NIH Public Access
Author Manuscript
Stat Biopharm Res. Author manuscript; available in PMC 2011 October 5.

Published in final edited form as:
Stat Biopharm Res. 2011 May 1; 3(2): 372–384. doi:10.1198/sbr.2011.10025.

Multiple Hypothesis Testing for Experimental Gingivitis Based on Wilcoxon Signed Rank Statistics

John S. Preisser, **Pranab K. Sen**, and **Steven Offenbacher**
John S. Preisser, is Research Professor, and Pranab K. Sen, is Cary C. Boshamer Distinguished Professor, Department of Biostatistics, and Steven Offenbacher, DDS, PhD, MMSc, is OraPharma Distinguished Professor of Periodontal Medicine, Director, Center for Oral and Systemic Diseases, North Carolina Oral Health Institute, University of North Carolina at Chapel Hill

Abstract

Dental research often involves repeated multivariate outcomes on a small number of subjects for which there is interest in identifying outcomes that exhibit change in their levels over time as well as to characterize the nature of that change. In particular, periodontal research often involves the analysis of molecular mediators of inflammation for which multivariate parametric methods are highly sensitive to outliers and deviations from Gaussian assumptions. In such settings, nonparametric methods may be favored over parametric ones. Additionally, there is a need for statistical methods that control an overall error rate for multiple hypothesis testing. We review univariate and multivariate nonparametric hypothesis tests and apply them to longitudinal data to assess changes over time in 31 biomarkers measured from the gingival crevicular fluid in 22 subjects whereby gingivitis was induced by temporarily withholding tooth brushing. To identify biomarkers that can be induced to change, multivariate Wilcoxon signed rank tests for a set of four summary measures based upon area under the curve are applied for each biomarker and compared to their univariate counterparts. Multiple hypothesis testing methods with choice of control of the false discovery rate or strong control of the family-wise error rate are examined.

Keywords

Area under the curve; Biomarker; False discovery rate; Family-wise error rate; Nonparametric; Permutation; Repeated measures

1. Introduction

Biomedical and dental research often involves the analysis of multiple outcomes recorded on repeated occasions (DeRouen, Hujoel, and Mancl 1995). In many studies, there is often interest in identifying outcomes that exhibit significant change in their levels over time as well as to characterize the nature of that change. In particular, periodontal research often involves the repeated measures analysis of molecular mediators of inflammation. Because such studies are typically performed on a small to moderate number of subjects, nonparametric multivariate methods, which have weaker assumptions and are less sensitive to outliers, may be favored over parametric ones. Additionally, there is a need for statistical methods that control an overall error rate for multiple hypothesis testing.

© American Statistical Association Statistics in Biopharmaceutical Research 2011

correspondence: jpreisse@bios.unc.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

This article presents a novel application of univariate and multivariate nonparametric hypothesis testing procedures to longitudinal data to assess changes over time in biomarker levels in a study of experimental gingivitis. Experimental gingivitis attempts to induce gingivitis by withholding brushing and flossing for a fixed period of time and then to restore good oral health by the resumption of dental care (Burrell and Walters 2008). Levels of multiple biomarkers are measured through repeated sampling of the gingival crevicular fluid over time both during induction phase and afterwards when routine oral health care has resumed. Multivariate tests for a set of summary measures for each biomarker are compared to their univariate counterparts. Special consideration is given to multiple hypothesis testing methods with choice of method directed at either control of the false discovery rate or strong control of the family-wise error rate.

Section 2 introduces the experimental gingivitis study. Section 3 defines summary statistics for change over time based upon area-under-the-curve computations, and the univariate and multivariate rank tests used to assess them (Ghosh, Grizzle, and Sen 1973; Dawson and Siegler 1996). Multiple testing procedures are also outlined. Section 4 reports on the results of the statistical analysis of the biomarkers. Finally, Section 5 concludes with discussion of the proposed nonparametric analysis approach and the utility in this setting of the investigated multiplicity methods.

2. Description of the Experimental Gingivitis Study

In this study of 22 humans, routine oral hygiene is temporarily discontinued to permit the study of localized changes in biofilm overgrowth and inflammation associated with 31 inflammatory mediators (biomarkers). The study recruited and enrolled subjects with naturally occurring gingivitis, defined as bleeding upon probing present, typically in at least 10% of dental sites, as these subjects were more likely to develop experimental gingivitis in the course of the study. Gingivitis is induced by withholding tooth brushing by the use of intraoral acrylic stents that cover selected teeth in each arch during tooth brushing to induce local gingival inflammation. Mediator levels are determined from the laboratory analysis of gingival crevicular fluid. At the end of the induction phase, stents are discontinued and hygiene on all teeth is reinstituted to resolve inflammation over a subsequent resolution period. The study design consists of a one-week hygiene phase (during which routine dental care is encouraged), a three-week induction phase (the period of removal of the benefits of brushing and flossing), and a four-week resolution phase of resumption of daily oral health care. Gingival crevicular fluid is collected from the same oral sites at the beginning of the hygiene phase (or Day −7, one week prior to baseline), the beginning of the induction phase (Day 0 or baseline), Day +7, Day +14, and Day +21 (end of the induction phase/baseline for the resolution phase), and during the resolution phase at Day +35 and Day +49. At the final time point, four weeks after brushing/flossing is resumed, baseline levels are expected to be restored for all biomarkers.

At each time point, gingival crevicular fluid is collected from eight dental sites from the stent teeth and the volume of fluid collected from each sample is recorded. There are four different assays run, and a given biomarker is assessed with only one of these assays, producing two measurements per subject at each time point. For each biomarker assay the mass of each mediator is determined by comparison to authentic standards and a concentration at each site determined based upon the volume collected. The average of the two concentration measurements is taken. These two measurements are always observed together; one is never observed without the other being observed. So the average of the two measurements comprises the data.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Although there is no clinical intervention per se, the goal of the experiment is to identify new candidate biomarkers that are sensitive to poor oral health care as identified by their patterns of change during induction and resolution of gingivitis. The data have been previously analyzed using parametric modeling (Offenbacher et al. 2010). The next section describes the proposed non-parametric statistical analysis of the pattern of response from Day 0 to Day +49 based on subject-level area-under-the-curve summary measures.

3. Statistical Methods

3.1 Summary Indices of Change: Area-Under-the-Curve

A metric scale is assumed for the biomarker levels; the procedures for constructing summary indices of change described here do not apply to strictly ordinal data. Because of heavy skewness, all biomarker data are transformed to log base 10 after adding 1 to the original data. Let L_{ijt} be the log gingival crevicular fluid level of the jth biomarker for the ith subject at the tth time point, for $t = 0, \ldots, 5$, corresponding to Day 0, +7, +14, +21, +35, and +49, respectively. Conceptually, we consider that each biomarker is one of three types. For a positively sensitive biomarker, L_{ij0}, L_{ij1}, L_{ij2}, and L_{ij3} are expected to be in increasing order during the induction (stent) phase, while L_{ij3}, L_{ij4}, and L_{ij5} are expected to be in decreasing order during the resolution (nonstent) phase. The opposite picture holds for a negatively sensitive biomarker, with a decreasing trend during induction followed by a increasing trend during the resolution phase. For positively and negatively sensitive biomarkers, L_{ij5} should be compatible with L_{ij0}. For insensitive biomarkers, the gingival crevicular fluid levels should be comparable across the six time points. Some biomarkers may deviate from these patterns. For example, the levels of a possibly positively sensitive biomarker, IL-1β, corresponding to a single subject shown in Figure 1a are compatible with a pattern whereby the levels peak earlier than Day 21. Next, the levels of a possibly negatively sensitive biomarker, Resistin, shown in Figure 1b, suggest marked asymmetry, whereby levels not only return to baseline after Day 21, but temporarily elevate above baseline. Figure 1 provides a visual aid to the specification of summary indices of change associated with experimental gingivitis; in the case of Figure 1b, areas C and A have negative values, area D has a positive value, and the two regions of B combine to form a region with essentially zero area.

These considerations dictate that the statistical methodology employed should allow for detecting change away from the null in any of these directions. We adopt an area-under-the-curve approach to approximate the average change between two observed time points of the continuously evolving biomarker levels. Formally, for the jth biomarker level from the ith subject, define the change from baseline to time t as $Y_{ijt} = L_{ijt} - L_{ij0}$. With week as the unit for time, define the following summaries of area under the curve:

$$
\begin{gathered}
A_{ij}=(Y_{ij1}+2Y_{ij2}+Y_{ij3})/2\\
=\text{area between day 7 and day 21}\\
B_{ij}=Y_{ij3}+Y_{ij4}\\
=\text{area between day 21 and day 35}\\
C_{ij}=Y_{ij1}/2\\
=\text{area between day 0 and day 7}\\
D_{ij}=Y_{ij4}+Y_{ij5}\\
=\text{area between day 35 and day 49}
\end{gathered}
$$

Next, define four variates of interest to be assessed in the statistical analysis described in the section to follow:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\begin{gathered} X_{ij1}=C_{ij}-\frac{1}{2}D_{ij}=\frac{1}{2}(Y_{ij1}-Y_{ij4}-Y_{ij5}) \\ X_{ij2}=A_{ij}-B_{ij}=\frac{1}{2}(Y_{ij1}+2Y_{ij2}-Y_{ij3}-2Y_{ij4}) \\ X_{ij3}=Y_{ij2} \\ X_{ij4}=Y_{ij4}-Y_{ij5}. \end{gathered} \tag{1}$$

The rationale underlying X_{ij1} and X_{ij2} is to examine whether the rate of induction is the same as the rate of resolution; rejection of the null hypothesis would point to asymmetry. The statistic X_{ij3} examines the rate of induction between Day 0 and Day 14. The statistic X_{ij4} examines the rate of resolution between Day 35 and Day 49. Together, these four variates provide a nearly complete picture of a biomarker's pattern of change over time. In the context of the pattern shown in Figure 1b, for example, X_{ij1} is computed as a positive area corresponding to region D subtracted from a negative area for region C, giving a large area (in absolute terms) with a negative sign. Such deviations from the null pattern of no change over time might not be detected by X_{ij3} or X_{ij4} since these variates do not directly compare induction to resolution phases.

Reducing the dimension of a subject's data for a biomarker from six to four data points should translate into increased statistical power for the alternative statistical hypotheses the summaries target. The logarthmic transformations employed are important to the rank analysis described in the next section because of the differencing used in the definitions in (1) and has the effect of defining the contrasts $(X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4})'$ on the multiplicative scale instead of the additive scale of the untransformed data.

3.2 Univariate and Multivariate Rank Tests

Let $k = 1, 2, 3, 4$ index the variate. We employ, for each biomarker $j = 1, \ldots, J (J = 31)$, a four-variate Wilcoxon Signed Rank Test to $\mathbf{X_{ij}} = (X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4})'$ to examine the four variates simultaneously for departure from their null median values of 0, appealing to its permutation distribution for generating an exact p-value. Let n_j denote the number of subjects without any missing elements in X_{ij} (i.e., the number with complete data vectors).

For the kth variate, let R^{+}_{ijk}, $i = 1, \ldots, n_j$ be the ranks of the absolute value $|X_{ijk}|$ of the jth biomarker, among the n_j subjects. Following Sen (1998), define

$$\operatorname{sign}(X_{ijk})=\begin{cases} -1 & \text{if } X_{ijk}<0 \\ 0 & \text{if } X_{ijk}=0 \\ 1 & \text{if } X_{ijk}>0, \end{cases} \quad \forall\{i, j, k\}$$

and

$$\begin{gathered} W_{jk}=n_j^{-1}\sum_{i=1}^{n_j}\operatorname{sign}(X_{ijk})R^{+}_{ijk}, \\ j=1,\ldots,31; k=1,2,3,4, \end{gathered} \tag{2}$$

and

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$v_{jkk'}=n_j^{-1}\sum_{i=1}^{n_j}\mathrm{sign}(X_{ijk}X_{ijk'})R_{ijk}^{+}R_{ijk'}^{+},\quad j=1,\ldots,31;\quad (k,k')=\{(1,1);(2,2);(3,3);(4,4);(1,2);(1,3);(1,4);(2,3);(2,4);(3,4)\}. \tag{3}$$

Note that assigning sign(X_{ijk}) = 0 for $X_{ijk} = 0$ implies that sign($X_{ijk}X_{ijk'}$) = 0 because sign($X_{ijk}X_{ijk'}$) = sign(X_{ijk}) × sign($X_{ijk'}$). This effectively removes zero values from contributing to calculations of variances and covariances in (3), as well as from means in (2). In the case of ties, only minor adjustments are needed, applying mid-ranks to Equations (2) and (3) (Appendix A).

The four-variate test that all four medians of the variates for a biomarker are simultaneously zero is

$$Q_j=n_j\mathbf{W}_{\mathbf{j}}'\mathbf{V}_{\mathbf{j}}^{-1}\mathbf{W}_{\mathbf{j}}, \tag{4}$$

which will have an asymptotic chi-square distribution with four degrees of freedom (DF), where $\mathbf{V_j} = \{v_{jkk'}\}$ is the 4 × 4 covariance matrix of $\mathbf{W_j} = (W_{j1}, W_{j2}, W_{j3}, W_{j4})'$. Reducing the dimension of a subject's six repeated biomarker measurements to the four variates in $\mathbf{X_{ij}}$ translates into increased statistical power for the alternative statistical hypotheses corresponding to patterns represented by $\mathbf{X_{ij}}$ (possibly with loss of power for patterns of changes not targeted by $\mathbf{X_{ij}}$). The increased power is most readily apparent in Equation (4), since the DF for its asymptotic chi-square distribution have been reduced from six to four.

Alternatively, univariate tests are defined corresponding to the X_{ijk}, four such tests for each biomarker. The univariate testing approach applies the univariate Wilcoxon signed rank test to each of the four measures for all 31 biomarkers resulting in 124 p-values. Whereas a subject with any X_{ijk} missing is omitted from the multivariate test in (4), all available data are used for the univariate tests, substituting n_{jk} for n_j in Equation (2) and for v_{jkk} in (3); note, for the test in (4), $n_j \le \min(n_{j1}, n_{j2}, n_{j3}, n_{j4})$. The univariate Wilcoxon signed rank statistic for the kth variate is $n_{jk}W_{jk}^{2}/v_{jkk}$ which has an asymptotic chi-square distribution with 1 DF under the null hypothesis that the median of X_{jk} is equal to zero.

It is noteworthy that the Wilcoxon signed rank tests have a broader application whereby the null hypothesis corresponds to the joint conditions of a zero median and symmetry of the distribution of differences, whereas the alternative hypothesis is either a nonzero median difference or asymmetry. The nonparametric tests are sensitive to asymmetry in the distribution of difference scores (even when the median is zero), since the difference in mean ranks would not be zero.

Whereas the asymptotic theory for the univariate test holds up fairly well for sample sizes as small as 20, sign invariance methods are needed for small sample sizes, and more so for the multivariate tests. Since the sample size per univariate test ranges from 11 to 22 (because of missing data), we apply exact permutation tests for all univariate and multivariate tests. Let $S_{ijk} = sign(X_{ijk})$ and define the following matrices for the jth biomarker,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathbf{S}_j=(\mathbf{S}_{1j},\ldots,\mathbf{S}_{nj})=\begin{bmatrix} S_{1j1} & \cdots & S_{nj1} \\ \vdots & S_{ijk} & \vdots \\ S_{1j4} & \cdots & S_{nj4} \end{bmatrix},$$

and

$$\mathbf{R}_j^+=(\mathbf{R}_{1j}^+,\ldots,\mathbf{R}_{nj}^+)=\begin{bmatrix} R_{1j1}^+ & \cdots & R_{nj1}^+ \\ \vdots & R_{ijk}^+ & \vdots \\ R_{1j4}^+ & \cdots & R_{nj4}^+ \end{bmatrix}.$$

In these definitions, and for the remainder of the section, we write n instead of n_j to simplify notation. The permutation distribution of the multivariate Wilcoxon signed-rank statistic is given by computing the test statistic Q_j where $\mathbf{R}_j^+$ are held fixed and $\mathbf{S}_j$ can have 2^n (conditionally) equally likely realizations $((-1)^{i1}, \mathbf{S}_{1j}, \ldots, (-1)^{in}\mathbf{S}_{nj})'$, $i_r = 0, 1, 1 \le r \le n$. Note that in this sign-invariance permutation procedure, each $\mathbf{S}_{ij}$ has a single complement $(-1)\mathbf{S}_{ij}$ and together the pair are the only possible outcomes for the *i*th subject in the permutation scheme (see Appendix B for further details and justification). The permutation based *p*-value is then the proportion of the resulting test statistics, say $Q_j^{(1)},\ldots,Q_j^{(2^n)}$ that exceed the value of the observed test statistic Q_j. Note that when n is large, the permutation distribution of Q_j may be approximated using monte carlo methods, that is, taking a large with-replacement sample of the 2^n realizations. However, in our application of either univariate or multivariate signed rank tests, the maximum number of permutations for any test was just over four million (2^{22} = 4, 194, 304), a number which was easily handled by complete enumeration of the permutation distribution using a SAS/IML macro (Sas Institute Inc. 2008) written for the explicit purpose; the macro took less than three minutes to run per test. For either the univariate or multivariate procedure, adjustment for multiple testing may proceed as described in the next section.

3.3 Adjustments for Multiple Testing

For the procedures described above, the set of M *p*-values so obtained, where $M = JK$ or $M = J$ (depending upon the context of univariate or multivariate testing, respectively) may be evaluated for statistical significance, taking into account multiplicity. Traditionally, the overall family-wise error rate (FWER) α may be controlled with the Bonferroni adjustment (Hochberg and Tamhane 1987). The FWER is defined as the probability that at least one true null hypothesis is rejected when any of the null hypotheses hold. Application of the union-intersection principle provides a Bonferroni correction requiring a statistical test to have *p*-value less than α/M to be labeled as significant. Indeed, the overall test size is $\le \alpha$ regardless of whether the M tests are independent or not (Simes 1986; Sarkar and Chang 1997). This test actually provides strong control of the FWER, in that for any subset of the M hypotheses, the probability of falsely rejecting the null hypothesis that all individual hypotheses are true is not greater than α (Hochberg 1988).

Motivated by the stringency of the Bonferroni adjustment we consider alternative multiplicity adjustments including the Bonferroni “step-down” procedure of Holm (1979). This procedure starts by examining the smallest *p*-value $P_{(1)}$. If $P_{(1)} > \alpha/M$, then all null hypotheses are $H_{01}, \ldots, H_{0M}$ are accepted (and procedure stops). Otherwise, if $P_{(1)} < \alpha/M$, then reject $H_{0[1]}$, the null hypothesis corresponding to $P_{(1)}$, and proceed to the second step. In the second step examine the second smallest *p*-value $P_{(2)}$. If $P_{(2)} > \alpha/(M - 1)$, then the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

null hypotheses $H_{0[2]}, \ldots, H_{0[M]}$ are accepted (and procedure stops). Otherwise, if $P_{(2)} < \alpha/(M - 1)$, then reject $H_{0[2]}$ and proceed to the third step. This continues until all remaining null hypotheses are accepted, or until the last hypothesis $H_{0[M]}$ is rejected. Like the standard Bonferroni correction, Holm's sequentially rejective procedure is applicable to both independent and dependent tests. And like the Bonferroni correction, it provides strong FWER control, while being less stringent (or more powerful).

Alternatively, the "step-up" procedure of Hochberg (1988), which provides strong control of the FWER in certain situations, is as follows. Start by examining the largest *p*-value $P_{(M)}$. If $P_{(M)} < \alpha$, then all null hypotheses are rejected. Otherwise, if $P_{(M-1)} < \alpha/2$, then all null hypotheses $H_{0[1]}, \ldots, H_{0,[M-1]}$ are rejected. If not, then compare $P_{(M-2)}$ with $\alpha/3$, and so on. The set of significant tests is given by the set $\{(m): m = 1, \ldots, m^*\}$ determined by the largest (m), say (m^*), such that $P(_{m}) < \alpha/(M - (m) + 1)$.

It is noteworthy that while the Holm and Hochberg procedures contrast the ordered *p*-values with the same set of critical values, the Hochberg procedure is sharper in the sense that the set of individual null hypotheses rejected will always contain the corresponding set of null hypotheses rejected by the Holm procedure. While Holm's procedure places no restrictions on the dependency of *p*-values, the use of Hochberg's method has some controversy in this regard, particularly for multiple endpoints, as in the case of the experimental gingivitis study.

When the aims of a study are exploratory, multiplicity may be alternatively addressed by controlling the false discovery rate or FDR (Benjamini and Hochberg 1995). The FDR is defined as the expected fraction of rejected null hypotheses for which the null hypothesis is true, conditional on the number of rejected hypotheses being greater than zero, multiplied by the probability of having at least one rejection. First, define the ordered set of the *M* *p*-values from smallest to largest, $\{P_{(m)}\}$, and define α as the FDR one is willing to accept. The set of "significant" tests is given by the set $\{(m): m = 1, \ldots, \mathrm{m}^*\}$ determined by the largest (m), say (m^*), such that $P_{(m)} < ((m)/M)\ \alpha$. Because $(m)/M \geq 1/(M - (m) + 1)$ (for $M \geq 2$), with equality only when $(m) = 1$ or $(m) = M$, one can easily see that the FDR bounds are less stringent than Hochberg's FWER bounds. In other words, the FWER controlling procedures described above cannot give more significant tests than the FDR procedure, and will often produce fewer "significant" results.

Both the FWER (Hochberg 1988) and FDR (Benjamini and Hochberg 1995) methods may only be justified under the assumption of independent tests, clearly not the case here since a subject's four variates $X_{ij1}, \ldots X_{ij4}$ for biomarker *j* are correlated as are the tests from the different biomarkers. Nonetheless, these multiple testing procedures are justified in our setting as approximate procedures through appeal to the Chen-Stein Theorem (Chen 1975) as shown by Sen (2008) and along the lines presented in Appendix C, with somewhat better justification for the procedure involving *JK* univariate tests over the one for *J* multivariate tests.

4. Results

Appendix D describes methods and results of simple imputation methods applied to the experimental gingivitis data. Not all cases of missing data had values imputed, as the imputations were based only on information from within-subjects. Missing data methods are not emphasized in this article, as the purpose of the statistical analysis is not to provide a definitive analysis of the gingivitis data, but rather to illustrate multiple hypothesis testing methods for nonparametric tests applied to area-under-the-curve summaries.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Figure 2 displays the mean level of log mediator levels for the 31 biomarkers, displayed in three plots by biomarker type. In Figure 2a, three of the six matrixmetalloproteinases (MMP1, MMP3, MMP13) show a decreasing trend prior to day 21, indicative of a pattern of suppression of levels that characterizes negatively sensitive biomarkers. Among the five adipokines, Figure 2b reveals Serpin-E1 to have a clear pattern of suppression throughout induction and resolution periods, returning to baseline levels at Day 35. The other remaining adipokines, particularly Complement-D, show suppression of levels through the induction period, followed by elevation of levels in the resolution period, before finally returning to baseline levels at Day 49. It should be noted that subject specific trends may exhibit substantial variation from the corresponding population averaged trends for biomarkers that are displayed in the plots of Figure 2.

Next, trends in mean log levels of 20 cytokines are shown in Figure 2c; for ease of interpretation, only five of the cytokines were considered to have sufficiently distinguished patterns to merit identification in the plot by a distinct line pattern. Two of these, IL-1α and IL-1β, appear to be positively sensitive biomarkers; they show increasing mediator levels during the induction period, followed by decreasing levels and a return to baseline values in the resolution period. Three cytokines are negatively sensitive biomarkers. The biomarker MIP-1β displayed the largest negative change in levels during the induction period, with a fairly speedy return to baseline levels in the resolution period as indicated by the asymmetry of patterns of suppression about Day 21. Two cytokines, IL-8 and TNFα, having patterns essentially coincident with one another, show a lesser degree of suppression, bottoming out prior to the end of the induction phase, and returning to baseline levels by Day 35.

Figure 3a displays the (ordered) exact permutation test p-values for the 24 univariate rank tests (among 124 total tests) identified as significant by the procedure controlling FDR at 0.05. Table 1 identifies the 12 biomarkers that have at least one significant univariate test under this criterion, and that together account for the 24 significant tests. Furthermore, tests for X_2 andX_3 are both significant for seven of the biomarkers suggesting asymmetry about Day 21, and suppression of biomarker levels within the first two weeks of induction.

Superimposed on the plot in Figure 3a is the strong FWER criterion at $\alpha = 0.05$ and 0.10 control levels. At an overall FWER $\alpha = 0.\ 05$, the seven univariate tests with the smallest p-values are declared statistically significant; at $\alpha = 0.10$, there are ten significant tests. The fact that the jagged line for observed p-values in Figure 3a crosses over the FWER $\alpha = 0.05$ (or $\alpha = 0.10$) critical boundary line only once indicates that the results for Hochberg and Holm’s procedures were the same. Table 1 identifies the seven biomarkers with at least one univariate test achieving significance at $\alpha = 0.10$.

Figure 3b reveals that 10 biomarkers have exact permutation multivariate test p-values identified as significant at the FDR 0.05 level. Superimposed on the plot is the strong FWER criterion at 0.05 and 0.10 control levels. Under FWER control, five multivariate tests are significant at $\alpha = 0.05$ and an additional three tests are also significant at $\alpha = 0.10$. As for Figure 3a, there are no “multiple crossings” of the jagged line over the boundary lines, indicating that the results for Hochberg and Holm’s FWER procedures were the same. Table 1 identifies the biomarkers having significant multivariate tests.

5. Discussion

In this article, a nonparametric multiple hypothesis testing approach is advocated for the analysis of repeated measures biomarker data. An attractive feature of the proposed approach is its ability to assess a large number of biomarkers relative to the number of subjects; this was illustrated with an experimental gingivitis study where the number of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

biomarkers exceeded the number of subjects. Moreover, computation of exact *p*-values from the permutation distribution of univariate and multivariate Wilcoxon signed rank tests provided an analysis of the small dataset based on only 22 subjects. In consideration of the greater amount of problems caused by missing data in the multivariate testing procedure described, the univariate method is advocated for the experimental gingivitis study, and similar studies. Another potential drawback of the multivariate methods is their susceptibility to singularities, especially for small sample sizes. The univariate methods also may accommodate one-sided tests, although only two-sided tests were employed in the experimental gingivitis study. The primary motivation for the analysis presented here is that analysis of ranks provide tests less sensitive to outliers and Gaussian distribution assumptions than provided by parametric analysis. It is possible to adapt the permutation methods to produce exact *t*-tests as an alternative to rank-based methods.

In the proposed approach, the *p*-values generated by signed rank analysis are analyzed with multiple hypothesis testing procedures. Specifically, the choice of procedure depends upon one's analysis goals. If a study is considered confirmatory in the sense that a high degree of confidence is desired in the identification of significant biomarkers, procedures that control the family-wise error rate in the strong sense are advocated. On the other hand, if the study is exploratory in the sense that it is considered to be a screening study for potentially important biomarkers that would be assessed definitively in a future confirmatory study, procedures that control the false discovery rate (i.e., Benjamini and Hochberg 1995) are advocated as these will yield a larger set of candidate biomarkers than would be obtained under an FWER controlling procedure. The FDR procedure is better suited than the FWER procedures for the experimental gingivitis study given the large number of biomarkers and tests.

Given either FWER or FDR α-control, the various multiple testing procedures gave similar results for the experimental gingivitis data. There were no differences among them for the univariate tests, with mild differences among them for the multivariate tests, where the number of tests was smaller ($M = 31$ vs. $M = 124$ univariate tests). We noted that, unlike the Bonferroni and Holm procedures, the Hochberg and Benjamini-Hochberg procedures strictly apply to independent tests with further justification necessary for their application to dependent variates; to this end, recent advances in multiple testing considerations of Sen (2008) provided justification, which is stronger for the univariate testing situation considered than the multivariate testing procedure. Further investigation, with attention given to the patterns of dependency among tests, is needed to compare the multiple hypothesis testing procedures in our setting and to determine whether Poisson approximation-based adjustments to Hochberg's or Benjamini-Hochberg's multiple hypothesis testing procedures along the lines described in Appendix C lead to improved inference for problems similar to the experimental gingivitis study.

With respect to experimental gingivitis, the interpretation of X_1, X_2, X_3, and X_4 as reflecting symmetry or asymmetry has potential implications relating to the biology of the system. Biomarkers that generally increase in concentration may demonstrate monotonic increases as a log function proportional to the stimulus concentration. After the removal of the stimulus a first-order decay in biomarker level might be expected to reflect passive diffusion or transport from the local tissues and would be reflected in symmetry. Asymmetric changes would reflect active transport from the site following the removal of the stimulus or metabolic destruction. When considering the decrease in biomarker level during the induction of disease the inhibition of basal synthesis would logically follow a similar pattern with an uncoupling of synthesis during induction and recoupling during resolution reflecting a symmetrical pattern. Asymmetry would likely be expected from secondary signals which are involved with feedback to suppress the inhibitors to reactivate synthesis and restore basal

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

secretory levels. This is undoubtedly an oversimplification of the process involved, but this example provides potential insight to discriminate whether there are differences in the homeostatic mechanisms which regulate the steady-state levels of different biomarkers.

An important limitation of the proposed nonparametric approach is that it is a hypothesis testing approach and so it does not provide estimation of the mean (or median) response pattern over time. To draw interpretations for specific biomarker patterns over time, the hypothesis testing results were supplemented with appeal to graphics showing average trends in the observed data. Thus, the proposed nonparametric multiple hypothesis testing procedures are complementary to fully parametric model-based methods in the sense that their prospective strengths and weaknesses are opposite. Whereas parametric or semiparametric longitudinal nonlinear modeling approaches provide direction estimation of the trends in biomarkers over time (which the nonparametric testing method does not), their validity depends on strong assumptions regarding the form of the mean model (including uncertainty pertaining to the transformation of the response) and the missingness process, as well as assumptions about the nature of left truncation of observations due to a lower detection limit. Conversely, the nonparametric method appears well suited for identifying biomarkers that have important roles in disease processes in the presence of such challenging and messy data.

In this regard, a limitation of the statistical analysis of the experimental gingivitis study was that it used deterministic (within-subject) imputations for selected missing data (Appendix D). However, we conjecture that the use of permutation tests in the proposed procedures ameliorates many of the usual problems (i.e., underestimation of variances) encountered when deterministic imputation methods are employed, say, with many parametric analysis methods. A comparison of the proposed nonparametric procedures with parametric ones in both balanced and unbalanced (missing data) situations like those addressed in this article warrants investigation.

Acknowledgments

This research was supported by NIH grants RR00046 (GCRC) and U54RR024383 (CTSA). The authors thank Yunro Chung for assistance producing the figures, and two anonymous referees for comments that led to improvements in the article.

References

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Pwerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300. 377, 379, 382.

Burrell RC, Walters JD. Distribution of Systemic Clarithromycin to Gingiva. Journal of Periodontology. 2008; 79(9):1712–1718. 372. [PubMed: 18771373]

Chen LHY. Poisson Approximation for Dependent Trials. Annals of Probability. 1975; 3:534–545. 377, 382.

Dawson DV, Siegler IC. Approaches to the Nonparametric Analysis of Limited Longitudinal Data Sets. Experimental Aging Research. 1996; 22:33–57. 373. [PubMed: 8665986]

DeRouen TA, Hujoel PP, Mancl LA. Statistical Issues in Periodontal Research. Journal of Dental Research. 1995; 74:1731–1737. 372. [PubMed: 8530733]

Ghosh M, Grizzle JE, Sen PK. Nonparametric Methods in Longitudinal Studies. Journal of the American Statistical Association. 1973; 68:29–36. 373.

Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1988; 75:800–802. 376, 377, 382.

Hochberg, Y.; Tamhane, AC. Multiple Comparison Procedures. New York: Wiley; 1987. p. 376

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979; 6:65–70. 376, 382.

Koch GG, Sen PK. Some Aspects of the Statistical Analysis of the ‘Mixed Model’. Biometrics. 1968; 24:27–48. 381. [PubMed: 5642408]

Offenbacher S, Barros S, Mendoza L, Mauriello S, Preisser J, Moss K, de Jager M, Aspiras M. Changes in Gingival Crevicular Fluid Inflammatory Mediator Levels During the Induction and Resolution of Experimental Gingivitis in Humans. Journal of Clinical Periodontology. 2010; 37:324–333. 373. [PubMed: 20447255]

Puri, ML.; Sen, PK. Nonparametric Methods in Multivariate Analysis. New York: Wiley; 1971. p. 381p. 382

Sarkar SK, Chang C-K. The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics. Journal of the American Statistical Association. 1997; 92:1601–1608. 376.

SAS Institute Inc. SAS/IML 9.2 User’s Guide. Cary, NC: SAS Institute Inc; 2008. p. 376

Sen, PK. Multivariate Median and Rank Sum Tests. In: Armitage, P.; Colton, T., editors. Encyclopedia of Biostatistics. New York: Wiley; 1998. p. 2887-2900.p. 375

Sen PK. Kendall’s Tau in High-Dimensional Genomic Parsimony. IMS Collections. 2008; 3:251–266. 377, 380, 382.

Simes RJ. An Improved Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 1986; 73:751–754. 376.

Westfall PH, Wolfinger RD. Multiple Tests with Discrete Distributions. The American Statistician. 1997; 51:3–8. 382.

Appendices

A. Adjustment of Test Statistics for Ties

If there are r_{0jk} observations for which $X_{ijk} = 0$, not only for each of them we have Sign(X_{ijk}) = 0 but also their mid-rank score is $(r_{0jk} + 1)/2$. In the same way, if r_{hjk} absolute values of the $X_{ijk}(1 \le i \le n_j)$ are tied at x_{hjk} and if there are R^{*+}_{hjk} observations with absolute values less than x_{hjk}, then the midrank score for these tied observations is $R^{*+}_{hjk}+(r_{hjk}+1)/2$. This adjustment for ties, though routine, affects the sum of squares $\sum_{i=1}^{n_j} i^2$ and reduces it to

$$\sum_{s=1}^{q_{jk}} r_{sjk}[R^{*+}_{sjk}+r_{sjk}/2]^2,$$

where q_{jk} = # of distinct order values of the $|X_{ijk}|$, $1 \le i \le n_j$. Therefore, the presence of ties reduces the null variance of the W_{jk} (without affecting their mean (=0)). When standardizing the W_{jk}, this adjustment is necessary to use asymptotic critical levels and p-values. In the same way, for the multivariate case, denoting $\mathbf{V}_{Hj}$ as the hypothetical covariance matrix in the absence of ties for which the diagonal elements equal $\sum_{i=1}^{n_j} i^2$, adjustment for ties leads to a covariance matrix $\mathbf{V}_j$ such that $\mathbf{V}_{Hj} - \mathbf{V}_j$ is positive semi-definite. Thus, in (4), for defining the Q_j we need to use the $\mathbf{V}_j$ not $\mathbf{V}_{Hj}$. In small samples, for W_{jk}, the permutation distribution will be generated by the 2^{n_j} sign-inversions as is elaborated in Appendix B. The permutation-based p-value can also be obtained with reduced computational effort using the permutation distribution generated by the $2^{n_j-r_{0jk}}$ sign-inversions with removal of the r_{0jk} observations for which $X_{ijk} = 0$. This equivalency arises because the set of test statistics resulting from the $2^{n_j-r_{0jk}}$ sign-inversions is replicated by the factor $2^{r_{0jk}}$ to produce the full set.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

B. Null Distributions of Test Statistics

In this study, in the complete case, we have, for each of n (= 22) subjects, J (=31) biomarkers and K (=4) response variates (X_1, X_2, X_3, X_4). Thus, we have the set of responses

$$X_{ijk}, i=1,\ldots,n; j=1,\ldots,J; \text{and } k=1,\ldots,K.$$

For any i, the X_{ijk} ($1 \le j \le J$, $1 \le k \le K$) are not independent. Thus, we let

$$\mathbf{X}_i=(X_{i11},\ldots,X_{i1K},\ldots,X_{iJ1},\ldots,X_{iJK})',$$

for $i = 1, \ldots, n$. The (joint) distribution of X_i is denoted by $G(\mathbf{x})$, $\mathbf{x} \in \Re^{JK}$. Under the global null hypothesis H_0, all the JK marginal distributions (of G) are symmetric about 0, thus prompting us to incorporate the classical Wilcoxon signed-rank statistic along with its multivariate generalizations (Puri and Sen 1971) in our testing problem. We define the coordinate-wise W_{jk} as in (2) and their multivariate versions $Q_j(i \le j \le J)$ as in (4).

For each marginal W_{jk}, we have the basic property of exact distribution-freeness (under H_0); however, as the W_{jk} are not necessarily independent, they fail to be jointly distribution-free (under H_0). For this reason, in the multivariate tests based on the Q_j, we advocate a basic permutation (sign-invariance) procedure (as in Koch and Sen 1968 and Puri and Sen 1971). This, however, needs a slightly more stringent regularity assumption that we formulate first.

Let G_j ($\mathbf{x}$), $\mathbf{x} \in \Re^K$ be the K-variate marginal distribution of the jth biomarker set ($1 \le j \le J$). We say that G_j ($\mathbf{x}$) is diagonally symmetric about $\mathbf{0}$ if $\mathbf{X}_{ij}$ and $(-1)\mathbf{X}_{ij}$ both have the same distribution $G_j(\mathbf{x})$. This assumption holds for many multivariate (symmetric) distributions, including the multi-normal law.

Let $\mathbf{S}_{ij} = (\text{Sign}X_{ij1}, \ldots, \text{Sign}X_{ijK})'$, $1 \le j \le J$, $1 \le i \le n$. Also, for each $j(= 1, \ldots, J)$ and $k(= 1, \ldots, K)$, we define the R^+_{ijk}, $1 \le i \le n$ as before in (2), and let

$$\mathbf{R}^+_{jk}=(\mathbf{R}^+_{1jk},\ldots,\mathbf{R}^+_{njk})', 1 \le j \le J, 1 \le k \le K.$$

Under the null hypothesis, letting

$$\mathbf{R}^+_j=(\mathbf{R}^+_{j1},\ldots,\mathbf{R}^+_{jK})', 1 \le j \le J,$$

we have (Puri and Sen 1971) $\mathbf{S}_j = (\mathbf{S}_{1j}, \ldots \mathbf{S}_{nj})'$ and $\mathbf{R}^+_j$ (conditionally) independent where $\mathbf{S}_j$ can have 2^n (conditionally) equally likely realizations $((-1)^{i1}, S_{1j}, \ldots, (-1)^{i\mathrm{n}}\mathbf{S}_{nj})'$, $i_r = 0,1,1 \le r \le n$. This conditional law generates a conditionally distribution-free test (based on Q_j). Further, if n is large, this permutation distribution can be well approximated by the (central) chi-square distribution with K degrees of freedom.

We may note that for the marginal W_{jk}, under H_0, the normal approximation is quite fast—even for $n \ge 12$, it works out well. However, for the Q_j, the chi-square (K DF) approximation may be a bit slower (in n), especially if K is not small. Thus, the use of the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

JK marginal W_{jk} can be advocated with two advantages. First, the marginal distributions are known, well tabulated, and quickly convergent to normal ones. Second, they also allow varying missing observations across the biomarkers as well as variates. In the case of the Q_j, missingness patterns varying over $j(= 1, \ldots, J)$ are allowed, but not over the *K* variates, all of which are required to be observed.

C. Control of Errors in Multiple Hypothesis Testing

For control of the FWER error rate, Holm's procedure (1979) was motivated by the stringency of the Bonferroni adjustment, as was Hochberg's method (1988). For dependent tests, Hochberg's procedure requires justification, which has been provided by Sen (2008). That justification is summarized here, first giving consideration to a simpler case.

A refinement of Bonferroni's bounds may be derived by using a Poisson approximation known as the Chen-Stein Theorem (Chen 1975). We define

$$W_M = \sum_{r=1}^{M} I(P_r < \alpha^* / M),$$

for some α^* that determines the significance criterion for the individual tests. Under the global null hypothesis, $W_M \sim \text{Poi}(\alpha^*)$. In words, the number of falsely rejected individual null hypothesis is approximately Poisson distributed with mean α^*. The approximation applies to many multiple testing problems, since asymptotic independence holds for general dependence patterns when the degrees of freedom (e.g., *M* for the multiple univariate tests) is large, more so for small α. Now, $P_0\{W_M \geq 1\} = 1 - \exp(\alpha^*)$. Setting the FWER to α, α^* is determined by solving $1 - \exp(\alpha^*) = \alpha$, giving the adjusted Bonferroni criterion for individual hypothesis tests as α^*/M. For example, $\alpha = 0.05$ gives $\alpha^* = 0.05129$. Then adjusted Bonferroni levels are 0.05129/31 = 0.001655 (compared to the unadjusted Bonferroni value of 0.05/31 = 0.001613) for the Q_j and 0.05129/124 = 0.000414 (compared to the unadjusted Bonferroni value of 0.05/124 = 0.000403) for the W_{jk}. These slight adjustments, when using $\alpha = 0.05$ (or $\alpha = 0.10$) applied to the experimental gingivitis data, produce identical results in terms of sets of significant tests as the unadjusted Bonferroni procedure. Lastly, adjustments for the discreteness of the distributions (e.g., Westfall and Wolfinger 1997) were not required because inference was based on a complete enumeration of the conditional permutation distribution.

In adapting the Hochberg (1988) and Benjamini and Hochberg (1995) procedures to control the FWER or FDR, respectively, based on Q_j (or W_{jk}), one must consider that these statistics are not generally independent (nor necessarily positively associated). Generally, in the context of high-level crossing probabilities for multivariate normal distributions, Sen (2008) demonstrated the applicability of the Chen-Stein Theorem and has shown that it leads to a discrete Poisson process. Heuristically, the theorem when applied to a bivariate normal distribution with nonzero correlation provides that the largest marginal order statistics are asymptotically independent. Applying the Chen-Stein Theorem to dependent binary variables, some advances in the multiple hypotheses testing problem have been recently made (Sen 2008), providing justification for both the FWER Hochberg and FDR Benjamini-Hochberg multiple hypothesis testing procedures. Future work will investigate modifications of the Hochberg (1988) and Benjamini and Hochberg (1995) procedures based on the Chen-Stein Theorem to the experimental gingivitis data and similar problems.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

D. Details of Data Imputation

Data were available for 33 biomarkers measured from the gingival crevicular fluid for 25 subjects. Due to instances of excessive amounts of either missing data or data below the lower detection limit, the analysis considers data from J= 31 biomarkers measured on n = 22 subjects. Because not all subjects had all six time points observed for all 31 biomarkers, some preliminary data imputations were carried out in order to avoid very small sample sizes in the data analysis and limit the amount of information that would be discarded due to incomplete data. There were three different types of imputations, and all used only within-subject information: imputations for left-censored data using the midpoint between the recorded value of zero and a lower detection limit, imputation by substitution, and imputation by linear interpolation. The latter two methods used only within-subject information for imputations.

It is noteworthy that while left-truncation has implications for the symmetry of the observed biomarker response data L_{ijt}, the data X_{ijt} being used in the univariate and multivariate Wilcoxon signed rank tests are much less affected by left truncation since these correspond to differences. Thus, the assumptions of the tests in Appendix B can be reasonably made.

First, since recorded biomarker levels of zero indicate values below a biomarker's lower detection limit, we imputed the log biomarker response level L_{ijt} as the log base 10 applied to half the lower detection limit plus 1. There were 122 subject-visits observed (5.54 visits per subject). The total amount of recorded zeros varied substantially across biomarkers: 15 biomarkers had no zeros recorded, 6 biomarkers had eleven or fewer (less than 10%) 0 observations, two biomarker had between 17 (14%) to 29 (24%) zeros, 4 biomarkers had between 51 (42%) and 68 (56%) zeros, and 4 biomarkers had between 77 (63%) and 85 (70%) zeros. Indeed, most ties in this data correspond to zero values for the variates X_{ijk} resulting from the differencing operations in (1) applied to fixed biomarker-specific values assigned for left truncation.

Next, a small number of missing biomarker values were imputed using a substitution method. Specifically, if a biomarker level at Day 0 was missing, we substituted the Day − 7 value, and if that was missing, the Day +49 value. If a biomarker level at Day +49 was missing, we substituted the Day 0 value, and if that was missing, the Day − 7 value. The justification for these substitutions is that levels of a biomarker are expected to return to baseline at the end of the resolution phase. Applying these imputations, one subject that did not have gingival crevicular fluid collected at Day 0 had Day − 7 data used for Day 0. Another subject had Day +49 data used for Day 0.

Finally, linear interpolations using data from previous and next visits were performed sequentially for biomarker levels missing at Day +7, Day +35, Day +21, and Day +14, in that order. When insufficient data (real or imputed) was unavailable at any give stage, the data was left missing. Two subjects did not have any data for one of the biomarkers, and there were 52 instances when a subject had zeros recorded for all visits for a biomarker; these observations were omitted. Thus, across all biomarkers, there are $22 \times 31 - 2 - 52 = 628$ sets of (possibly incomplete) longitudinal biomarker levels.

The final results of imputation are as follows. Over all six time points, there are a possible $628 \times 6 = 3768$ observations, of which 3222 (85.5%) were observed, 480 (12.7%) were imputed, and 66 (1.8%) were missing. Table A.1 shows that highest amount imputations were made at Day 7, followed by Day 0 and then Day 35. Few imputations (< 5%) were made at Days 14, 21 and 49.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

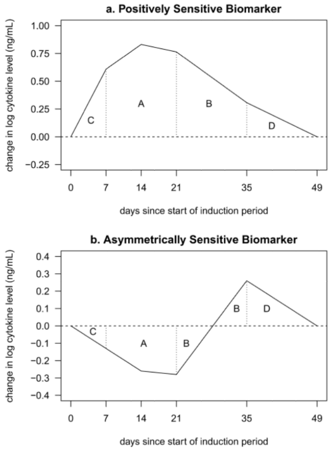

Figure 1.

Experimental gingivitis biomarker levels from single subjects for IL-1β (Figure 1a) and Resistin (Figure 1b). Letters denote a partition of area under the curve from which four summary indices of interest (X_{i1}, X_{i2}, X_{i3}, X_{i4}) are computed.

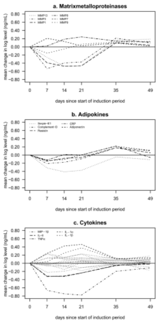

Figure 2.

Mean changes in biomarker levels from Day 0.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

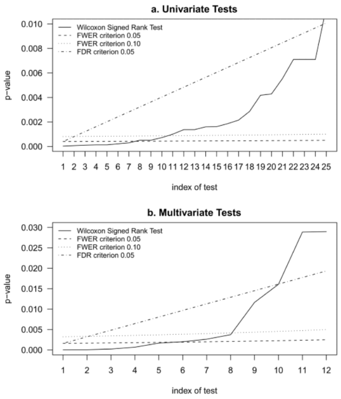

Figure 3.

Exact permutation *p*-values from (a) univariate and (b) multivariate Wilcoxon Signed Ranks tests. *P*-values below lines correspond to significant tests under the procedure indicated.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Biomarkers with significant univariate (X_1, X_2, X_3, or X_4) or multivariate (Q_j) tests in a confirmatory statistical analysis with strong control of the family wise error rate (FWER), or in an exploratory analysis controlling the false discovery rate (FDR).

	FWER control					FDR control				
biomarker[†]	X_1	X_2	X_3	X_4	Q_j	X_1	X_2	X_3	X_4	Q_j
Matrixmetalloproteinases										
MMP1			2		2		1	1		1
MMP3	2	1			2	1	1	1		1
MMP8								2		
MMP13					1	1	1	1		1
Adipokines										
Adiponectin						1	1			2
Complement-D	1	1			1	1	1	2		1
CRP				1					1	1
Resistin		1			2	1	1			1
Serpin-E1			1		1		1	1		1
Cytokines										
IL-1α								1		2
IL-1ra										2
IL-8					1	2	1	1		1
MIP-1β		2	1		1		1	1		1
TNFα							1	1		1

an entry of 1 denotes significance at an overall control rate of 0.05.

an entry of 2 denotes significance at an overall control rate of 0.10.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table A.1

Amount of imputed biomarker data across visits

	observed	imputed	missing	total
Day 0	485 (77.2%)	143 (22.8%)	0	628
Day 7	426 (67.8%)	170 (27.1%)	32 (5.1%)	628
Day 14	565 (90.0%)	31 (4.9%)	32 (5.1%)	628
Day 21	605 (96.3%)	22 (3.5%)	1	628
Day 35	521 (83.0%)	106 (16.9%)	1	628
Day 49	620 (98.7%)	8 (1.3%)	0	628
All time points	3222 (85.5%)	480 (12.7%)	66 (1.8%)	3768