Novel Methods for Multivariate Ordinal Data applied to Genetic Diplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity

Knut M. Wittkowski

The Rockefeller University

2003

Novel Methods for Multivariate Ordinal Data Applied to Olympic Medals, Risk Profiles, Genomic Pathways, Genetic Haplotypes, Pattern Similarity, and Array Normalization

KNUT M. WITTKOWSKI

The Rockefeller University, General Clinical Research Center
1230 York Ave Box 322, New York, NY 10021, U.S.A.
kmw@rockefeller.edu, Tel: +1 (212) 327-7175; Fax: +1 (212) 327-8450

Abstract

In many applications, e.g., safety, security, biology, medicine, and pattern recognition, it is rare that a single variable is sufficient to represent all aspects of activity, risk, or response. Since complex systems tend to be neither linear, nor hierarchical in nature, but correlated and of unknown relative importance, the assumptions of traditional multivariate statistical methods can often not be justified on theoretical grounds. Establishing validity through empirical validation is not only problematic, but also time consuming. This paper proposes the use of u-statistics for scoring multivariate ordinal data and a family of simple non-parametric tests for analysis. The scoring method is demonstrated to be applicable to scoring profiles of Olympic medals, adverse events of different severity, and side effects of different category. It is then applied to identifying determinants (genomic pathways) that best correlated with complex responses to an intervention (treatment of psoriasis).

Key words: multivariate, rank test, isotonic regression, hierarchical data, gene expression, overall benefit

1. INTRODUCTION

When analyzing complex phenomena by means of statistical methods, a single measure does often not appropriately reflect all relevant aspects to be considered, so that several measures of influences and/or outcomes need to be considered. Sometimes the definite measure is not easily obtained, so that a set surrogate measures has to be evaluated. At other times, e.g., when the aim is to ameliorate a complex phenomenon, a definitive measure may not even exist. Such problems may arise in many applications, such as assessment of quality-of-life, face recognition, identification of terror attacks, low and high level gene expression analysis, and improvements of database security.

In our first example, we will focus first on a familiar situation, the ranking of countries based on the number of Olympic gold, silver, and bronze medals. Of course, this example could easily be generalized to other applications, where measures with different grades of severity are to be integrated, such as grave, severe, and (relatively) benign side effects of medical treatment or indications of imminent terrorist attacks. Our second example will draw on experiences in the field of medicine, although methods for integrating measures for gene expression along biological pathways could also be applied to indicators of problems along the a line of transmission or the sequential steps of manufacture. We will

* This research was supported in part by a General Clinical Research Center grant (M01-RR00102) from the National Center for Research Resources at the National Institutes of Health and investigator grants AI49572 and AI49832 from the National Institute of Health. The authors would like to thank ASIFA HAIDER, JIM KRUEGER, EDMUND LEE, FELIX NAEF who contributed through inspiring discussions and ALEX PESHANSKY and NOEL GODDARD, who carefully read the manuscript.

focus on the effect of treatment on chronic diseases, in general, and psoriasis, in particular. Psoriasis is a skin disease caused by activation of multiple cell types including keratinocytes, vascular cells, and various types of leukocytes. Treatment efficacy can be measured by histological criteria, by intradermal expression of inflammatory cytokines, or by clinical characteristics, such as redness (vascular response) and scaling (keratinocyte response). Since the advent of micro arrays and RT-PCR, researchers are now interested in genes whose expression is controlled in a concerted fashion and related to the response.[1]

Most multivariate methods are based on the linear model, either explicitly, as in regression, factor, discriminant, and cluster analysis, or implicitly, as in neural networks. One scores each variable individually on a comparable scale, either present/absent, low/intermediate/high, 1 to 10, or z-transformation, and then defines a global score as a weighted average of these scores. In other words, data are interpreted as points in a Euclidian space of (independent) dimensions. The number of dimensions is reduced by assuming the dimensions to be related by a specific function of known type (linear, exponential, etc.), allowing one to determine for each point the Euclidian distance from a hyperspace.

While mathematically elegant and computationally efficient, this approach has shortcomings when applied to real world data. Since the relative importance of the variables, the correlation among them, and the functional relationship of each variable with the immeasurable latent factor 'efficacy', 'safety', 'risk', or 'overall usefulness' are typically unknown, construct validity[2] cannot be established on theoretical grounds. Instead, one needs to resort to empirical 'validation', choosing weights and functions to provide a reasonable fit with a 'gold standard' when applied to a sample. While this allows for a comparison between studies where the researchers agreed to use the same scoring system, the diversity of scoring systems used attests to the subjective nature of this process.

As an alternative, hierarchical procedures have been proposed, where subjects are ordered based on a 'primary' variable first, and only if this fails, a 'secondary' variable is considered. While this may seem less subjective at first, it also has shortcomings. Often, variables can be graded, although there is no absolute hierarchy. For instance, one may count the number of grave, severe, and (relatively) benign events observed during a given period. If there were just one additional grave event in one subject, one may find it unreasonable that the other subject is considered less affected, regardless of the number of severe (yet not grave) events experienced.

Even when the assumptions of the linear model regarding the contribution to and the relationship with the underlying immeasurable factor are questionable, it is often reasonable to assume that each variable has at least an 'orientation', i.e., that, if all other conditions are held constant, an increase in this variable is either 'good' or 'bad'. The direction of this orientation can be known (hypothesis testing) or unknown (selection procedures). In genetics, for instance, having more 'abnormal' alleles may increase the risk (or magnitude) of a disease phenotype. In genomics, a higher expression of several related genes may indicate increased disease activity. When screening for security risks, more indicators for atypical behavior may raise concern to a higher level, in face or voice recognition, more indicators being similar may increase the likelihood of correct identification.

When we were faced with the analysis of anal vs. vaginal contacts as risk factors for sexual transmission of HIV,[3] we presented a partial ordering for dealing with graded and ungraded variables, which allowed to incorporate preexisting knowledge that anal contacts carry more risk without having to ignore the number of vaginal contacts reported. Using the marginal likelihood principle with this partial ordering, we developed a nonparametric method to assess the overall risk of HIV infection based on different types of behavior[3] or the overall protective effect of barrier methods against HIV infection.[4] More recently, we applied this approach to assessing immunogenicity in cancer patients.[5] In short, one determines all rankings compatible with the partial ordering of the observed

multivariate data and then computes a vector of scores as the average across these rankings. While this constituted the first objective approach to the analysis of multivariate ordinal data, because it did not rely on questionable assumptions, it lacked computational efficiency. The computational effort required could be prohibitive even for moderately sized samples, so that approximate solutions had to be sought.

Here, we propose a closely related approach based on u-statistics,[6] which is computationally more efficient. With this approach, individual analyses can often be performed even using spreadsheet software and screening for optimal subsets of explanatory variables becomes feasible without the restrictions imposed by commonly used hierarchical strategies. We then demonstrate, how this method leads to a family of simple non-parametric statistical tests for comparing treatments with respect to several ordinal outcomes, some of which may be graded. For censored data, the resulting tests reduce to those of GEHAN,[7, 8] SCHEMPER,[9] FINKELSTEIN-SCHOENFELD[10] and MOYE[11]. The proposed family of tests applies to stratified designs with two or more treatments, including the WILCOXON/MANN-WHITNEY (WMW) test[12], the KRUSKAL-WALLIS test,[13] and the FRIEDMAN test.[14] It also allows for SCHEFFÈ-type multiple comparisons.[15]

The scoring mechanism and the test are introduced in Section 2. In Section 3, we will demonstrate, how the proposed scores can be utilized within the framework of familiar nonparametric tests. Sections 4 and 5 illustrate the use of the test for two different examples. The first example details the application of the procedure for a scoring system. We have chosen Olympic medals, because of this example allows for a discussion of traditional hierarchical scoring systems[10, 11] in a particularly familiar setting. We will illustrate the shortcomings of these scoring systems, while demonstrating how knowledge about a grading of variables can be accommodated within the proposed u-statistics framework. We will then turn to the analysis of a study in psoriasis, where a treatment response is first scored based on both clinical and histological outcomes and then genomic pathways are sought, which best correlate with the overall treatment response.

2. U SCORES FOR MULTIVARIATE ORDINAL DATA

Our aim is to first develop a computationally efficient procedure to score multivariate ordinal data. We then present simple non-parametric tests for comparing these scores between groups, with an option for stratification and paired comparisons. We will not make any assumptions regarding the functional relationships between variables and the latent factor, except that each variable has an orientation, i.e., that if all other variables are held constant, an increase in this variable is either 'good' or 'bad'.

Throughout this paper, the index *j* will be used for groups and the index *k* for subjects within each group. Thus, each combination (*jk*) characterizes one subject. Whenever this does not cause confusion, we will identify subjects with their vector of $L \geq 1$ observations to simplify the notation.

For the proposed scoring mechanism, each subject $\{x_{jk} = (x_{jk1}, \ldots, x_{jkL})'\}_{j=1,\ldots,p;\, k=1,\ldots m_j}$ is compared to every other subject in a pairwise manner. For stratified designs, these comparisons will be made within each stratum only. When the observed outcomes can be assumed to be correlated with an unobservable latent factor, a partial ordering[16] among the subjects is easily defined. If the second of two subjects has values at least as high among all variables $\ell = 1, \ldots, L$, but higher in at least one variable, it will be called 'superior':

$$x_{jk} < x_{j'k'} \Leftrightarrow \{\forall_{\ell=1,\ldots,L}\, x_{jk\ell} \leq x_{j'k'\ell} \wedge \exists_{\ell=1,\ldots,L}\, x_{jk\ell} < x_{j'k'\ell}\}. \qquad (1)$$

If univariate observations ($L = 1$) are all different, subjects can be ordered (Figure 1a). If ties, i.e., identical observations, are present, two cases need to be considered. Ties may be

due to the underlying phenomenon. Often, however, they are caused by discretization or by observing a discrete surrogate variable for a continuous phenomenon. In both cases, there are three possibilities for each pair of subjects. In the former case, they are '<', '>', or '=' (Figure 1b), in the latter, where ties reflect some ambiguity,[17] they are '<', '>', or '≅' (Figure 1c). Intervals, however, can only be ordered, if they are disjoint, thus their pairwise order may be undetermined. In Figure 1d, it is not known, if the event happening between the first and the third cut-off point (1..3) was, in fact, earlier than the event happening between the second and the third (2..3). The same rationale applies to several ($L>1$) variables (Figure 1e). In either case, the ordering may becomes 'partial', rather than 'complete'. For interval censored data, the order between two subjects is undetermined if $x_{jk1} < x_{j'k'2}$. For multivariate data, the order between two subjects is undetermined if $x_{jk\ell} < x_{j'k'\ell}$ for some variable ℓ, while $x_{jk\ell'} > x_{j'k'\ell'}$ for another variable ℓ'.

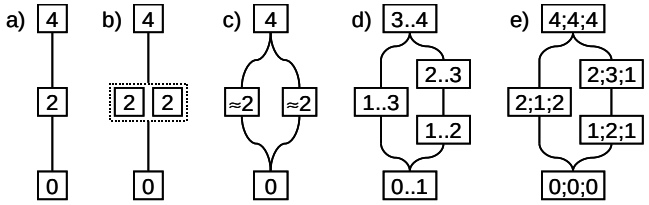


Figure 1: Orderings: a) simple, b) exact, c) inexact, d) interval, e) multivariate.

We will call this partial order 'weak', as compared to its 'strong' cousin

$$x_{jk} \;\; x_{j'k'} \Leftrightarrow \{\forall_{\ell=1,\,\ldots,\,L} x_{jk\ell} < x_{j'k'\ell}\} \tag{2}$$

.

Many partial orderings can be defined, provided that they are transitive, i.e., that $(a<b)\wedge(b<c)\Rightarrow(a<c)$. Since both (1) and (2) treat all variables evenly, we will call these partial orderings 'regular'. The weak regular ordering is the natural ordering for discrete variables. At first sight, the strong ordering may seem to be more appropriate for discretized variables, because the true order of the observations discretized into a tie is unknown. If each variable is presumed to be a surrogate for the same latent factor, however, the strong regular ordering has an undesirable feature. The more variables are included, the more likely it is that at least one of them is tied, i.e., the more pairwise orderings would become undetermined. If, on the other hand, one would make the assumption that the ordering is adaptive hierarchical, i.e., that ties are broken by untied variables, then one obtains the weak partial ordering also for discretized variables. Thus, the weak regular partial ordering (1) will be called 'natural' for applications where each variable can be assumed to be a surrogate for the same underlying latent factor.

Even though a partial ordering does not guarantee that all subjects can be ordered on a pairwise basis, they can all be scored. Let I be an indicator function, i.e.,

$$\mathrm{I}\left(x_{j'k'} < x_{jk}\right) = \begin{cases} 1 & \text{if} \quad x_{j'k'} < x_{jk} \\ 0 & \text{if} \quad x_{j'k'} \text{ and } x_{jk} \text{ cannot be ordered} \\ 0 & \text{if} \quad x_{j'k'} > x_{jk} \end{cases}$$

One can then assign a scores u(x_{jk}) to each subject x_{jk} by simply counting the number of subjects being inferior and subtracting the number of subjects being superior

$$\mathrm{u}\left(x_{jk}\right) = \sum_{j'k'} \mathrm{I}\left(x_{j'k'} < x_{jk}\right) - \sum_{j'k'} \mathrm{I}\left(x_{j'k'} > x_{jk}\right) \tag{3}$$

Figure 1d and e already suggests that interval-censored and multivariate ordinal data can be treated in a similar fashion. To further clarify the relation between censored and multi-

variate data, it is convenient to consider the most general case, interval censored observations. Such data may arise when the exact date of an event is not known, but the event is known to have happened after date x_{jk1} and before date x_{jk2}. Right-censored data are a special case ($x_{jk2} = x_{jk1}$: event, $x_{jk2} = \infty$: censoring). With this interpretation, pairs of subjects can be ordered, if their intervals do not overlap, or, equivalently, if both time points in one subject are earlier than both time points in the other subject:

$$\left\{x_{jk} < x_{j'k'}\right\} \Leftrightarrow \left\{x_{jk2} < x_{j'k'1}\right\} \Leftrightarrow \left\{\max_{\ell}\left(x_{jk\ell}\right) < \min_{\ell}\left(x_{j'k'\ell}\right)\right\}.$$

Subjects, whose intervals overlap, cannot be ordered and, thus, this ordering is partial.

From the third part of this equivalence, it is easy to see that the same partial ordering could be applied to situations, where a measurement is made several times and subject A is considered less affected than subject B if all measurements of subject A are lower than each of the measurements of subject B. Depending on the circumstances, one might alternatively compare subjects based on the average or medians, which yields the WILCOXON/MANN-WHITNEY test based on the within-subject averages or medians:

$$\left\{x_{jk} < x_{j'k'}\right\} \Leftrightarrow \left\{\operatorname{avg}_{\ell}\left(x_{jk\ell}\right) < \operatorname{avg}_{\ell}\left(x_{j'k'\ell}\right)\right\},\ \left\{x_{jk} < x_{j'k'}\right\} \Leftrightarrow \left\{\operatorname{med}_{\ell}\left(x_{jk\ell}\right) < \operatorname{med}_{\ell}\left(x_{j'k'\ell}\right)\right\}.$$

If the sequence in which the measurements was taken provides useful information, one can look at the distribution by comparing the within-subject order statistics:

$$\left\{x_{jk} < x_{j'k'}\right\} \Leftrightarrow \left\{\left(x_{jk1} < x_{j'k'1}\right) \wedge \ldots \wedge \left(x_{jkL} < x_{j'k'L}\right)\right\}.$$

In summary, u statistics can be used whenever a partial ordering can be defined that meaningfully reflects how the observed variables relate to the latent factor.

Some applications may ask for specific partial orderings. For instance, when estimating the signal value for a particular gene on a microarray from a probe set of pairs of perfect and mis-matches, several parametric and semi-parametric ('robust') methods have been proposed. A mis-match (MM) differs from a perfect match (PM) in that a single nucleotide is exchanged for its WATSON-CRICK complement to estimate the non-specific portion of the binding. With low expression levels it is to be expected that random errors in $x_{k,\mathrm{MM}}$ and $x_{k,\mathrm{PM}}$ result in $x_{k,\mathrm{PM}} < x_{k,\mathrm{MM}}$. To allow for a linear model be used based on the logarithms of the differences, it has been suggested by one manufacturer, Affymetrix,[18] to artificially decrease $x_{k,\mathrm{MM}}$ of such probe pairs to a heuristically motivated level that ensures each difference to be positive. Of course, this causes a severe bias for genes with low expression levels, because even a gene that is not expressed at all is guaranteed to yield a positive estimate. When using u statistics, this bias can easily be overcome by employing the following partial ordering:

$$\left\{x_{k} < x_{k'}\right\} \Leftrightarrow \left\{\left(x_{k,\mathrm{PM}} < x_{k',\mathrm{PM}}\right) \wedge \left(-x_{k,\mathrm{MM}} < -x_{k',\mathrm{MM}}\right)\right\}.$$

From this, one selects the pair with a score of zero as the most 'typical', or, if necessary, the average or median among those closest to zero. As this guarantees 'outliers' to be excluded, the believed need for taking logarithms has been overcome. If one is now to request that this estimate be non-negative, the resulting bias would be much lower than if one decreases $x_{k,\mathrm{MM}}$ for each pair where $x_{k,\mathrm{PM}} < x_{k,\mathrm{MM}}$.

When searching for genetic contribution to a disease there are three reasons for the need of multivariate analyses. The first is that the disease gene may be at some distance from the closest marker locus. Thus, both neighboring marker loci may contain information about the disease locus. We will call two adjacent markers a marker 'interval'. Secondly, a disease gene may contain several marker loci, allowing to distinguish between genetic variants differing in pathologic potency. Alternatively, a set of related genes (promoter, etc.) may be located in close proximity, so that each marker in a sequence of markers

contributes information about a sequence of genes contributing to the same phenotype. We will call a set of consecutive markers a 'haplotype'. Finally, a phenotype may be caused by genes being several markers apart, or even on different chromosomes. We will call a set of haplotypes causing a phenotype an 'epistatic set'. The need for a special partial ordering arises from the specific meaning of the term 'interval' in this context.

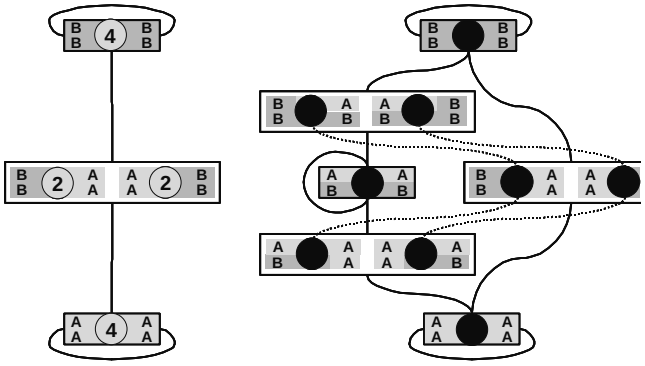


Figure 2: Partial orderings of genetic evidence for an interval between two markers to contain a disease gene G, left: inbred strains, right: outbread strains. Numbers indicate the number of notes that are smaller, greater, or exactly tied. Nodes within boxes are comparable only with nodes connected through a dashed line or through the lines connecting the box, but not among each other.

Let A and B denote the low- and high-risk allele, respectively. One may then scores each marker as $AA = 0$, $AB = 1$, and $BB = 2$. With a traditional, linear model approach one might adjust the score for heterozygous markers between 0 and 2 to linearize the relationship between the score and the assumed risk. If the two alleles were multiplicative, for instance, one might chose $AB = \sqrt{2} \approx 1.41$. Second, one might compute a haplotype score as the weighted average of these marker scores. The pitfalls of this approach are obvious. First, the risk of heterozygous subjects compared to homozygous subjects of either type is typically unknown. Second, even with markers spaced at equal distance in terms of cM, several adjacent markers in a highly conserved region are not more informative as either of them. In a region with more variation, additional markers provide more information. With u statistics, we do not need to make any assumptions regarding the functional relationship between sets of markers and the latent factor 'risk', except that each marker is assumed to have a monotonous relationship with risk, i.e. that $r(AA) \le r(AB) \le r(BB)$. Intervals can then be partially ordered (Figure 2) and these interval scores can then be handled in the manner described above.

In the linear model (Euclidian space) two objects A and B can be ordered with respect to their distance from a reference X by assessing the absolute size of their difference from this reference. For ordinal variables, however, the magnitude of a difference has no meaning and, thus, 'distance' cannot simply be defined in terms of the absolute size of a difference, as in the linear model. If at least some of the variables lack orientation, we will use the term 'pattern', rather than 'profile'. While pattern, in contrast to profiles, cannot be partially ordered with respect to their level, they can still be partially ordered with respect to there distance from a reference. Object A is considered 'closer' to X than B, if the distance of A from X is smaller than the distance of B from the X for each variable, provided that the direction of the deviation from X is the same. This class of partial orderings applies, in particular, to pattern (face or voice) recognition in authentication.

3. ASYMPTOTICS AND TEST STATISTICS

When MANN-WHITNEY,[19] in 1947, proposed their version of what is now known as the WILCOXON[12]/MANN-WHITNEY test, it was one of the first uses of u-statistics. HOEFFDING formalized this concept in 1948 for the one-sample case[6] and in 1951, LEHMANN considered the two sample case. Here, we quote from the latter paper with only minor adjustments with respect to notation and a few simplifications, as did GEHAN[7, 8] in 1965:

Theorem (HOEFFDING/LEHMANN/GEHAN): *Let* $\{\mathbf{X}_{1k}\}_{k=1,\ldots,m_1}$ *and* $\{\mathbf{X}_{2k}\}_{k=1,\ldots,m_2}$ *be independently distributed random vectors, and* $\mathrm{t}(\mathbf{x}_k, \mathbf{x}_{k'})$ *a real-valued function. Let a u-statistic of kernel* t *be defined as*

$$U = \frac{1}{m_1 m_2} \sum_{k=1}^{m_1} \sum_{k'=1}^{m_2} \mathrm{t}(\mathbf{x}_{1k}, \mathbf{x}_{2k'}).$$

If $\mathrm{E}\{\mathrm{t}(\mathbf{X}_1, \mathbf{X}_2)\}$, $\mathrm{E}\{\mathrm{t}^2(\mathbf{X}_1, \mathbf{X}_2)\}$, *and* $\lim_{m_+ \to \infty} m_1/m_2$ *exist, then*

$$\sqrt{m_+}\,(U - \mathrm{E}_0(U)) \xrightarrow[m_+ \to \infty]{} \mathrm{N}(0,1).$$

For univariate data, kernels often are an indicator function, e.g., $\mathrm{t}(x_1, x_2) = \mathrm{I}(x_1 < x_2)$

Note that the observations $\mathbf{x}_k$ were allowed to be multivariate. In the last section of his paper, LEHMAN even pondered tests for several variables, but only for testing the hypotheses of independence and symmetry. When GEHAN[7, 8] applied u statistics to censored observations, however, he viewed them as univariate observations (x_{jk1}: time under study), accompanied by an indicator of precision ($x_{jk2} = 1$: event, $x_{jk2} = 0$: censoring), rather than as multivariate data. (To avoid the notational difficulties related to ties, we assume for the moment a continuous scale.) In 1990, LEE explicitly stated that "*there is nothing in the above theory that requires [the random variables to take values in ℝ], and in fact they may take values in any suitable space.*"[20 p.7] Even so, the potential of u-statistics for the analysis of multivariate data has yet to be fully realized.

Since stratifying the data to allow for scores be computed separately among more comparable blocks of subjects often reduces error variance, we will allow for designs, where subjects are stratified into blocks $i = 1, \ldots, n$. Then a test statistic can be constructed from $U_{ij} = \sum_{k=1}^{M_{ij}} u_{ijk}$ based on the vector $\mathbf{T} = \sum_i \frac{1}{M_i + 1} \mathbf{U}_i$ as a quadratic form $W = \mathbf{T}' \mathbf{V}_0^- \mathbf{T}$, where $\mathbf{V}_0^-$ is a generalized inverse of the variance-covariance matrix of $\mathbf{T}$ under the null hypothesis, as described in reference.[15] The hypothesis of interest is tested by comparing W to a χ^2 distribution with $p - 1$ degrees of freedom.

For uncensored data, this test reduces to a stratified rank test with marginal likelihood block weights,[15] in general, and for binary data to the stratified MCNEMAR[21] test,[22] as a replacement for the TDT.[23] For censored data, the unstratified version of this test reduces to GEHAN's[7, 8] and SCHEMPER's[9] generalizations of the WILCOXON/MANN-WHITNEY and KRUSKAL-WALLIS[13] tests, with additional longitudinal measures, to the test proposed by FINKELSTEIN-SCHOENFELD.[10] The latter paper also proposed a version for stratified designs based on $\mathbf{T}^{(\mathrm{FS})} = \sum_i \mathbf{U}_i$. Since this version does not normalize scores to reflect differences in block size, [15, 24] however, it applies only for designs with equal block sizes.

4. APPLICATION 1: RANKING SUBJECTS BY PROFILES OF GRADED EVENTS

Risk indicators can often be graded by severity. For instance, one might consider the frequency of different types of attempts to break a fire wall (sophisticated, less sophisticated, trivial) as indicators of an attack on a computer system, the frequency of reported prescriptions by type (prescription, non-prescription) as an indicator of an emerging epidemic, the frequency anal, vaginal, and oral contacts as indicators for the risk of HIV transmission,[3] or the number of grave, severe, and (relatively) benign side effects or adverse events as an indicator or risks associated with a treatment. Here, we will consider countries and their medal counts from the 2002 Winter Olympics in Salt Lake City, Utah. Several competing approaches are currently used to rank countries by their medal counts, with different rankings publish in different media based on the same medal counts. We will demonstrate that the results based on u statistics cover a 'middle ground', but, more importantly, that they allow for a ranking being determined that is independent of any subjective weights assigned to the different types of medals.

A total of n = 25 countries C_i won at least one medal at the 2002 Winter Olympics. Four different weighting schemes are commonly used:

Identical:	$IScr = g + s + b$
Linear:	$LScr = 3g + 2s + 1b$
Exponential:	$EScr = 2^2 g + 2^1 s + 2^0 b$
Hierarchical:	$HScr = \left(\lceil \max_i b_i \rceil \lceil \max_i s_i \rceil\right) g + \left(\lceil \max_i b_i \rceil\right) s + b$

where a ceiling $\lceil x \rceil$ is an arbitrary integer larger than x.

Hierarchical weighting schemes are often introduced in a different fashion. Subjects are to be ranked first by the most important criterion (here: gold medals). Variables of lower importance are only used if subjects are tied based on variables of higher importance. Since no weights are explicitly assigned, this seems to avoid some of the shortcomings of linear model weighting schemes. By rewriting hierarchical weighting schemes in the above fashion, however, they are seen as merely a special case of linear model weighting schemes. Since no country had more than 100 medals in any category, 100 can be used here as the ceiling. Table 1 gives the medal counts and the different rankings for above four linear model weighting schemes (uniform, linear, exponential, hierarchical).

Figure 3 shows how the scoring methods affect the ranking of the countries. In this example, the weighting schemes agree only for the most extreme cases, Germany, Slovenia, and Belarus. The difference in ranks may be as high as 6.5 (Austria) and 7.0 (Sweden).

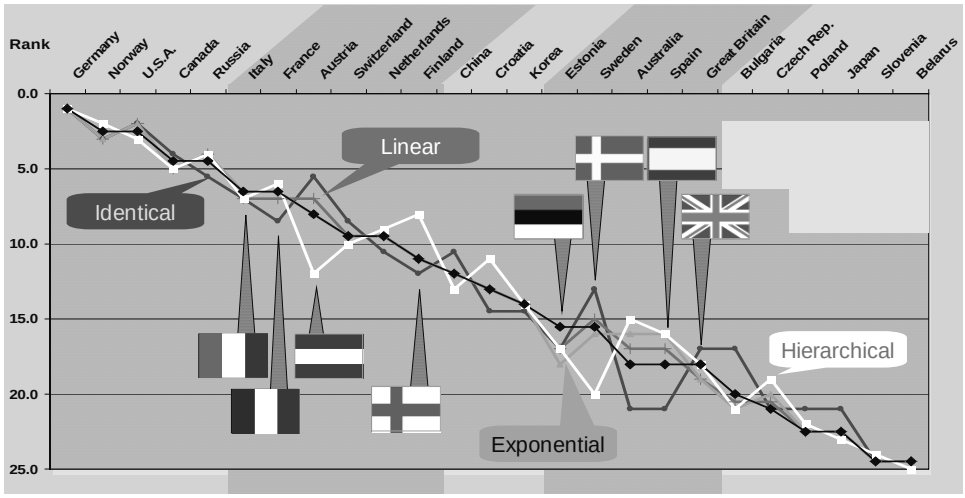
Comparing the U.S. to Norway highlights a shortcoming of hierarchical weights. The U.S. has almost twice as many silver and bronze medals as Norway. Yet, for Norway to have a single gold medal more than the U.S. is sufficient to put Norway in front. The same description (one gold medal more vs. half the number of silver and bronze medals) holds true for Estonia vs. Sweden. Yet, given the absolute numbers, one might argue that one silver and three bronze medals do not compensate for the lack of a gold medal, i.e., that the hierarchical weighting scheme is most appropriate for this comparison. In general, however, weighting hierarchically is not appropriate when it would be desirable to allow the lower importance variables to 'overwrite' the order of the higher importance variables. On the other hand, as the cases U.S. vs. Norway and Estonia vs. Sweden have shown, any fixed set of weights may be difficult to justify. Nonetheless, the higher value of gold vs. bronze medals should be reflected when scoring countries.

Table 1: Medals won at the 2002 winter Olympics in Salt Lake City by country with scores and ranks for different linear model weighting schemes.

Country	g	s	b	IScr	IRg	LScr	LRg	EScr	ERg	HScr	HRg	MnRg	MxRg	dRg
Germany	12	16	7	35	1.0	75	1.0	87	1.0	121607	1.0	1.0	1.0	.0
Norway	11	7	6	24	3.0	53	3.0	64	3.0	110706	2.0	2.0	3.0	1.0
U.S.A.	10	13	11	34	2.0	67	2.0	77	2.0	101311	3.0	2.0	3.0	1.0
Canada	6	3	8	17	4.0	32	5.0	38	5.0	60308	5.0	4.0	5.0	1.0
Russia	6	6	4	16	5.5	34	4.0	40	4.0	60604	4.0	4.0	5.5	1.5
Italy	4	4	4	12	7.0	24	7.0	28	6.5	40404	7.0	6.5	7.0	.5
France	4	5	2	11	8.5	24	7.0	28	6.5	40502	6.0	6.0	8.5	2.5
Austria	2	4	10	16	5.5	24	7.0	26	8.0	20410	12.0	5.5	12.0	6.5
Switzerland	3	2	6	11	8.5	19	9.5	22	9.5	30206	10.0	8.5	10.0	1.5
Netherlands	3	5	0	8	10.5	19	9.5	22	9.5	30500	9.0	9.0	10.5	2.5
Finland	4	2	1	7	12.0	17	11.0	21	11.0	40201	8.0	8.0	12.0	4.0
China	2	2	4	8	10.5	14	12.0	16	12.0	20204	13.0	10.5	13.0	2.5
Croatia	3	1	0	4	14.5	11	13.0	14	13.0	30100	11.0	11.0	14.5	3.5
Korea	2	2	0	4	14.5	10	14.0	12	14.0	20200	14.0	14.0	14.5	.5
Estonia	1	1	1	3	17.0	6	17.0	7	18.0	10101	17.0	17.0	18.0	1.0
Sweden	0	2	4	6	13.0	8	15.0	8	16.0	204	20.0	13.0	20.0	7.0
Australia	2	0	0	2	21.0	6	17.0	8	16.0	20000	15.5	15.5	21.0	5.5
Spain	2	0	0	2	21.0	6	17.0	8	16.0	20000	15.5	15.5	21.0	5.5
Great Britain	1	0	2	3	17.0	5	19.0	6	19.0	10002	18.0	17.0	19.0	2.0
Bulgaria	0	1	2	3	17.0	4	20.5	4	21.0	102	21.0	17.0	21.0	4.0
Czech Rep.	1	0	1	2	21.0	4	20.5	5	20.0	10001	19.0	19.0	21.0	2.0
Poland	0	1	1	2	21.0	3	22.5	3	22.5	101	22.5	21.0	22.5	1.5
Japan	0	1	1	2	21.0	3	22.5	3	22.5	101	22.5	21.0	22.5	1.5
Slovenia	0	0	1	1	24.5	1	24.5	1	24.5	1	24.5	24.5	24.5	.0
Belarus	0	0	1	1	24.5	1	24.5	1	24.5	1	24.5	24.5	24.5	.0

Legend: g/s/b: Number of gold, silver, and bronze medals, respectively. IScr/IRg, LScr/LRg, EScr/ERg, HScr/HRg: Scores and ranks for identical (1:1:1), linear (3:2:1), exponential (4:2:1), and hierarchical (10000:100:1) weighting, respectively. MnRg/MxRg: Minimum and maximum among the four ranks. Shading indicates examples discussed in the text.

Figure 3: Comparison of the four rankings of countries by medal profiles based on the linear model (Table 1).



The above discussion demonstrates the importance of making correct assumptions when devising a scoring scheme. Here, we know that gold medals have an additional, yet unknown, value over silver medals, and silver medals an additional value over bronze medals. The partial 'medal' ordering for such composite variables can be easily defined:

$$C_i >_{medals} C_{i'} \quad \Leftrightarrow \quad (g_i, s_i, b_i) >_{medals} (g_{i'}, s_{i'}, b_{i'}) \quad \Leftrightarrow \begin{cases} \left[\{g_i + s_i + b_i \geq g_{i'} + s_{i'} + b_{i'}\} \wedge \{g_i + s_i \geq g_{i'} + s_{i'}\} \wedge \{g_i \geq g_{i'}\}\right] \\ \text{and} \\ \left[\{g_i + s_i + b_i > g_{i'} + s_{i'} + b_{i'}\} \vee \{g_i + s_i > g_{i'} + s_{i'}\} \vee \{g_i \geq g_{i'}\}\right] \end{cases}$$

Of course, one can obtain the same ordering by describing the success of a county through three derived cumulative variables, the total number of medals ([≥]B), the total number of silver and gold medals ([≥]S), and the total number of gold medals (G). Since these derived variables are now ungraded, the partial ordering of the countries can then be obtained by applying the natural ordering (1).

From this point on, i.e., after obtaining the derived variables, there is nothing specific about Olympic medals in the analysis. In a clinical scenario, by analogy, one might assume that the rows in Table 1 and Table 2 indicating patients instead of countries. The variables G, S, and B could then be thought not only as describing the cumulative counts of other graded variables, such as adverse events, but also as being direct counts of ungraded variables, such as gastrointestinal, sleeping, and breathing problems, respectively. In either case, the ordering of such profiles is only 'partial', because there may exist pairs of subjects that cannot be ordered. Austria and Switzerland, for instance, cannot be ordered. Even subjects that cannot be ordered may receive different u scores, because u scores are based on the pairwise order of each subject with respect to all other subjects.

The (mid) rankings URng based on the u statistic scores UScr for the data in Table 1 are given in Table 2. Ties (countries having the same rank) can be classified into two categories.[17] Australia and Spain form 'inexact' ties, because they have the same pairwise orderings with respect to all other countries. Great Britain, however, though having the same score, can be ordered with respect to Bulgaria and Estonia, while Australia and Spain cannot. Thus, these ties are merely 'circumstantial'. Table 1 shows how the different scoring methods affect the ranking of the countries. Austria and Finland, for instance, are ranked 6:12 with identical weighting or 12:8 with hierarchical weighting.

Table 2: UStat and MrgL Ranking of Countries based on the data in Table 1.

Country	g	s	b	G	S	B	G	N	U	C	R	I	F	A	S	N	F	C	C	K	E	S	A	S	G	B	C	P	J	S	B	#<	#>	UScr	URng
Germany	12	16	7	12	28	35		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	24	0	12	1.0
Norway	11	7	6	11	18	24	-1		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	22	1	11	2.5
U.S.A.	10	13	11	10	23	34	-1	0		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	22	1	11	2.5
Canada	6	3	8	6	9	17	-1	-1	-1		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	3	9	4.5
Russia	6	6	4	6	12	16	-1	-1	-1	0		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20	3	9	4.5
Italy	4	4	4	4	8	12	-1	-1	-1	-1	-1		0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	5	6	6.5
France	4	5	2	4	9	11	-1	-1	-1	-1	-1	0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	5	6	6.5
Austria	2	4	10	2	6	16	-1	-1	-1	-1	-1	0	0		0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	13	5	4	8.0
Switzerland	3	2	6	3	5	11	-1	-1	-1	-1	-1	-1	-1	0		0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14	7	4	9.5
Netherlands	3	5	0	3	8	8	-1	-1	-1	-1	-1	-1	-1	0	0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14	7	4	9.5
Finland	4	2	1	4	6	7	-1	-1	-1	-1	-1	-1	-1	0	0	0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	13	7	3	11.0
China	2	2	4	2	4	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0		0	1	1	1	1	1	1	1	1	1	1	1	1	1	12	10	1	12.0
Croatia	3	1	0	3	4	4	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0		1	1	0	1	1	1	1	1	1	1	1	1	1	1	11	10	1	13.0
Korea	2	2	0	2	4	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		1	0	1	1	1	1	1	1	1	1	1	1	1	10	13	-2	14.0
Estonia	1	1	1	1	2	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	0	0	1	1	1	1	1	1	1	1	1	7	14	-4	15.5
Sweden	0	2	4	0	2	6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0		0	0	0	1	0	1	1	1	1	1	1	5	12	-4	15.5
Australia	2	0	0	2	2	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0		0	0	0	1	1	1	1	1	1	5	14	-5	18.0
Spain	2	0	0	2	2	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0		0	0	1	1	1	1	1	1	5	14	-5	18.0
Great Britain	1	0	2	1	1	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0		1	1	1	1	1	1	1	6	15	-5	18.0
Bulgaria	0	1	2	0	1	3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	-1		0	1	1	1	1	1	4	17	-7	20.0
Czech Rep.	1	0	1	1	1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0		1	1	1	1	1	4	18	-7	21.0
Poland	0	1	1	0	1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	1	1	1	2	21	-10	22.5
Japan	0	1	1	0	1	2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0		1	1	1	2	21	-10	22.5
Slovenia	0	0	1	0	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	1	0	23	-12	24.5
Belarus	0	0	1	0	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	0	23	-12	24.5
				12	11	10	6	6	4	4	2	3	3	4	2	3	2	1	0	2	2	1	0	1	0	0	0	0							
				28	18	23	9	12	8	9	6	5	8	6	4	4	4	2	2	2	2	1	1	1	1	1	0	0							
				35	24	34	17	16	12	11	16	11	8	7	8	4	4	3	6	2	2	3	3	2	2	2	1	1							

g/s/b: Number of gold, silver, and bronze medals, respectively.
G/S/B: Number of gold, gold or silver, and total medals, respectively.
G/N/U/…: Matrix of pairwise orderings.
#</#>: Number of less/more successful countries.
UScr/URng: Scores and ranks by U statistics.

Both the scores based on u statistics and the earlier scores based on the marginal likelihood are valid. While MrgL scores utilize slightly more information than UStat scores, computation of MrgL scores is computationally intensive. 25 ranks, for instance, can be arranged in $25! = 1\times2\times3\times4\times\ldots\times25 = 1.5\times10^{25}$ permutations. Even with testing 1 000 000 permutations per second, this would require 500 billion years. Both scoring systems yield

similar results, however, because the lattice structures are topologically equivalent, i.e., the nodes, the edges, and their direction are the same. In particular, the same pairs of countries are considered exact ties (Norway/USA, Canada/Russia, Italy/France, Switzerland/Netherlands, Australia/Spain, Poland/Japan, Slovenia/Belarus) and, thus, given identical ranks. Under the different linear models, however, the rank ratio for countries within an exact tie may differ. In the above example, the Helvetia:Holland rank ratio ranges from 8.5:10.5 to 10:9. Inexact ties, however, may be affected. Replacing UStat scores by MrgL scores gives Sweden an advantage over Estonia, while eliminating the difference between Bulgaria and Czech Republic.

5. APPLICATION 2: RELATING COMPLEX OUTCOMES TO ACTIVITY PATHWAYS

5.1. Introduction

When trying to identify the factors that, by working together, cause a complex phenomenon such as quality-of-life, overall safety, or overall security, or, at least, allow to predict it, we are faced with several problems. First, most complex phenomena cannot be 'measured' in the traditional sense, because of the lack of a physical scale. Instead, we are faced with several indicators. While it is often reasonable to assume that 'more' is 'worse' for each of them, it may not be easy to determine, how much 'more' is how much 'worse'. Once the effect has been scored, we can identify the set of independent variables that indicate the most likely pathway or constellation causing the complex phenomenon. Again, a 'measure' has to be found to describe the contribution of several factors.

Table 3: Knowledge, data, and intermediate results for the psoriasis example.

Nam	ID	IL12	IFNg	IL8	iNOS	Stat1	ECD3	Km	ET	KH	UKm	UET	UKH	U2	U3	Z2	Z3
SIU									um								
C/S	ID	S AC	S AC	S AC	S AC	S AC	S AC	O AC	O AC	O OC	O OC	O OC	O OC	O OC	O OC	O OC	O OC
Max										1	12	12	12	12	12	12	12
Min		0	0	0	0	0	0	0	0	0	-12	-12	-12	-12	-12	-12	-12
Dat	CG	2	120	2	1	79	149	113	159	0	-6	-4	-7	-5	-5	-4	-4
Dat	JA	60	173	4	11	56	16	105	147	0	-8	-6	-7	-7	-7	-6	-6
Dat	JS	200	0	3	10	125	2	194	103	0	-4	-12	-7	-8	-8	-10	-10
Dat	PO	182	197	289	4	70	53	36	115	0	-12	-10	-7	-11	-11	-12	-12
Dat	CC	48	280	11	0	54	21	50	142	0	-10	-8	-7	-9	-9	-8	-8
Dat	MM	11	227	156	406	160	97	518	223	1	2	-2	6	0	1	6	-2
Dat	JN	253	175	990	51	82	51	926	265	0	6	4	-7	5	2	-2	2
Dat	AR	2	43	35	707	75	97	494	265	1	0	2	6	1	2	8	0
Dat	JR	193	363	37	263	175	49	246	276	1	-2	6	6	2	2	4	8
Dat	PR	148	847	211	409	172	67	612	248	1	4	0	6	2	3	0	4
Dat	DV	144	290	91	176	80	34	964	292	1	8	8	6	8	8	2	6
Dat	DR	933	1	1392	213	182	169	4771	637	1	12	12	6	12	12	12	12
Dat	MG	389	712	249	19	115	170	2013	361	1	10	10	6	10	10	10	10
ACT		M/0	M/0	M/0	M/0	M/0	M/0	Y/+ Y/+ Y/+ Y/+	Y/+ Y/+ Y/+ Y/+	Y/+	U U Test	U U Test	U Test	Um Test	Um Test	RAZ	RAZ

Legend: Beneath the header row with variable names, this table is organized in three horizontal sections, as described.[25] To conserve space, only relevant items are included. The top section contains the knowledge available before the data was conducted; SIU: SI units, C: causality (S: stratum, O: observation), S: scale level (OC: ordinal/continuous, AC: absolute/continuous). Independent and dependent variables are separated by a double line. The second section contains the data, while the bottom section describes the actions to be performed on the data. Each row describes a separate action and the area of derived variables is shaded. The first actions are transformations; RAZ: rank of averaged z-scores, U/Um: univariate/multivariate u-score. "Y/+" indicates that the polarity of these dependent variables is known to be positive, i.e., that higher values are indicative of more inflammatory activity. The independent variables are supposed to have a monotonous influence of unknown polarity (M/0) on the variables to be selected by the Test indicator.

Psoriasis is a complex inflammatory disease characterized by hyperproliferation of keratinocytes and accumulation of activated T-cells in the epidermis and dermis of le-

sions. Treatments with various immunomodulatory or -suppressive agents (e.g., cyclosporine and methotrexate) have a therapeutic index, which precludes long-term treatment. Therefore, there is an ongoing interest in reducing toxicity through targeting cells mediating this disease more specifically. Use of specific antibodies could decrease or inhibit the inflammatory process by blocking activation of T-cells and/or the extravasation of leukocytes.[26] Below, we demonstrate how this goal can be achieved by utilizing u statistics twice, first to score patients with respect to profiles of clinical outcome variables, and then to score various subsets of cytokines to identify the pathway by which the particular agent exerts its anti-inflammatory activity. The data is shown in Table 3, where the lower part outlines the transformations and analyses described below in detail.

5.2. Scoring patients by clinical outcome

Disease improvement in treatment of psoriasis is not easily quantified. The PASI (Psoriasis Area Severity Index) and its variants, while frequently used, are crude measures at best. It is computed by scoring thickness, redness, and scaling on a scale from 0 (none) to 4 (exceptionally striking) for four body areas independently. The sum of these scores is then multiplied by the size of the area (legs: 40%, trunk: 30%, arms: 20%, head:10%) and a score for the estimated percentage of skin involved from 0 (none) to 6 (90-100%). These weighted sums of individual scores are then added to an overall score. One characteristic of the linear model is that the difference between slight and no redness, for instance, is assumed to have the same meaning as the difference between moderate and striking scaling. In the absence of more rational approaches, the PASI has been widely used, although it's shortcomings are well-known. For instance, the accumulation of layers of dead skin cells (scaling), can make it difficult to see the redness underneath. Conversely, with extreme inflammation, scale may be non-adherent and, thus, lesions may appear relatively scale-free. Then, when treatment reduces inflammation, scaling can increase somewhat paradoxically, even though disease activity improves.

One of the major strengths of studying psoriasis as an inflammatory model is the potential to measure therapeutic improvement by more objective criteria. At The Rockefeller University, we have previously defined and categorized clinical response endpoints through immunohistochemical techniques for a large number of standard and experimental therapies,[27-29] so that this is now a well established technique. In this phase I study, responses were measured after treatment as expression of K16 mRNA (*Km*), epidermal thickness (*ET*), and K16 histology (*KH*, 0: negative, 1: positive). The first goal here is to score patients with respect to their overall clinical outcome.

The scores of individual variables (U_{Km} to U_{KH}, Table 3) differ, and none of them stands out as 'the best'. When following traditional approaches based on the linear model, one might derive at a response score by computing the average z-scores of Km and ET and of Km, ET, and KH. Since the proposed approach is non-parametric, we will assume that these scores were then analyzed by non-parametric tests, i.e., we present the ranks of these linear model scores (columns Z2 and Z3, respectively, of Table 3). For patients MM to DV, the linear combination $Z_2 = \text{rank}\left(\text{avg}\left(\text{z}\left(Km\right), \text{z}\left(ET\right)\right)\right)$ has ranks that are outside of the range spanned by $U_{Km} = \text{rank}(Km)$ and $U_{ET} = \text{rank}(ET)$. Looking at the effect of adding K16 histology to form an overall response score $Z_3 = \text{rank}\left(\text{avg}\left(\text{z}\left(Km\right), \text{z}\left(ET\right), \text{z}\left(KH\right)\right)\right)$ highlights this problem with linear model scores. Adding *KH* reverses the order of patients AR and JR, even though K16 histology was positive for both. The reason for this undesirable behavior is that the assumptions of the linear model, which are implicitly made when (a) computing z-scores and (b) averaging them, are not justified here.

Table 4: Computation of u scores U2 for Km and ET (left) and U3 for Km, ET, and KH (right). Patients are ordered by the u scores. Adding variable KH reverses the order of patients JN and PR. Dashed lines separate blocks of patients that can be independently scored. The dashed box in the right diagram indicates an (inexact) tie. In the left diagram, patients PR and JR, in contrast form an accidental tie only.

ID	Km	ET	DR	MG	DV	JN	PR	JR	AR	MM	CG	JA	JS	CC	PO	U2
DR	4771	637		1	1	1	1	1	1	1	1	1	1	1	1	12
MG	2013	361	-1		1	1	1	1	1	1	1	1	1	1	1	10
DV	964	292	-1	-1		1	1	1	1	1	1	1	1	1	1	8
JN	**926**	**265**	-1	-1	-1		1	0	1	1	1	1	1	1	1	5
PR	**612**	**248**	-1	-1	-1	-1		0	0	1	1	1	1	1	1	2
JR	246	276	-1	-1	-1	0	0		0	0	1	1	1	1	1	2
AR	494	264	-1	-1	-1	-1	0	0		0	1	1	1	1	1	1
MM	518	223	-1	-1	-1	-1	-1	0	0		1	1	1	1	1	0
CG	113	159	-1	-1	-1	-1	-1	-1	-1	-1		1	0	1	1	-5
JA	105	147	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	1	1	-7
JS	194	103	-1	-1	-1	-1	-1	-1	-1	-1	0	0		0	0	-8
CC	50	142	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0		1	-9
PO	36	115	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1		-11

ID	Km	ET	KH	DR	MG	DV	PR	JN	JR	AR	MM	CG	JA	JS	CC	PO	U3
DR	4771	637	1		1	1	1	1	1	1	1	1	1	1	1	1	12
MG	2013	361	1	-1		1	1	1	1	1	1	1	1	1	1	1	10
DV	964	292	1	-1	-1		1	1	1	1	1	1	1	1	1	1	8
PR	**612**	**248**	**1**	-1	-1	-1		0	0	0	1	1	1	1	1	1	3
JN	**926**	**265**	**0**	-1	-1	-1	0		0	0	0	1	1	1	1	1	2
JR	246	276	1	-1	-1	-1	0	0		0	0	1	1	1	1	1	2
AR	494	264	1	-1	-1	-1	0	0	0		0	1	1	1	1	1	2
MM	518	223	1	-1	-1	-1	-1	0	0	0		1	1	1	1	1	1
CG	113	159	0	-1	-1	-1	-1	-1	-1	-1	-1		1	0	1	1	-5
JA	105	147	0	-1	-1	-1	-1	-1	-1	-1	-1	-1		0	1	1	-7
JS	194	103	0	-1	-1	-1	-1	-1	-1	-1	-1	0	0		0	0	-8
CC	50	142	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0		1	-9
PO	36	115	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1		-11

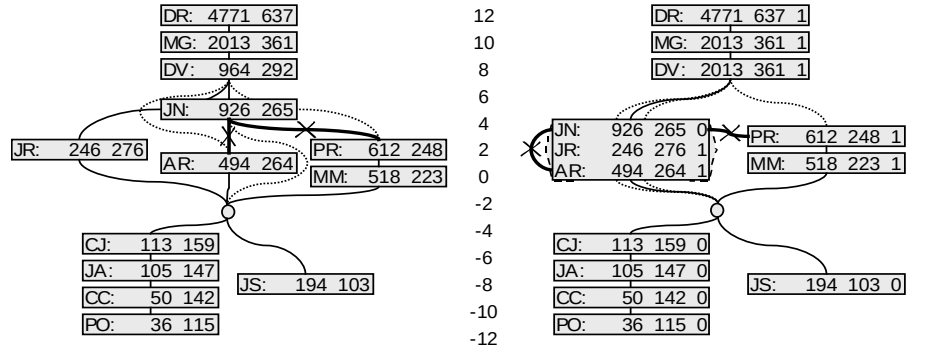


Figure 4: Partial order of bivariate (Km, ET) and trivariate (Km, ET, KH) observations (Table 3). Vertical position determined by u-scores (Table 4) U2 (left) and U3 (right). Crossed out bold lines indicate pairwise orderings being invalidated and dotted lines indicate those becoming relevant when KH=0 is added for patient JN.

U scores are less sensitive to adding a highly correlated binary variable, as one would expect. Only by coincidence is U2 is the average of UKm and UET. However, if variables have a linear relationship with the underlying latent factor, are of known relative importance, and have a fixed correlation, u scores are expected to be close to the ranks of averages weighted for relative importance and correlation.

Looking at the effect of adding K16 histology highlights how u scores differ from linear model scores. Table 4 shows that adding K16 histology reduces the number of pairwise orderings within the center block of patients JN, PR, JR, AR, and MM. In fact, the subset (JN, JR, AR) turns into an inexact tie of indistinguishable patients. Figure 4 illustrates that K16 histology affects neither the complete ordering of the top block of patients (DR, MG, DV) nor the partial ordering of the bottom block of patients (CJ, JA, CC, PO, JS). Within the center block, however, adding KH=0 to patient JN renders the pairs (JN, AR) and (JN, PR) unordered, so that the bold connections in Figure 4a are no longer valid. Still, DV is larger than AR and PR, and JN is larger than the bottom block. Thus, these pairwise orderings, which had previously been inherited through association, are now becoming explicit and are indicated by dashed lines. As a result, JN, JR, and AR now have the same superiors and inferiors, i.e., they form an inexact tie. In Figure 4b, the patients are ordered according to the (fewer) pairwise relations based on the set of three variables. Thus, in contrast to average z-scores, which can seem to be counter-intuitive, scores based on u statistics are easily interpreted.

Moreover, as can also be seen from the definition, u-scores are invariant to scale transformations (logarithms, weights, etc.). Finally, adding highly correlated variables has little effect on the results. If the K16 histology for patient JN had been positive, or the results for patients AR, PR, and MM had been negative, the additional variable would not have affected the results at all.

5.3. Scoring genomic pathways

The variations among patients in their response to treatments can now be used to better characterize the genes directly regulated by the various experimental antibodies. Thus, to gain a better understanding of the potential of cytokines and receptors to contribute to inflammation in psoriatic lesions, we have studied mRNA levels for relevant family members by RT-PCR. We hypothesize that a major inflammatory pathway in psoriasis is regulated by Type 1 T-cells (Th1 & Tc1 subsets). In this case, IL-12 stimulates IL-12R$^+$ T-cells to produce γ-interferon (defining Type 1 T-cells). In turn, γ-interferon acts on keratinocytes to induce synthesis of, among others, IL-8. Hence, we can define one functional pathway as IL-12 → γ-interferon → IL-8. In some patients, however, IL-4 mRNA levels changed in a way that was unrelated to the type 1 genes, suggesting a possible alternative response axis as γ-interferon → STAT1.

In the phase I study described here, inflammatory activity was measured as mRNA expression levels of interleukin-12 (IL12), interferon-γ (IFNg), interleukin-8 (IL8), iNOS, and Stat1. In addition, we measured the concentration of epidermal CD3$^+$ cells (ECD3). The second step in our analysis now is to describe differential gene expression changes in patients and to relate these changes to clinical outcomes as scored above. Since the response on these pathways appear to be controlled by increased mRNA expression for each of their products, expression measures for individual products can be combined into 'profiles', which can then be used to create 'pathway scores' for type 1 inflammatory genes. Of course, the concept of 'pathway scores' pertains to gene expression measures derived from both RT-PCR and expression arrays. Thus, we were interested in identifying pathways of inflammatory processes[30] that explain best differences in multivariate responses to anti-inflammatory antibodies that bind to T-cell surface proteins.

In the first four rows of the bottom block of Table 3, the dependent variables to be scored are indicated by 'Y' and the column for the resulting u-scores by 'U' (univariate) and 'Um' (multivariate), respectively. As each variable is known to have a positive correlation with disease activity (more is worse), the 'polarity' is set to '+'. If one variable had a negative correlation (more is better), one would indicate this by setting the polarity to '-' and the sign of the outcome would be changed before the u-scores are computed.

Clearly, activity profiles along a pathway can be scored in essentially the same fashion as response profiles, except for one additional level of complexity. When scoring responses, it was reasonable to assume that we know, whether 'more' is 'better' or 'worse'. With activity, this is not necessarily true. If treatment were to shift activity from one pathways to the other 'alternative', less effective pathway, 'better' effects may be associated with less activity along the former pathway and more activity along the other. On the other hand, if pathways are synergistic, more activity on either pathway may be 'worse'. Thus, one may wish to allow for various combinations of signs (polarities) to be associated with each set of activity variables. In the bottom line of Table 3, the '0' associated with each independent variable indicates that the polarity is allowed to vary, i.e., for each subset all possible combinations of polarities are to be considered.

5.4. Correlating Activity Pathways with Response Profiles

At the bottom line of Table 3, the response scores computed above are interpreted as dependent variables to determine, which set of pathway variables, when taken together, best explains the response outcomes. Testing for a monotone relationship, as indicated by assigning 'M' to the independent variables, implies SPEARMAN rank correlation. The output of the program is given in Table 5. For each set of independent variables, Table 5 contains one row. Within each of these sets, the all polarities are considered independently for each response variable and the result for the polarity giving the best correlation is given.

U-scores of K16 mRNA expression (RKm) and epidermal thickness (RET) have the highest correlation with a pathway score consisting of IL12, iNOS, and epidermal $CD3^+$ cells (RKm: 0.789 – 0.790 when IL12 is also included, RET: 0.789). When K16 mRNA and epidermal thickness are evaluated together, the highest correlation (0.815) is seen for the same set of inflammatory factors. That the correlation is higher for the combination than for each variable alone further supports the standing hypothesis that changes in these three inflammatory factors affect both response characteristics in a 'concerted action'. K16 histology is remarkably different. Including K16 histology in the response profile reduces the correlation for the set (IL12, iNOS, ECD3), although only marginally from 0.815 to 0.814. For RKH alone, however, a higher correlation (0.849) is seen for a different set of inflammatory factors (IFNg, iNOS, Stat1). Interestingly, for sets of inflammatory factors to have a high correlation with K16 histology, a high level of IL12 expression has to be considered 'protective' as indicated by the '1' in the polarity column [,9].

This suggests that K16 histology is related to a different pathway than K16 mRNA expression and epidermal thickness, a pathway which may be independent of IL12. For instance, K16 histology may reflect effects that preceded the effects measured by acute K16 mRNA expression. Thus, the rows in Table 5 are sorted by column R2, on which the following results will focus. Adding either IFNg, iNOS, or Stat1 to the IL12-iNOS, ECD3 pathway reduces the correlation marginally, but including all six inflammatory agents reduces the correlations from 0.815 to 0.692. Eliminating either IL12 or iNOS lowers correlations to a similar degree, but eliminating ECD3 from the pathway has a small effect on the correlation (0.741). Thus, the data suggests IL12 and iNOS as the most relevant indicators of anti-inflammatory activity in psoriasis.

Interestingly, IL12, by itself, has a very low correlation with U2, the correlation of 0.480 being the second lowest among all pathways. The correlation of iNOS alone (0.582) is only the third highest with respect to R2. If one had selected the inflammatory parameters based on univariate correlations, the average of the RKm and RET scores, one would have chosen (IL8, Stat1) with a correlation of only 0.661, rather than (IL12, iNOS) with a correlation of 0.741. Notably, the set selected by screening all possible sets of inflammatory factors (IL12, iNOS) was disjoint from the set that would have been selected by univariate methods (IL8, Stat). Which pathway to choose has tremendous implications for biological processes. One would predict that both Stat-1 and IL-8 mRNA could be transmitted by the Stat-1 transcription factor activated by IFNg. However, a set including IL-12 and iNOS pair implicates NFκB and Stat-4 as transcription factors.

As an alternative to the exhaustive search through all sets of inflammatory factors, one might have employed a hierarchical approach. In a typical decision tree, one would have first selected the most important factor in univariate analysis, which, in this, case, is Stat:0.613 > max(IL12: 0.480, IFNg: 0.238, IL8: 0.590, iNOS: 0.582, ECD3: 0.541). Among the bivariate sets including Stat, one would have selected Stat/IL8: 0.716 > max(Stat/IL12: 0.624, Stat/ECD3: 0.699, Stat/iNOS: 0.673, Stat/IFNg: 0.567). Thus, a hierarchical analysis would have had no advantage over the simple univariate analysis.

Table 5: Selected pathways of inflammatory genes and number of CD3+ cells in the epidermis, and correlation of their multivariate (1–6 variables) inflammation u scores with multivariate (1–3 variables) u score for response (see Table 3 for variable names and response u scores) sorted by response score U2. Right part: The highest correlation per column is indicated in bold, all correlations at least as high as the smallest among them (0.789) are shaded. Left part: Pathways with the highest correlation with U2 by multivariate u-scores, forward selection, and univariate analysis are boxed. The pathways with the next highest correlations by bi- and univariate correlation are shaded.

Pathway						Correlation with Response ...				
						UKm	UET	UKH	U2	U3
IL12			**INOS**		**ECD3**	0.789	**0.789**	0.676	**0.815**	**0.814**
IL12		IL8	INOS		ECD3	**0.790**	0.773	0.663	0.808	0.806
IL12			INOS	STAT	ECD3	0.765	0.726	0.696	0.771	0.778
IL12		IL8	INOS	STAT	ECD3	0.763	0.705	0.656	0.758	0.765
IL12	IFNG		INOS		ECD3	0.704	0.741	0.668	0.747	0.742
		IL8	INOS		ECD3	0.715	0.727	0.750	0.745	0.767
IL12	IFNG	IL8	INOS		ECD3	0.711	0.727	0.660	0.743	0.737
IL12			INOS		----	0.755	0.678	0.662	0.741	0.735
IL12			INOS	STAT		0.768	0.657	0.658	0.736	0.734
IL12		IL8	INOS	STAT		0.765	0.659	0.653	0.736	0.733
		IL8	INOS	STAT		0.742	0.668	0.754	0.729	0.741
IL12		IL8	INOS			0.733	0.667	0.649	0.723	0.717
		IL8		**STAT**		0.739	0.647	0.589	0.716	0.705
		IL8	INOS			0.710	0.673	0.729	0.714	0.723
		IL8	INOS	STAT	ECD3	0.704	0.677	0.745	0.713	0.736
			INOS	STAT	ECD3	0.707	0.664	0.816	0.708	0.751
IL12			----	STAT	ECD3	0.706	0.655	0.573	0.703	0.700
IL12	IFNG		INOS	STAT	ECD3	0.688	0.671	0.684	0.702	0.708
				STAT	ECD3	0.704	0.650	0.676	0.699	0.728
		IL8		STAT	ECD3	0.699	0.653	0.606	0.699	0.705
IL12					ECD3	0.691	0.657	0.423	0.696	0.679
----			INOS		ECD3	0.663	0.677	0.813	0.692	0.741
IL12	IFNG	IL8	INOS	STAT	ECD3	0.692	0.647	0.675	0.692	0.698
...										
	IFNG		INOS	STAT		0.614	0.605	**0.849**	0.630	0.672
IL12				STAT		0.662	0.546	0.512	0.624	0.602
				STAT		0.654	0.533	0.619	0.613	0.630
	IFNG	IL8		STAT	ECD3	0.579	0.579	0.640	0.598	0.610
IL12	IFNG	IL8		STAT	ECD3	0.608	0.541	0.556	0.594	0.595
	IFNG	IL8			ECD3	0.523	0.625	0.565	0.593	0.598
IL12	IFNG			STAT	ECD3	0.590	0.557	0.590	0.592	0.604
		IL8				0.588	0.555	0.371	0.590	0.556
IL12		IL8				0.603	0.540	0.293	0.590	0.540
			INOS			0.577	0.549	0.825	0.582	0.630
...										
IL12						0.505	0.423	0.124	0.480	0.424
	IFNG					0.176	0.286	0.412	0.238	0.258

6. DISCUSSION

Multivariate ordinal data are frequently used to assess semi-quantitative characteristics, such as risk profiles (genetic, genomic, security) or similarity of pattern (faces, voices, behaviors). Traditional approaches for combining different measures into a utility function or by estimating a common parameter of a joint model require that a relative weight be assigned to each measure. This occurs explicitly when scores are computed as linear combinations of specific functions of the variables, [31] e.g., $s^{(j)}(\mathbf{y}) = \sum_i w_i f_i^{(j)}(x_i)$. Typically, neither the choice of a family $\mathbf{f}^{(j)}$ transformation functions (linear, exponential, polynomial, …), nor a specific family member $f_i^{(j)}$, nor the choice of weights w_i is easily justified. With physical models, these choices may be justified on theoretical grounds. Often, however, the complexity of the system makes such a justification problematic. If an inappropriate model is chosen an analysis based on such a utility function may be misleading. "*This is not very reassuring considered that most models are chosen for their mathematical convenience rather than their biological plausibility*".[32 p.1352f] Other approaches make such assignment implicitly through a hierarchical decision strategy.[11, 32]

When fitting linear models, variables are frequently added or dropped sequentially. For instance, one may look for the most 'significant' variable in univariate analyses first, and

then add more variables in a 'step-up' fashion. Such a strategies, however, may not even come close to the optimum, as we have demonstrated. Tree based approaches (CART[33]), are an alternative, where subjects are separated by the most significant variable first, and each subset is then separated by another subset-specific variable. While this may result in easily communicated decision strategies, step-functions are not more easily justified on theoretical grounds than linear, exponential, or polynomial functions.

In the Olympic medals example, we have demonstrated that hierarchical ranking also may have shortcomings, even when variables are graded. A different approach, also termed hierarchical, is to find the most 'significant' variable in univariate analyses first, and then add more variables in a 'step-up' fashion. This strategy may miss the optimal result and often does not even come close, as we have demonstrated in the psoriasis example. A third class of method, also termed 'hierarchical', are tree based approaches, where subjects are separated by the most significant univariate criterion first, and each subset is then separated by a subset-specific variable next in the hierarchy. This approach, often referred to as CART,[33] has the advantage of resulting in easily communicated decision strategies. Yet, the justification for a step-function is as questionable on theoretical grounds as the justification for a linear, exponential, or polynomial function.

A frequently used attempt to resolve the dilemma of not having a theoretical justification for the model chosen is to use a 'training set' to determine transformations and weights that yield optimal results within this set, and then to check, if the results are 'reasonably good' when this specific scoring system is applied to an 'evaluation set'. If not, one selects another family and/or optimality criterion and tries again. Of course, a set of functions and weights that seems to be 'reasonably good' in the evaluation set is not guaranteed to be optimal. Thus, it has also been suggested that "*if [a] method is to be used, its statistical properties should be examined under different, biologically plausible, alternative distributions by simulation.*"[32 p.1352]

Aside from the lack of a theoretical justification, empirical validation faces practical problems. Many applications require the data to be analyzed in a timely fashion. Consider determining, whether current observations of a set of parameters suggests the onset of a terror and/or hacker attack. If one were to use a 'training set' to determine transformations and weights that yield 'optimal' results within this set, and then to check, if the scoring system is 'reasonably good' when applied to an 'evaluation set', one would need to observe several terror attacks first to train the model (neural network, classification and regression tree, …). Then, one would need to observe even more terror attacks to evaluate the model. Unfortunately, nobody could guarantee that the terrorists are not changing their strategy over time, as implicitly required by this training/evaluation paradigm.

Yet another approach[34] relies on combining univariate test statistics[35] either by forming an omnibus test or a linear combination of test statistics. With such approaches, however, only part of the information contained in the actual profiles is utilized when the data is reduced to univariate statistics and their covariance.

The approach proposed here overcomes the limitations of the above approaches. The advantage of the proposed approach is that no additional assumptions need to be made and validated. Once making the initial transformations has incorporated available knowledge, the proposed scores are valid by construction, as long as each variable increases (or decreases) with the unobservable latent factor. Thus, no empirical evaluation is needed. Since no assumption regarding the functional form of the relationship is made, u scores are scale independent. Moreover, no assumptions need to be made regarding relative importance of variables or correlation among the variables. Relative importance and correlation do not even need to be constant, but may vary with the level of the underlying latent factor. If the variables describe different risk indicators, for instance, other variables may be relevant for low risk subjects, than for high-risk subjects. Adding a highly correlated

variable is unlikely to affect any of the existing pair-wise orderings and, thus, has little or no effect on the scores.

The proposed approach encompasses several important special cases. Interval censored observations can be included by adding both ends of a time interval as individual variables. Thus, a single methodology, in some cases with specific partial orderings, can be used for the comprehensive analysis of many types of indicators, e.g., for quality, safety, security, or risk. If some variables should, in fact, be hierarchical, these variables can be collapsed in the traditional hierarchical fashion, where lower hierarchy variables are used for the sole purpose of breaking ties in variables at higher hierarchy. Whenever justifiable on theoretical grounds, the number of variables may be reduced by replacing some variables by min, max, median, or mean. If events with different grades of severity are counted, be it Olympic medals or adverse events, derived variables can be created that allow for this knowledge to be reflected. In short, one defines one variable that counts the total number of events, a second variable, that counts the number of events having at least grade 2, and so on. These variables of cumulative grade can then be ordered using the natural partial ordering. Interval censored observations can be included by adding the begin and the end of the time interval as individual variables. When considered appropriate for the situation at hand, the number of variables may be reduced by replacing some variables by min, max, median, or mean.

When dealing with other nonparametric method, the computational effort can be prohibitive. The proposed method, however, is computationally simple. From (1), it is clear that the computational effort increases only linearly with the number of variables. Table 2 illustrates an important feature of (3): Adding another subject means that the matrix in the center increases by one row and one column. Thus, the computational effort increases only with the square of the number of subjects.

Having a highly efficient algorithm available allows the method to be used in two important ways. First, individual analyses of relatively small data sets can be conducted in environments better suited for interactive inspection of the data and intermediate results. In the analysis of Olympic medals data, a commercially available spreadsheet program was used to compute u statistics, which can provide profound insight into the nature of the algorithm and, thus, into the understanding of the results. Second, in screening situations (selection procedures), where a large number of combinations of variables is to be analyzed, computational limitations often restrict the points in the multivariate solution space that can be actually evaluated. Hierarchical methods are often employed, because they limit the number of situations to be considered, even though it is well known that such strategies may easily miss the optimal solution. Having a more efficient algorithm reduces the need for employing such sub-optimal strategies. In the analysis of psoriasis data we demonstrated how qualitatively different interpretations may result from an exhaustive search compared to a hierarchical analysis.

The same approach could be directly applied for the comprehensive analysis of other types of indicators, e.g., for quality, safety, security, or risk, as indicated in our recent work on interaction of venom components.[36] Some situations may require specific partial orderings to be chosen. We have demonstrated the use of a particular problem specific partial ordering in the Olympic medal example and proposed partial orderings for genetic haplotypes, where the physical sequence of loci on a chromosome needs to be considered, in pattern recognition (e.g., face or voice recognition), where the directional distance of patterns from a reference needs to be minimized, and in signal value estimation from probe sets on cDNA arrays.

7. OUTLOOK

Previous expert systems aimed at mimicking human decisions through combining univariate heuristics by means of Bayesian law, neural networks, or fuzzy logic.[37, 38] They have not been widely used, however, mainly for two reasons. First, the human experts often did not agree on the heuristical 'certainty factors' to be assigned to the univariate rules and, even if they did, they found the knowledge acquisition process cumbersome. Second, the users did not understand how the certainty factors assigned to the univariate rules affected the multivariate decision process and, therefore, could not control this process. Having an intrinsically valid approach for multivariate data allows for revisiting the 'expert system' paradigm, this time allowing for more transparency (Figure 5).

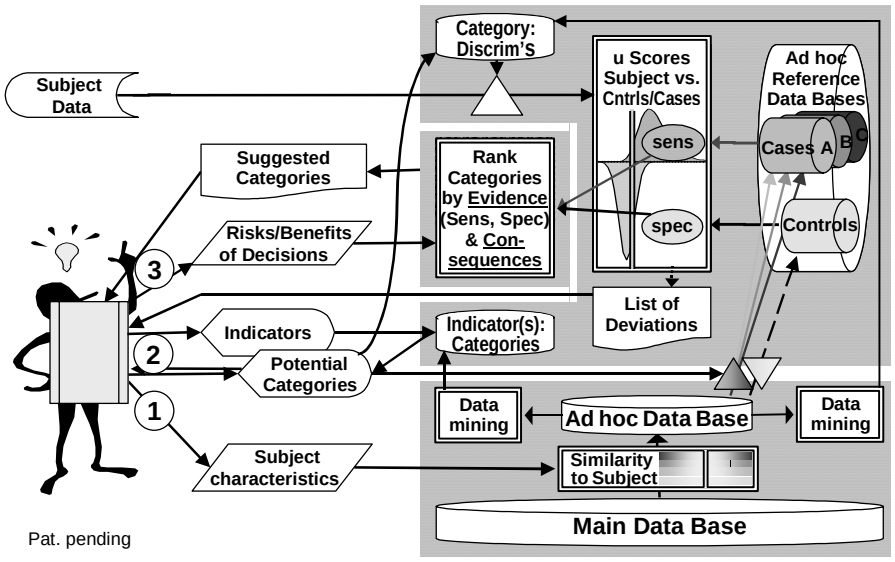


Figure 5: An Expert System approach made possible through the availability of intrinsically valid methods allowing for multivariate ordinal data in selection of 'similar' subsets, data mining to identify categories/discriminators, positioning subjects relative to cases/controls, and ranking categories by evidence and consequences.

Let us assume that a subject (e.g., patient) needs to be categorized (e.g., diagnosed). In the first step, the decision maker (e.g., physician) selects a set of characteristics considered relevant for this subject. Using u statistics for the first time, the system then selects from a main database a subset of reference subjects, which are 'similar' with respect to these characteristics. In a second step, the system uses data mining strategies based on u statistics to present the decision maker with a list of potential categories (e.g., diagnoses), consistent with the subject indicators (e.g., symptoms). Based on prior knowledge, the decision maker extends or curtails this list. The system then extracts reference populations for each of these potential categories from the ad hoc database and determines, which variables best discriminate between these populations. This profile of the subject data is then used to determine the sensitivity and specificity for each category. Finally, after the decision maker assigns relative risks and benefits to the different decisions, the system uses u statistics for the last time to rank decisions by their overall benefit.

With this, we have closed an interesting loop. Earlier, we had suggested to use expert systems to improve statistical database management;[39, 40] now we have demonstrated how to use statistical methods to improve expert systems.

References

1. Nussbaum R, Krueger JG. Treatment of inflammatory dermatoses with novel biologic agents: a primer. *Adv Dermatol* 2002; 18:45-89.
2. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin* 1955; 52:281-302.
3. Susser E, Desvarieux M, Wittkowski KM. Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. *American Journal of Public Health* 1998; 88:671-674.
4. Wittkowski KM, Susser E, Dietz K. The protective effect of condoms and nonoxynol-9 against HIV infection. *American Journal of Public Health* 1998; 88:590-596, 972.
5. Banchereau J, Palucka AK, Dhodapkar M, et al. Immune and clinical responses after vaccination of patients with metastatic melanoma with CD34+ hematopoietic progenitor-derived dendritic cells. *Cancer Research* 2001; 61:6451-8.
6. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 1948; 19:293-325.
7. Gehan EA. A generalised two-sample Wilcoxon test for doubly censored samples. *Biometrika* 1965; 52:650-653.
8. Gehan EA. A generalised Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965; 52.
9. Schemper M. A nonparametric k-sample test for data defined by intervals. *Statistica Neerlandica* 1983; 37:69-71.
10. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Stat Med* 1999; 18:1341-54.
11. Moye LA, Davis BR, Hawkins CM. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Stat Med* 1992; 11:1705-17.
12. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics* 1954; 1:80-83.
13. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952; 47:583-631.
14. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 1937; 32:675-701.
15. Wittkowski KM. Friedman-type statistics and consistent multiple comparisons for unbalanced designs. *Journal of the American Statistical Association* 1988; 83:1163-1170.
16. Wittkowski KM. An extension to Wittkowski [letter]. *Journal of the American Statistical Association* 1992; 87:258.
17. Wittkowski KM. Versions of the sign test in the presence of ties. *Biometrics* 1998; 54:789–791.
18. Hubbell E, Liu W-M, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002; 18:1585-1592.
19. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; 18:50-60.
20. Lee AJ. *U-Statistics*. New York, NY: Marcel Dekker; 1990.
21. McNemar Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrica* 1947; 12:153-157.
22. Wittkowski KM, Liu X. A statistically valid alternative to the TDT. *Hum Hered* 2002; 54:157-64.
23. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; 52:506-16.
24. Wittkowski KM. Small sample properties of rank tests for incomplete unbalanced designs. *Biometrical Journal* 1988; 30:799-808.
25. Wittkowski KM. A structured visual language for a knowledge-based front-end to statistical analysis systems in biomedical research. *Computer Methods and Programs in Biomedicine* 1991; 35:59-67.
26. Krueger JG. The immunologic basis for the treatment of psoriasis with new biologic agents. *J Am Acad Dermatol* 2002; 46:1-23; quiz 23-6.
27. Gottlieb SL, Gilleaudeau P, Johnson R, et al. Response of psoriasis to a lymphocyte-selective toxin (DAB389IL-2) suggests a primary immune, but not keratinocyte, pathogenic basis. *Nat Med* 1995; 1:442-7.
28. Oestreicher JL, Walters IB, Kikuchi T, et al. Molecular classification of psoriasis disease-associated genes through pharmacogenomic expression profiling. *Pharmacogenomics J* 2001; 1:272-87.
29. Trepicchio WL, Ozawa M, Walters IB, et al. Interleukin-11 therapy selectively downregulates type I cytokine proinflammatory pathways in psoriasis lesions. *J Clin Invest* 1999; 104:1527-37.
30. Gottlieb AB, Krueger JG, Wittkowski K, Dedrick R, Walicke PA, Garovoy M. Psoriasis as a Model for T-Cell-Mediated Disease: Immunobiologic and Clinical Effects of Treatment With Multiple Doses of Efalizumab, an Anti-CD11a Antibody. *Arch Dermatol* 2002; 138:591-600.
31. Li K-C, Aragon Y, Shedden K, Thomas Agnan C. Dimension reduction for multivariate response data. *Journal of the American Statistical Association* 2003; 98:99-109.
32. Finkelstein DM, Goggins WB, Schoenfeld DA. Analysis of failure time data with dependent interval censoring. *Biometrics* 2002; 58:298-304.
33. Breiman L. *Classification and regression trees*. Belmont, CA: Wadsworth; 1984.
34. DiRienzo AG, DeGruttola V. Design and analysis of clinical trials with a bivariate failure time endpoint, with application to AIDS Clinical Trials Group Study A5142. *Control Clin Trials* 2003; 24:122-134.
35. Puri ML, Sen PK. *Nonparametric methods in multivariate analysis*. New York: Wiley; 1971.
36. King TP, Jim SY, Wittkowski KM. Inflammatory role of two venom components of yellow jackets (Vespula vulgaris): a mast cell degranulating peptide mastoparan and phospholipase A1. *International Archives of Allergy and Immunology* 2003; 131:25-32.
37. Buchanan BG, Lederberg J. The Heuristic DENDRAL Program for Explaining Empirical Data, IFIP Congress 71, Ljubljana, Yugoslavia, 1971. Vol. 1. North-Holland.
38. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. *Artificial Intelligence* 1993; 61:209-261.
39. Elliman AD, Wittkowski KM. The impact of expert systems on statistical database management. *Statistical Software Newsletter* 1987; 13:14-27.
40. Wittkowski KM. An expert system approach for generating and testing statistical hypotheses. In: Phelps B, ed. Interactions in artificial intelligence and statistical methods. Aldershot, GB: Unicom, 1987:45-59.