectedly, the permutation procedure's robustness properties seem not to vary systematically with the overall sample size, as has already been observed by Kaizar et al. (2011). For $p = 100$ (Fig. 2B), we come to basically the same conclusions, but the deviations from the nominal type I error rate are partly considerably more pronounced than for $p = 20$, which is readily visible from Fig. 2B. For very unbalanced group sizes, for example, the actual type I error rate ranges from 0.005 to 0.040 when $\rho_1 < \rho_2$ and from 0.066 to 0.200 when $\rho_1 > \rho_2$. With the permutation null distribution of the sum statistic $\widetilde{S}'$, we arrive at very similar results throughout, which becomes evident when we compare the heat maps in Fig. 3 with those in Fig. 2. When our sum statistics are employed, it thus seems that the permutation procedure cannot be recommended under SMH unless it holds $n_1 = n_2$. However, one should note that many scenarios in which the permutation procedure seriously fails are unlikely to be encountered in practice (e.g. those with $\rho_1 = 0$ and $\rho_2 = 0.75$ or vice versa), while its failure in more realistic scenarios (e.g. those with $\rho_1 = 0.25$ and $\rho_2 = 0.5$ or vice versa) seems to be less dramatic, in particular for moderately unbalanced group sizes. Therefore, if potentially some more type I errors than desired do not pose enormous problems in the application at hand and the group sizes are not exceedingly unbalanced, we believe that the permutation procedure may still be used.

Similarly, the heat maps in Fig. 4 now illustrate the results obtained with the permutation null distribution of the max-T based on chi-squared test statistics. Remarkably, here the permutation null distribution seems to remain 'practically valid' even under non-exchangeability and unbalancedness, with nearly all deviations from the nominal type I error rate lying within the simulation margin of error. In contrast to that, Fig. 5 sug-

gests that the permutation null distribution of the max-T based on CA test statistics is less robust. For very unbalanced group sizes, it is particularly prone to anticonservative behaviour when the higher correlation is combined with the larger group: for $p = 20$ (Fig. 5A), the actual type I error rate ranges from 0.049 to 0.075 across the respective scenarios, while its range for $p = 100$ (Fig. 5B) is from 0.046 to 0.085. Compared to the extent to which permutation tests based on our sum statistics may fail, this appears almost negligible. Nevertheless, max-T tests are not per se the better choice, especially when many weak rather than few strong individual effects are expected in the set of interest.

In follow-up simulations, we repeated the complete study with $c = 2$ in order to see whether the robustness properties identified above depend on the number of levels per variable (see online supplement B). In the context of this paper, the case $c = 2$ is of relatively little interest, but it is computationally convenient because here our sum statistics and their max-T counterparts, respectively, are equivalent. Hence, it is sufficient to examine one sum and one max-T statistic. For the sum statistic, we find that the results are similar to those in the case $c = 4$ (see Fig. 2 and 3). For the max-T statistic, the results are similar to those for the CA-based max-T statistic in the case $c = 4$ (see Fig. 5). The max-T permutation test based on chi-squared test statistics thus has different robustness properties for $c = 2$ than for $c = 4$. We expect permutation tests that rest upon squared CA test statistics to have similar robustness properties for any choice of c because, unlike the chi-squared test statistic which has df $= c - 1$, the squared CA test statistic has df = 1 independent of c.

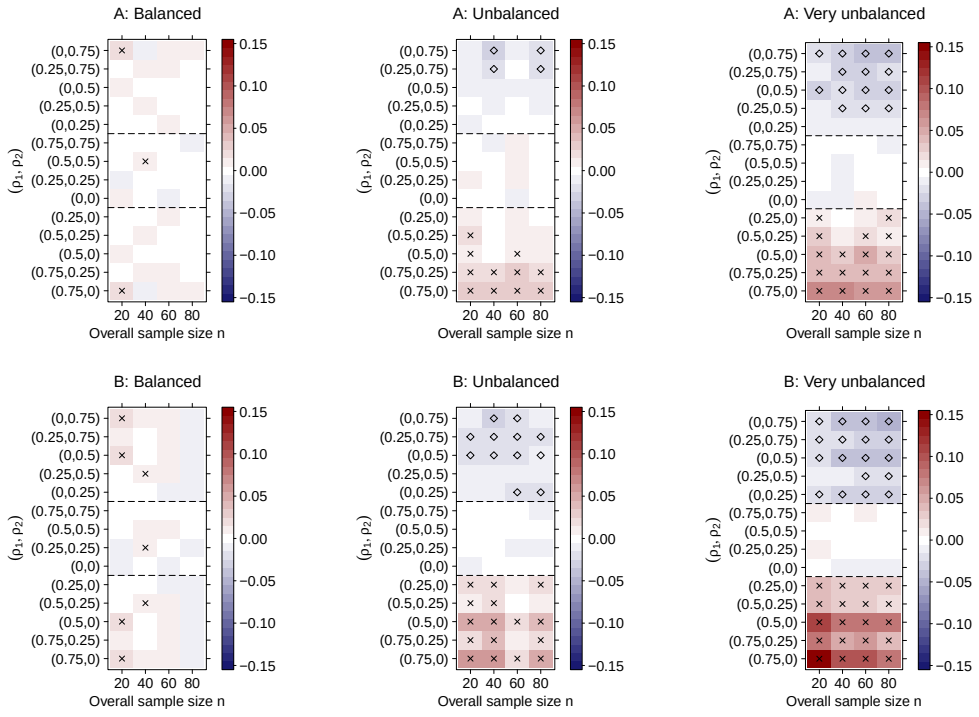

Figure 2. Actual minus nominal type I error rate with the permutation null distribution of S for **A:** $p = 20$ and **B:** $p = 100$. Each heat map cell corresponds to one of the 336 simulation scenarios. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. Values outside the margin of error are marked: diamonds indicate systematic conservativeness and crosses systematic anticonservativeness. The colour scale is chosen such that a direct visual comparison of Fig. 2–5 is enabled.

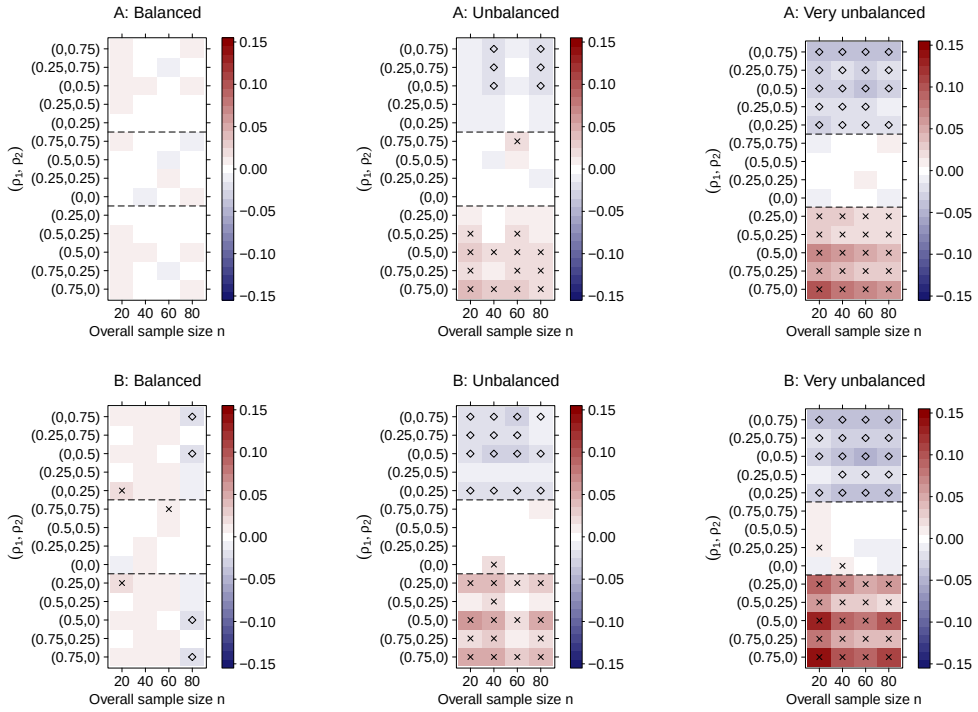

Figure 3. Actual minus nominal type I error rate with the permutation null distribution of $\widetilde{S}'$ for **A:** $p = 20$ and **B:** $p = 100$. Further annotations as explained in Fig. 2.

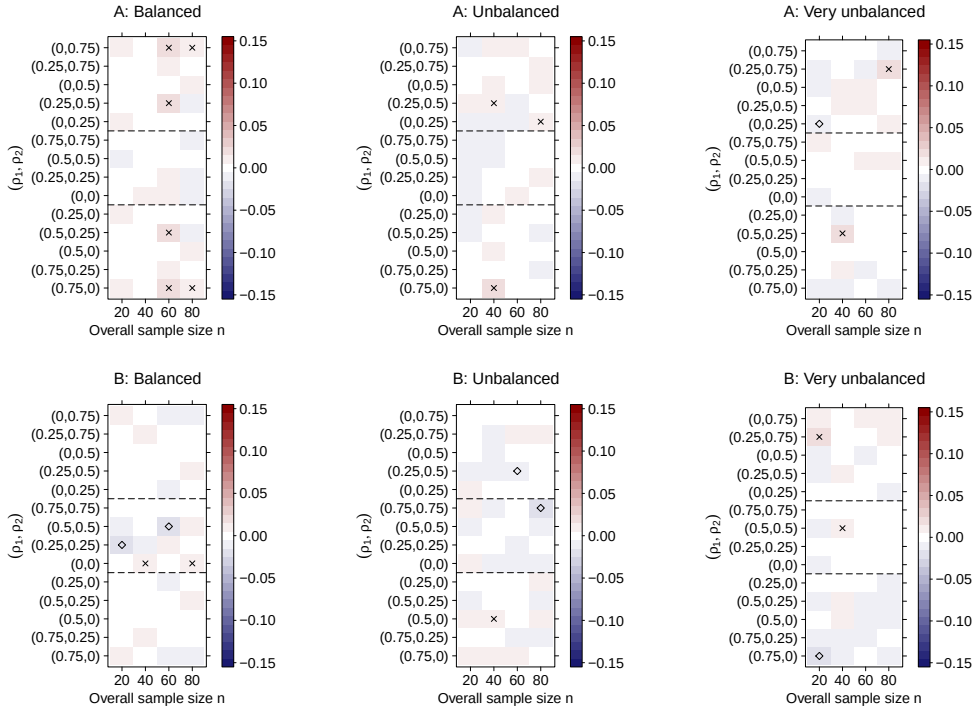

Figure 4. Actual minus nominal type I error rate with the permutation null distribution of max-T (chi-squared) for **A:** $p = 20$ and **B:** $p = 100$. Further annotations as explained in Fig. 2.

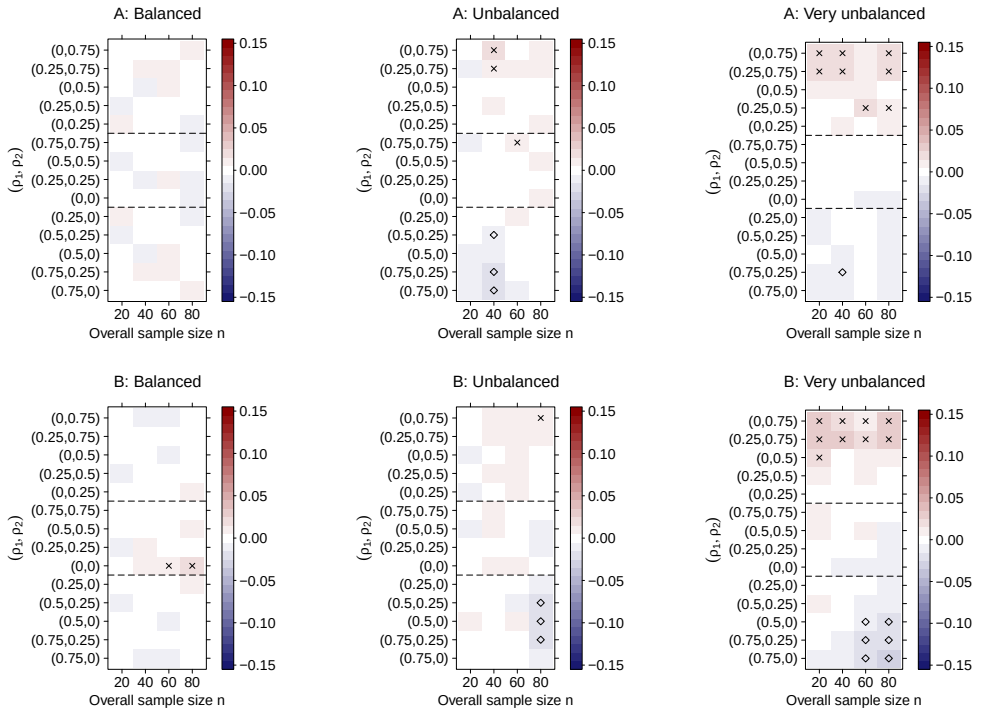

Figure 5. Actual minus nominal type I error rate with the permutation null distribution of max-T (CA) for **A:** $p = 20$ and **B:** $p = 100$. Further annotations as explained in Fig. 2.

6. Analysis of ICF core set data for stroke patients

6.1. Multiplicity adjustment for tree-structured hypotheses

In Sections 2–5 we have considered scenarios where one set is tested. As soon as multiple sets are to be tested simultaneously, multiplicity adjustment procedures become relevant. For the special case of tree-structured hypotheses such as depicted in Fig. 1, Meinshausen (2008) introduced a simple 'top-down' multiple testing procedure which offers FWER control simultaneously over all tree levels. The procedure starts with testing the root set (i.e. complete set) at level α. If the null hypothesis is rejected, it continues by testing the child sets at the subsequent tree level and descends only into child sets of rejected null hypotheses. This means that child sets of sets whose null hypotheses could not be rejected are *not* tested. Provided that an effect has been ascertained in the root set, Meinshausen's procedure thus tries to attribute this effect to more specific sets or even individual variables. Essentially, it thereby opens the door to a compromise between global and classical multiple testing. For any set M that is tested in the top-down approach, the adjusted p-value is the raw p-value multiplied by $p/|M|$, where $|M|$ denotes the cardinality of M and p is the cardinality of the root set (i.e. the p-value of the root set is unadjusted, whereas univariate p-values receive the Bonferroni adjustment). Each tree level can thus be tested at level α, even though the FWER is controlled simultaneously over *all* tree levels at level α. Recently, Goeman and Solari (2010) and Goeman and Finos (2012) developed more elaborate procedures for tree structures which are uniformly more powerful than that of Meinshausen. For the sake of brevity and simplicity, their procedures are not considered in the present paper. Computationally, when our proposed tests are used to test for set effects, Meinshausen's procedure seems to involve as many permutation rounds as there are sets in the tree. However, p-values for an entire tree structure can be computed efficiently based on one permutation round for the root set (i.e. from the resultant $(p \times R)$

matrix that contains the p marginal test statistics for the R permutation resamples). This is beneficial, in particular when extensive tree structures are studied.

As stated above, in Meinshausen's procedure the multiplicity penalty for any tested set M is $p/|M|$. Sets that comprise many variables will thus be easier to reject than sets that comprise few variables. In some applications, such an implicit priorisation of large sets may be inconvenient. In ICF-based applications, however, this is even desirable because it reflects the expert opinion based on which ICF core sets are composed. In the ICF core set for stroke, the ICF components 'body structures' (s) and 'environmental factors' (e) receive the multiplicity penalties 130/5 and 130/33, which is plausible because social and attitudinal aspects are considered more relevant for stroke patients than anatomical aspects (Geyh et al., 2004). (Otherwise, more than just five ICF categories describing anatomical aspects would have been included by the health experts in the core set.) This is different for patients suffering from ankylosing spondylitis, for example. In the respective ICF core set, the ICF components s and e therefore receive the multiplicity penalties 80/19 and 80/14 (Boonen et al., 2010).

6.2. Data description and methods

Our ICF-based dataset includes patients after first stroke from a multicentre study conducted in post-acute rehabilitation facilities from 2004 to 2007. To recap: the ICF core set for stroke based on which the data were collected comprises $p = 130$ ICF categories (listed in online supplement A). Of the $n = 104$ patients $n_1 = 46$ came from high-income Asian countries (China, Malaysia, South Korea, Thailand) and $n_2 = 58$ from European countries (Austria, Germany, Italy, Norway, Sweden, Switzerland). At the time of data collection, all patients were ≥ 50 years old and their Body Mass Index (BMI) was ≤ 30. The two groups did not differ substantially in the distribution of sex, age, and BMI, rendering adjustment for these typical confounders unnecessary. As has been recommended by Bostan et al. (2012) for the five-level ordinal scale used in the ICF components b, s, and d, we coarsened both the five-level and the nine-level scale used in the ICF component e to $c = 3$ levels: the scheme 0 1 2 3 4 was coarsened to 0 1 1 2 2, whereas the scheme -4 -3 -2 -1 0 1 2 3 4 was coarsened to 2 2 2 2 1 0 0 0 0. This potentially reduces the number of variables for which one or more levels could not be observed in both groups, which is an appreciable side effect because such variables lead to degenerate $\widehat{\mathbf{\Sigma}}_{0k}$s. With the three-level scheme this occurs only for the ICF category 'blood pressure functions' (b420) where the third level (severe to complete impairment) has never been observed. We set the associated univariate test statistic to zero. Less conservative strategies to handle the ICF category b420 are to exclude it from the analysis or to treat it as binary; both led to the same conclusions as our strategy. For the comparison of Asian and European stroke patients with respect to their ICF pattern, we contrasted five approaches (A1–A5) with each other:

(A1) We combined Meinshausen's top-down procedure with our permutation test based on the chi-squared sum statistic S. We approximated the permutation null distribution of S based on 10000 resamples. A complete enumeration of all 104!/ (46!58!) $\approx 7.96 \times 10^{29}$ possible permutation resamples was too computationally intensive.

(A2) See A1, however with the max-T permutation test based on chi-squared test statistics to test any set considered in Meinshausen's procedure.

(A3) We carried out the traditional univariate chi-squared test for each ICF category and subsequently applied the Bonferroni-Holm procedure (Holm, 1979) to adjust for multiplicity. This approach is rather simplistic, yet it is widely used.

(A4) See A3, however with the permutation null distribution (approximated based on

10000 resamples) which, unlike the asymptotic chi-squared null distribution used in A3, respects the dependence between the individual test statistics.

(A5) We used the permutation-based 'discrete Bonferroni method' of Westfall and Troendle (2008), again with 10000 resamples. This approach is similar to the popular max-T-based stepdown approach of Westfall and Young (1993), with the crucial difference being that it provides FWER control under SMH, at the price of potentially less power. Unlike A1–A4, this approach dispenses with mid-p-values.

Because, in this particular ICF example, it was of interest to detect MI rather than MO, analogous approaches based on the CA sum statistic $\widetilde{S}'$ or its max-T counterpart were not taken into consideration. The results obtained with A1–A5 are summarised in Table 1 and discussed in Section 6.3. It should be emphasised that, unlike A1 and A2, A3–A5 do not exploit the datas' tree structure inferentially. However, we can exploit it ex post for interpretation by treating the smallest adjusted p-value in a set as set-specific test.

6.3. Results

Fig. 6 shows, for the complete ICF core set, the permutation null distributions of the test statistics S and max-T (chi-squared), together with their asymptotic null distributions under the assumption of independence between the 130 ICF categories. Strictly speaking, the permutation and asymptotic null distributions are not fully comparable because the former are conditional on the observed table margins for each ICF category, whereas the latter are unconditional distributions. Under independence, however, the conditional and unconditional null distributions asymptotically behave similarly (or even the same under certain conditions (Romano, 1990)). For this reason, because $n = 104$ is sufficiently large, the comparison between the two null distributions in Fig. 6 provides reliable information on how valid or invalid results based on the asymptotic null distributions would be. As becomes evident from the figure, the asymptotic null distributions are inappropriate in the present application. Regardless of which test statistic is chosen, we find that the ICF core set is significant (i.e. MI is confirmed between the overall ICF pattern of stroke patients from Asian and European countries). Table 1 now tells us which sets (i.e. ICF components, chapters, and categories) this significant difference can be attributed to. For clarity, it contains only the ICF components, chapters, and categories that have been identified as significant by at least one of the five approaches A1–A5 described in Section 6.2. We find that the results are fairly consistent across all approaches apart from A3. Mostly owing to ignored dependencies between the ICF categories and thus between the associated test statistics, A3 yields the most conservative conclusion with four significant ICF categories: 'structure of upper extremity' (s730), 'acquiring, keeping and terminating a job' (d845), 'products and technology for personal indoor and outdoor mobility and transportation' (e120), and 'architecture and construction services, systems and policies' (e515). The latter are also found to be significant by A1, A2, A4, and A5, together with the ICF categories 'structure of lower extremity' (s750), 'housing services, systems and policies' (e525), and 'associations and organisational services, systems and policies' (e555). For the ICF category 'doing housework' (d640), MI is revealed merely by A2, A4, and A5. As displayed in Table 1, A1 which is based on the sum statistic S does not reject SMH for the ICF chapter 'domestic life' (d6); Meinshausen's procedure hence does not descend further into individual ICF categories one of which is d640. This potential type II error may be explained by the fact that A1 has different power properties than A2–A5. Conversely, it is solely A1 which detects MI for the ICF chapters 'neuromusculoskeletal and movement-related functions' (b7) and 'support and relationships' (e3), whereas none of

the nine and seven ICF categories contained is found to be marginally significant by neither of the approaches. Apparently, the ICF categories in b7 and e3 only *jointly* provide evidence against SMH. This result indicates that A1 outperforms the other approaches in the presence of many weak individual effects, as has been expected. Note that while it is unlikely that A1, A2, and A4 are theoretically valid in the present application, we assume that they are practically valid, for two reasons. Firstly, the group sizes are relatively weakly unbalanced. Secondly, at the level of individual ICF categories, A1, A2, and A4 (which do not guarantee FWER control under SMH) lead to nearly the same conclusions as A5 (which guarantees FWER control under SMH).

The fact that many of the differences we have found are in the environment suggests that the kind of support people receive after stroke differs between Asian and European countries, which in turn reflects that the two country groups differ in their health and social policies. This does not come as a surprise and supports the validity of our results. The latter may now serve the WHO or other international organisations to uncover those inequalities in health service provision that directly affect stroke patients. Information of this kind may help policy makers to eliminate or reduce such inequalities and ultimately to improve the quality of post-stroke rehabilitation services. The difference found in support and relationships is particularly noteworthy. Astin et al. (2008) reported that cardiac patients are more frequently cared at home by their family in Asian than in European countries where residential care is much more common. Both results put together form a good basis for more detailed studies on the role of family and non-family relationships in post-stroke rehabilitation. The differences we have found in body functions and structures require additional explanation. The question is whether they are due to different evaluation approaches more than really due to differently affected body functions and structures. Further studies are needed to answer this question.

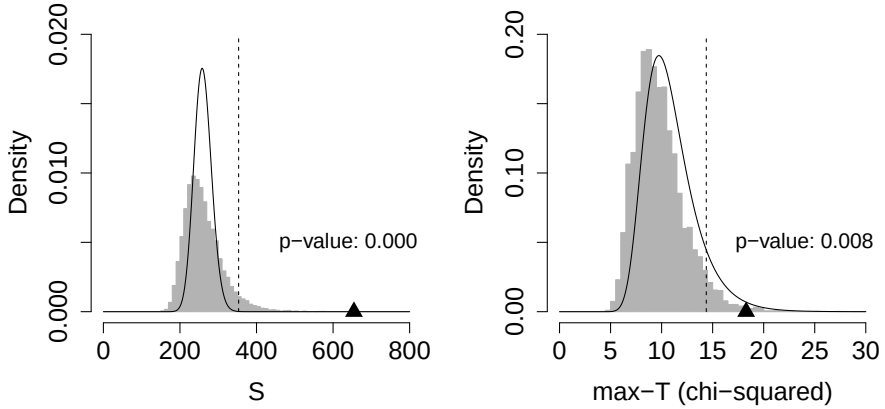

Figure 6. Grey areas show the permutation null distributions (approximated based on 10000 resamples) and superimposed black curves the asymptotic null distributions (assuming independence) of the chi-squared sum statistic S and the max-T based on chi-squared test statistics for the complete ICF core set. Dashed lines indicate critical values (0.95-quantiles) of the permutation distributions. Filled triangles indicate observed values of S and max-T (chi-squared).

Table 1. Multiplicity adjusted p-values for the ICF components, chapters, and categories that have been identified as significant by at least one of the approaches A1–A5 (see Section 6.2 for detailed explanations), with $\alpha = 0.05$. Adjusted p-values > 0.05 are indicated by '–'.

	A1	A2	A3	A4	A5
Body functions (b)	0.017	–	–	–	–
Neuromusculoskeletal and movement-related functions (b7)	0.004	–	–	–	–
Body structures (s)	0.000	0.005	0.014	0.013	0.008
Structures related to movement (s7)	0.000	0.009	0.014	0.013	0.008
Structure of shoulder region (s720)	0.032	0.032	–	0.031	–
Structure of upper extremity (s730)	0.013	0.013	0.014	0.013	0.008
Structure of lower extremity (s750)	0.046	0.046	–	0.043	0.038
Activities and participation (d)	0.028	0.013	0.020	0.000	0.011
Domestic life (d6)	–	0.035	–	0.000	0.029
Doing housework (d640)	–	0.000	–	0.000	0.029
Major life areas (d8)	0.003	0.010	0.020	0.025	0.011
Acquiring, keeping and terminating a job (d845)	0.026	0.026	0.020	0.025	0.011
Environmental factors (e)	0.000	0.011	0.022	0.013	0.011
Products and technology (e1)	0.002	0.015	0.022	0.013	0.011
Products and technology for personal use in daily living (e115)	–	–	–	–	0.028
Products and technology for personal indoor and outdoor mobility and transportation (e120)	0.013	0.013	0.022	0.013	0.011
Support and relationships (e3)	0.017	–	–	–	–
Services, systems and policies (e5)	0.004	0.012	0.034	0.019	0.016
Architecture and construction services, systems and policies (e515)	0.039	0.039	0.034	0.037	0.016
Housing services, systems and policies (e525)	0.020	0.020	–	0.019	0.030
Associations and organisational services, systems and policies (e555)	0.026	0.026	–	0.025	0.027

7. Discussion and conclusion

Motivated by the need for statistical tools to analyse data collected by means of the ICF, we have discussed two-sample permutation tests for sets of possibly high-dimensional multivariate ordinal data. Specifically, we have addressed the closely related problems 'SMH against MI' (Eq. 1 and 2) and 'SMH against MO' (Eq. 1 and 3). While ICF-based data have been our main motivation, the proposed tests can likewise be used to analyse items in psychodiagnostic tests (e.g. structured into sets by the subdimension they describe), side or adverse effects in drug safety or toxicity studies (e.g. structured into sets by means of the Medical Dictionary for Regulatory Activities (MedDRA)), or single-nucleotide polymorphisms (SNPs) in genome-wide studies (e.g. structured into sets by genes).

To capture MI and MO, we have proposed sum statistics (Eq. 7 and 8), derived as multivariate test statistics under the assumption that the variables be independent. Under this assumption, we have found that the test statistic of Klingenberg et al. (2009), which our test statistic for MO is based on, is equivalent to the sum of univariate CA trend test statistics. Given that the independence assumption is inevitable in most practical situations, this equivalence argues for broader exploration of tests based on simple sum statistics constructed from other traditional univariate test statistics for ordinal data. Compared to tests based on max-T statistics, such tests usually have more power against alternatives with many weak individual effects, which is an important class of alternatives in ICF-

based applications and beyond. This is well known and has been reinforced in our power studies (not shown). Regarding the tests proposed in this work, there is an additional intuitive explanation why they are expected to be powerful against this class of alternatives: both the Pearson chi-squared and CA test statistic, from which our sum statistics are constructed, are score test statistics, and score test statistics operate under the null hypothesis, resulting in good power against alternatives 'close' to the null hypothesis.

By means of simulations, we have explored the behaviour of our permutation tests and their max-T counterparts under SMH, that is, in null scenarios where the multivariate observations may be non-exchangeable across groups (Fig. 2–5). The motivation behind has been that despite the theoretically well-founded criticism towards permutation-based inference in such data scenarios, researchers commonly face the problem that no superior (e.g. bootstrap-based) inference methods exist. Of the bootstrap procedures we have considered (not shown), none has proved to be a promising alternative to the permutation procedure. We have called this common situation 'null dilemma'. As expected, our simulations have suggested that how deficient the permutation procedure can become depends on the difference in the group-specific covariance matrices, the proportion between the group sizes, and the number of variables in the set. It has come as an initially unexpected observation, however, that the choice of the test statistic and the number of levels per variable seem to play a crucial role as well. For instance, max-T permutation tests have shown remarkable robustness properties under SMH, especially when the max-T based on chi-squared test statistics is used and the number of levels is not too small (Fig. 4). Subject to our simulations, it can thus be concluded that theoretical invalidity does not necessarily imply practical invalidity. It is unrealistic to expect simple and generally valid guidelines, but we believe that systematic studies such as ours will help to establish some useful practical recommendations regarding the use of permutation tests under SMH.

As the ICF is more and more used worldwide to collect data on functional limitations and disabilities, the need for statistical tools tailored to such data will continue to rise. Recently, this has been realised by other authors as well (Kalisch et al., 2010; Gertheiss et al., 2011). The tests presented in this paper enable researchers to analyse ICF-based data at different levels of detail (e.g. ICF components or ICF chapters). They are useful by themselves and, in addition, can be fruitfully combined with available multiple testing procedures for tree-structured hypotheses (Meinshausen, 2008; Goeman and Solari, 2010; Goeman and Finos, 2012). However, the proposed methodology has its limitations and can be improved in several directions. Firstly, it may be extended to scenarios in which not all variables are measured on the same ordinal scale. This seems unproblematic if the marginal test statistics maintain the same number of degrees of freedom, such as is the case for the CA test statistic. When the chi-squared test statistic is used, some standardisation will be needed. Secondly, it is desirable to extend it to scenarios in which two groups shall be compared after adjustment for covariates. In non-randomised ICF studies, for instance, the two groups to be compared often differ substantially with respect to age and BMI, the major confounders in studies on human functioning and disability. To avoid false positive results due to such confounders, it is of utmost importance to be able to adjust for the them in the analysis. A simple way to achieve this is to apply the proposed unadjusted tests in covariate-defined strata and subsequently correct for multiplicity over the strata. However, such an approach usually becomes infeasible when there are several potential confounders to adjust for, since the typical sample sizes are too small to construct multivariate strata. Alternatively, the comprehensive theory on generalised linear models may be exploited to form relevant sum statistics, yet their permutation null distribution will require more assumptions than in the unadjusted case in order to be valid.

Acknowledgements

MJ is grateful to the German National Academic Foundation for a PhD scholarship by which this work was supported. We thank Jelle Goeman, Bernhard Klingenberg, and Reinhard Meister for helpful discussions.

A. Equivalence of Klingenberg's test statistic and the sum of CA test statistics

Under the assumption of independence between variables, the test statistic of Klingenberg et al. (2009) is the sum statistic $S' = \sum_{k=1}^{p} \widehat{\delta}_{0k}^{-\frac{1}{2}} s_k$ (divided by p). It is therefore sufficient to show the equivalence in the univariate case. We refer to the notation introduced in Section 3. For the kth component we obtain

$$\begin{aligned}
\widehat{\delta}_{0k}^{-\frac{1}{2}} s_k &= \left(\boldsymbol{u}_k^\top \widehat{\mathrm{Cov}}(\boldsymbol{d}'_k)\boldsymbol{u}_k\right)^{-\frac{1}{2}} \boldsymbol{u}_k^\top \boldsymbol{d}'_k \\
&= \left\{ \frac{n}{n_1 n_2} \left[\boldsymbol{u}_k^\top \left(\mathrm{diag}(\widehat{\boldsymbol{\pi}}'_{0k}) - \widehat{\boldsymbol{\pi}}'_{0k}\widehat{\boldsymbol{\pi}}'^{\top}_{0k}\right)\boldsymbol{u}_k\right] \right\}^{-\frac{1}{2}} \boldsymbol{u}_k^\top \boldsymbol{d}'_k \\
&= \left\{ \frac{n}{n_1 n_2} \left[\frac{1}{n}\sum_{v=1}^{c} u_k(v)^2 n_{\cdot k}(v) - \left(\frac{1}{n}\sum_{v=1}^{c} u_k(v) n_{\cdot k}(v)\right)^2 \right] \right\}^{-\frac{1}{2}} \times \\
&\quad \sum_{v=1}^{c} u_k(v) \left(\frac{n_{2k}(v)}{n_2} - \frac{n_{1k}(v)}{n_1} \right) \\
&= \left\{ \frac{1}{n n_1 n_2} \left[n\sum_{v=1}^{c} u_k(v)^2 n_{\cdot k}(v) - \left(\sum_{v=1}^{c} u_k(v) n_{\cdot k}(v)\right)^2 \right] \right\}^{-\frac{1}{2}} \times \\
&\quad \left(\frac{n_1^2 n_2^2}{n^2}\right)^{-\frac{1}{2}} \sum_{v=1}^{c} u_k(v) \left(\frac{n_1 n_{2k}(v)}{n} - \frac{n_2 n_{1k}(v)}{n} \right) \\
&= \left\{ \frac{n_1 n_2}{n^3} \left[n\sum_{v=1}^{c} u_k(v)^2 n_{\cdot k}(v) - \left(\sum_{v=1}^{c} u_k(v) n_{\cdot k}(v)\right)^2 \right] \right\}^{-\frac{1}{2}} \times \\
&\quad \sum_{v=1}^{c} u_k(v) \left(\frac{n_1 n_{2k}(v)}{n} - \frac{n_2 n_{1k}(v)}{n} \right).
\end{aligned}$$

Since the right-hand side corresponds to $\frac{T_k}{\sqrt{\mathrm{Var}(T_k)}}$ with $T_k = \sum_{v=1}^{c} u_k(v)\left(\frac{n_1 n_{2k}(v)}{n} - \frac{n_2 n_{1k}(v)}{n}\right)$, which is the most common form of the CA test statistic (Neuhäuser, 2010), equivalence of S' and the sum of CA test statistics follows directly.

References

Ackermann, M. and K. Strimmer (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics 10*, 47.

Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine 20*, 2709–2722.

Agresti, A. and B. Klingenberg (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C 54*, 691–706.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*, 375–386.

Astin, F., K. Atkin, and A. Darr (2008). Family support and cardiac rehabilitation: a comparative study of the experiences of South Asian and White-European patients and their carer's living in the United Kingdom. *European Journal of Cardiovascular Nursing 7(1)*, 43–51.

Boonen, A., J. Braun, I. E. van der Horst Bruinsma, F. Huang, W. Maksymowych, N. Kostanjsek, A. Cieza, G. Stucki, and D. van der Heijde (2010). ASAS/WHO ICF Core Sets for ankylosing spondylitis (AS): how to classify the impact of AS on functioning and health. *Annals of the Rheumatic Diseases 69(1)*, 102–107.

Bostan, C., C. Oberhauser, and A. Cieza (2012). Investigating the dimension functioning from a condition-specific perspective and the qualifier scale of the International Classification of Functioning, Disability and Health based on Rasch analyses. *American Journal of Physical Medicine and Rehabilitation 91(suppl)*, S129–S140.

Chung, J. H. and D. A. S. Fraser (1958). Randomization Tests for a Multivariate Two-Sample Problem. *Journal of the American Statistical Association 53(283)*, 729–735.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics 10*, 417–451.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society: Series C 60*, 377–395.

Geyh, S., A. Cieza, J. Schouten, H. Dickson, P. Frommelt, Z. Omar, N. Kostanjsek, H. Ring, and G. Stucki (2004). ICF core sets for stroke. *Journal of Rehabilitation Medicine 36*, 135–141.

Goeman, J. J. and L. Finos (2012). The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology 11(1)*, 1–18.

Goeman, J. J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Annals of Statistics 38(6)*, 3782–3810.

Herrmann, K. H., I. Kirchberger, F. Biering-Sørensen, and A. Cieza (2011). Differences in functioning of individuals with tetraplegia and paraplegia according to the International Classification of Functioning, Disability and Health (ICF). *Spinal Cord 49(4)*, 534–543.

Hirji, K. F. (1991). A comparison of exact, mid-p, and score tests for matched case-control studies. *Biometrics 47*, 487–496.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–803.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Holper, L., M. Coenen, A. Weise, G. Stucki, A. Cieza, and J. Kesselring (2010). Characterization of functioning in multiple sclerosis using the ICF. *Journal of Neurology 257(1)*, 103–113.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*, 383–386.

Huang, Y., H. Xu, V. Calian, and J. C. Hsu (2006). To permute or not to permute. *Bioinformatics 22*, 2244–2248.

Kaiser, S. and F. Leisch (2010). orddata: Generation of Artificial Ordinal and Binary Data. *R package, version 0.1*.

Kaizar, E. E., Y. Li, and J. H. Hsu (2011). Permutation Multiple Tests of Binary Features Do Not Uniformly Control Error Rates. *Journal of the American Statistical Association 106(495)*, 1067–1074.

Kalisch, M., B. A. G. Fellinghauer, E. Grill, M. H. Maathuis, U. Mansmann, P. Bühlmann, and G. Stucki (2010). Understanding human functioning using graphical models. *BMC Medical Research Methodology 10*, 14.

Klingenberg, B., A. Solari, L. Salmaso, and F. Pesarin (2009). Testing Marginal Homogeneity Against Stochastic Order in Multivariate Ordinal Data. *Biometrics 65*, 452–462.

Lancester, H. O. (1961). Significance Tests in Discrete Distributions. *Journal of the American Statistical Association 56 (294)*, 223–234.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika 95*, 265–278.

Neuhäuser, M. (2010). *Computer-intensive und nichtparametrische statistische Tests*. Oldenbourg.

Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. Wiley.

Pollard, K. S. and M. J. van der Laan (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference 125*, 85–100.

Romano, J. (1990). On the behaviour of randomization tests without a groupsymmetry assumption. *Journal of the American Statistical Association 85*, 686–692.

Stucki, G. and G. Grimby (2004). Applying the ICF in medicine. *Journal of Rehabilitation Medicine 44(suppl)*, 5–6.

Troendle, J. F., E. L. Korn, and L. M. McShane (2004). An Example of Slow Convergence of the Bootstrap in High Dimensions. *The American Statistician 58(1)*, 25–29.

Tschiesner, U., C. Oberhauser, and A. Cieza (2011). ICF Core Set for head and neck cancer: do the categories discriminate among clinically relevant subgroups of patients? *International Journal of Rehabilitation Research 34(2)*, 121–130.

Ustün, T. B., S. Chatterji, J. Bickenbach, N. Kostanjsek, and M. Schneider (2003). The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation 25(11-12)*, 565–571.

Westfall, P. H. and J. F. Troendle (2008). Multiple Testing with Minimal Assumptions. *Biometrical Journal 50(5)*, 745–755.

Westfall, P. H. and S. S. Young (1993). *Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley.

World Health Organisation (2001). International Classification of Functioning, Disability and Health: ICF.

Zheng, G., J. Joo, and Y. Yang (2009). Pearson's test, trend test, and MAX are all trend tests with different types of scores. *Annals of Human Genetics 73(2)*, 133–140.