Research Article

Biomarkers in Medicine

Using a multivariate model to assess the interactive effects of demographics and lifestyle on the hematological profile

Aim: To assess the extent to which a multivariate approach to modeling interrelated hematological indices provides more informative results than the traditional approach of modeling each index separately. **Materials & methods:** The effects of demographics and lifestyle on ten hematological indices collected from a Dutch population-based sample (n = 3278) were studied, jointly using multivariate distance matrix regression and separately using linear regression. **Results:** The multivariate approach highlighted the main effects of all predictors and several interactions; the traditional approach highlighted only main effects. **Conclusion:** The multivariate approach provides more power than traditional methods to detect effects on interrelated biomarkers, suggesting that its use in future research may help identify subgroups that benefit from different treatment or prevention measures.

First draft submitted: 12 October 2016; Accepted for publication: 20 March 2017; Published online: 23 June 2017

Keywords: age • BMI • hematological profile • MDMR • multivariate analysis • multivariate multiple regression • sex • smoking

Daniel B McArtor*[‡,1], Bochao D, Lin[‡2], Jouke-Jan Hottenga[2], Dorret I Boomsma[2], Gonneke Willemsen[2] & Gitta H Lubke[1,2]

[1]Department of Quantitative Psychology, University of Notre Dame, Notre Dame, IN, USA

[2]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

*Author for correspondence: Tel.: +1 410 829 2136 dmcartor@nd.edu.

[‡]Authors contributed equally

Hematological indices are complex, heritable [1–3] and tightly regulated human phenotypes [4]. The set of blood cells targeted in a standard laboratory blood test provides information on a wide range of functions, including immune response, hormone regulation, osmotic balance and coagulation regulation [5,6]. Abnormal values on hematological indices that fall outside the reference range may be indicative of underlying disease [7]. Because the standard hematological profile is relatively easy and inexpensive to obtain, it provides the basis for many commonly used tests in diagnoses.

Many hematological variables are related to demographics and lifestyles. For instance, red blood cell count, hematocrit and hemoglobin have been shown to be associated with age, sex, smoking and BMI [8]. Age and sex have also been found to be strongly related to platelet count, and age- and sex-specific reference ranges have even been proposed [9]. White blood cell count and platelet numbers are increased in obese participants [10], and in fact most hematological parameters show an association with BMI [11]. Smoking has also been associated with increases in white blood cell count, and changes in smoking behavior result in changes in the number of white blood cells [12].

Most studies concerning associations with hematological variables have taken an approach of investigating one hematological variable at a time. This approach is appropriate if a researcher is interested in the effects on a specific individual blood characteristic. If, however, the goal is to identify predictors associated with multiple blood characteristics, then the strategy of modeling each hematological variable in isolation is suboptimal. A more efficient strategy to accomplishing this goal is to test the association between a set of covariates and subjects' hematological profiles. Here, a hematological profile is defined as a set of scores on multiple observed

Future Medicine part of fsg

hematological variables. Analyzing blood characteristics jointly rather than individually is theoretically appealing because it facilitates the identification of predictors (and interactions between predictors) that influence multiple blood traits jointly. Furthermore, considering blood counts as a multivariate outcome is statistically beneficial because it allows researchers to conduct a single test assessing the effects of a set of covariates on all outcomes simultaneously rather than separate association tests for each blood count variable. The latter, more traditional, approach requires an adjustment to the standard criterion for statistical significance in order to correct for multiple testing [13]. No such correction is necessary when conducting a single multivariate test, thereby resulting in the potential for increased statistical power.

An important benefit of the multivariate approach is that it facilitates the identification of characteristic profiles for subgroups, for example, characteristic profiles of subjects with high versus normal BMI levels in males and females. Differences among these profiles can be highly informative and useful in the distillation of personalized treatments. For example, they can characterize risk for maladaptive levels of particular blood counts in subgroups of the population that may otherwise tend to appear normal on many other hematological indices. The first step in this process is to identify predictors that are relevant in explaining differences in the hematological profiles. Once established, profiles can be compared across subgroups.

In this study, we establish hematological profiles and investigate whether hematological profiles are associated with age, sex, BMI, smoking and any potential interactions between these covariates. This question is addressed using the standard univariate approach, as well as a multivariate approach that employs Multivariate Distance Matrix Regression (MDMR) [14,15] to test the association of the predictors with individual differences between the blood profiles as a whole. Beyond the specific application presented in this paper, the invocation of the multivariate approach and its comparison to the more traditional univariate approach is a critical focus of the current research. By illustrating the inferential advantages of the multivariate approach, the analyses presented here are partially intended to serve as a proof of concept motivating researchers to utilize multivariate regression techniques in future studies of interrelated biological indices.

Materials & methods

Participants

The participants in the study are registered with The Netherlands Twin Register (NTR) and took part in NTR biobank projects. The study design and settings have been discussed in detail in references [16–19]. In these projects, blood samples were collected during a home visit, and a brief interview was conducted to collect information on health status, lifestyle and body composition. All participants provided informed consent and the project was approved by The Medical Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam.

Hematological data were collected within a large biobank study that included twin families from the general population in The Netherlands. Individuals of 18 years or older from the NTR were invited into the study. Letters were sent to 14,093 participants. Of the 11,753 individuals who could be contacted by phone, 8126 (69%) agreed to participate, 193 (2%) had problems deciding and 3434 (29%) did not want to participate. A home visit could not be scheduled for 453 individuals willing to participate, so 7673 subjects were included following this procedure. Additionally, twin mothers and their family members who were recruited for a twinning project (n = 1059) were contacted, as well as young adults who were registered as children by their parents (n = 434). A number of spouses and family members of those contacted spontaneously indicated that they were also willing to enter the study (n = 364). For a second Biobank study [16,18], adult participants were invited by letter followed by a phone call. Of those participants reached by phone, 71% (517 individuals) agreed to take part in the study. In total, 9989 individuals participated in one or both biobank projects, and usable hematological data were available for 9672 of these individuals.

Among the group of individuals with hematological data (n = 9672), several exclusion criteria were applied. First, we excluded subjects with: blood C-reactive protein greater than or equal to 15 (mg/l), basophil count ≥ 0.3 (10^9/l), illness within 1 week of measurement, cancer treatment, use of anti-inflammatory medication and use of iron supplementation. The resulting sample was comprised of 8176 subjects. Participants may have had reported or unreported medical conditions beyond those enumerated above, but these conditions were not encoded or utilized in the analysis. Note that no patient groups were explicitly sampled in this population-based sample.

Next, subjects who had at least one blood cell score beyond ± 5 standard deviations from that variable's mean were excluded. The resulting sample was comprised of 7999 subjects from 3278 families. The final dataset (n = 3278) was formed by randomly sampling one member from each family to ensure independence of observations because MDMR cannot currently be adjusted to account for familial clustering.

Data collection

Blood sampling & hematological indices

Participants were visited at home between 7:00 and 10:00 am after overnight fasting. They were asked to refrain from strenuous exercise and, if possible, medication as of the evening before the visit, and smokers were instructed to refrain from smoking 1 h prior to the home visit. Prior to the visit, participants received a confirmation letter asking them to have all their medication available at the time of data collection. For all medicines, doses, brand names and chemical names were recorded by a nurse from the medication packaging. Fertile women who were not using oral contraceptives were visited on a fixed day of the cycle when possible. Women using oral contraceptives were visited in a pill-free week, and they were asked about the brand of the oral contraceptives that they were taking, and use of other kinds of hormonal contraceptives was noted. Participants were interviewed about their physical health [18].

During the home visit, peripheral venous blood samples were drawn by safety-lock butterfly needles into anticoagulant vacuum tubes in the following sequence: 2 × 9 ml EDTA, 2 × 9 ml lithium heparin (only one tube in a subset), 1 × 9 ml sodium heparin (in a subset only), 1 × 4.5 ml CTAD, 1 × 2.5 ml PAX (in a subset only), 1 × 4.5 ml serum and 1 × 2 ml EDTA tube. After collection, all tubes were inverted about ten-times to prevent clotting and then transported to the laboratory in Leiden.

Hematological parameters

The 2-ml EDTA tubes were transported at room temperature and upon arrival in the laboratory used to determine the hematological parameters using the Coulter system (Coulter Corporation, FL, USA). These parameters consisted of the total white blood cell count, percentage and absolute cell counts of five subtypes of white blood cells (neutrophils, lymphocytes, monocytes, eosinophils and basophiles), red blood cell count, hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, red blood cell distribution width, platelet count and mean platelet volume.

C-reactive protein

C-reactive protein (CRP) level was obtained from a plasma subsample that came from a 9 (ml) heparin tube that was transported in melting ice to the laboratory. The plasma subsample was snap-frozen and stored at -30°C. Upon defrosting one of these subsamples, CRP was determined by the 1000 CRP assay (Diagnostic Product Corporation, CA, USA).

Health, BMI & smoking

During the visit, a brief interview was conducted. Participants provided information on their current medication use and disease status and were asked about their smoking history. Height was reported and weight was measured. BMI (kg/m^2) was calculated from weight (kg) divided by the square of height (m^2). Based on their current self-report smoking behavior, participants were divided nonsmokers and current smokers. Current smokers were defined as those who reported to smoke regularly, while ex-smokers were categorized as nonsmokers independent of whether or not they smoked in the past.

Statistical analyses

Selecting outcome variables

To avoid the possibility of analyzing highly collinear variables that measure extremely similar traits, we excluded several hematological variables that displayed large (>0.70) correlations with other candidate outcome variables. This cutoff was selected partially to target theoretically redundant variables and partially based on the fact that only a few pair-wise correlations were above 0.70, whereas the rest were substantially lower. Specifically, we removed white blood cell count, red blood cell count, mean corpuscular hemoglobin and hematocrit ratio. In addition, basophil level was not included because variation in the basophil numbers was limited. This resulted in ten hematological outcome variables in total: neutrophil count (*#neut*), lymphocyte count (*#lymph*), monocyte count (*#mono*), eosinophil count (*#eos*), hemoglobin level (*hemo*), mean corpuscular volume (*MCV*), mean corpuscular hemoglobin concentration (*MCHC*), red blood cell distribution width in percent (*RDW%*), platelet count (*#plt*) and mean platelet volume (*MPV*). Figure 1 illustrates the correlations among these outcome variables.

Univariate association tests

Standard multiple regression with a univariate hematological outcome was used to investigate the effects of age, sex (62.6% female), smoking (77.6% current nonsmoker), BMI and their two-way interactions on ten hematological variables (see Table 1 for descriptive statistics). The main interest of the analyses was to investigate interactions in order to explore potential risk groups. There was no missing data on any of the four predictors, but some response profiles were partially missing. Each multiple regression model was fit using as much data as possible rather than excluding subjects from all models based on partial missingness. Missingness rates for each blood index are given in Table 1.

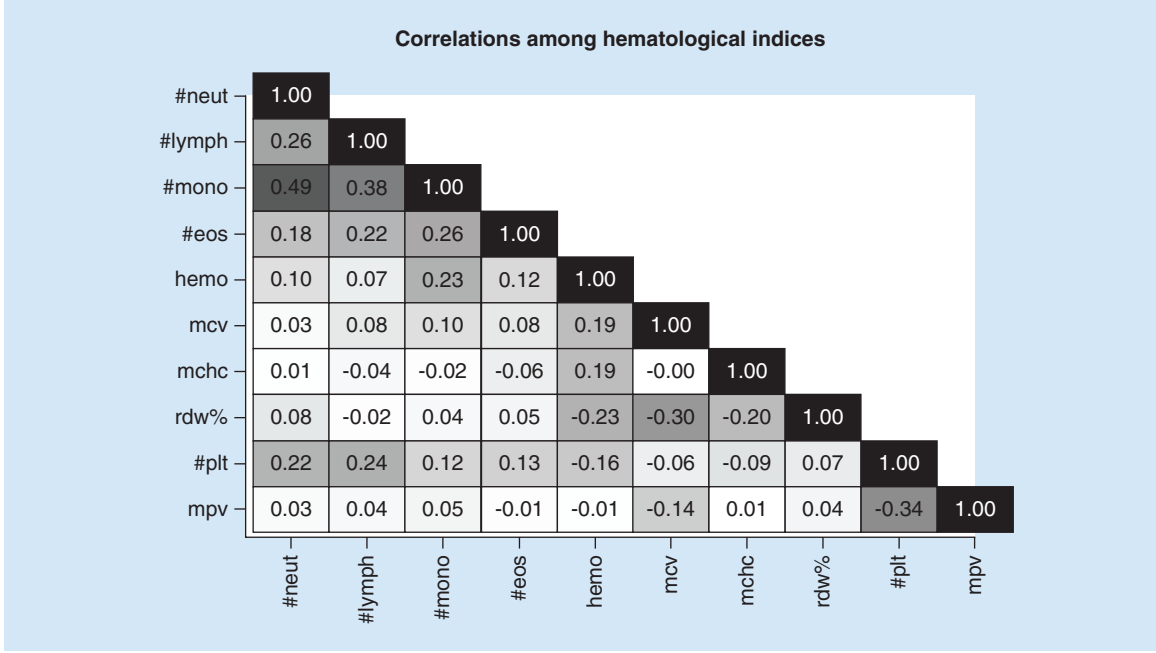

Figure 1. Correlations among the analyzed blood indices.

Because ten different models are to be fit using a single sample, each of which involves the estimation of many parameters, it is critical to control for multiple testing in these analyses to avoid inflated Type-I error rates. To do so, the statistical significance of each predictor on each outcome was evaluated using two different univariate significance criteria, each of which uses a Bonferroni correction. First, the criterion α_f^u = 0.05/10 = 0.005 controls the family-wise Type-I error rate at 0.05. That is, α_f^u sets the probability of committing a Type-I error on a particular predictor to 0.05, which is accomplished by dividing 0.05 by the number of fitted regression models. Second, α_{pc}^u = 0.05/(10 × 10) = 0.0005 controls the per-comparison Type-I error rate at 0.05. This stricter criterion controls the probability of committing any Type-I errors across all ten models that each use ten predictors.

Multivariate association tests

MDMR [13,14] is a procedure that permits testing the association of hematological profiles based on multiple blood cell indices with predictor variables. More specifically, differences between each pair of subjects' profiles are collected in symmetric $n \times n$ 'distance matrix'. Distance matrices are often subjected to cluster analysis [20], but MDMR utilizes them in a regression framework instead in order to test the effects of covariates on the profiles. This is done by partitioning the sums of squares of the distance matrix into a portion due to regression and a portion due to error. This decomposition is analogous to the partitioning of the sums of squares of a univariate outcome in standard linear regression.

Importantly, differences between profiles on multiple variables can be quantified using different measures of dissimilarity (i.e., distance). In this study, two different distance metrics were computed to characterize the dissimilarity between subjects' blood profiles, and the two resulting distance matrices were regressed onto the set of predictors using MDMR. The first metric considered was the Euclidean distance. If y_i and y_j denote vectors of scores along q outcome variables for subjects i and j, the Euclidean distance between these two subjects' response profiles is defined as,

(1)

$$d_e(i,j) = \sqrt{\sum_{k=1}^{q} (y_{ik} - y_{jk})^2}$$

It can be shown that Euclidean MDMR is the same model as multivariate multiple regression, so this approach also represents the natural multivariate extension to the standard linear regression used in the univariate analyses described above. Second, Manhattan distances were considered. The Manhattan distance between subjects i and j is defined as the sum of their absolute item-wise differences:

(2)

$$d_m(i,j) = \sum_{k=1}^{q} |y_{ik} - y_{jk}|$$

These distances are less sensitive to outliers and therefore more robust than Euclidean distances because they are based on absolute rather than squared differences. That is, the use of Euclidean distances (and standard linear regression) can result in spuriously significant effects due to outlying observations, but Manhattan distances are less prone to this phenomenon.

When conducting MDMR, one model is fit to all ten outcome variables jointly. This approach therefore requires a less stringent correction for multiple testing than the univariate approach. More specifically, the criterion $\alpha_f^m = 0.05$ controls the family-wise Type-I error rate (i.e., family-wise false discovery rate) of MDMR at 0.05 because only one model is fit to all outcome items jointly. Similarly, $\alpha_{pc}^m = 0.05/10 = 0.005$ controls the probability of committing a Type-I error on any of the ten predictors at 0.05 (i.e., per-comparison false discovery rate).

All analyses were conducted in *R* [21] using the *MDMR* package, which is freely available on The Comprehensive R Archive Network [22]. This package is the software companion to McArtor *et al.* [23], where the reader can also find a more detailed discussion of MDMR. Note that all 3278 individuals were utilized to fit the MDMR models despite some response profiles being partially missing because distance matrix computations can address partial missingness through a pseudo-imputation procedure. When computing the distance between two response profiles in cases where one or both are partially missing, the observed scores values are utilized and the resulting distance is then inflated proportionally to the number of missing observations.

Results

Univariate association tests

Table 2 reports the result of the univariate analyses in the form of p-values, and Table 3 gives the standardized regression coefficients and variance explained. One or more main effects were significant for all hematological variables. Results indicated that age, sex and high BMI are associated with elevated levels of most hematological variables, while smoking was found to be related to roughly half of outcomes. The effects of smoking and BMI tended to be positive in direction such that smokers with higher BMI tended to have higher blood counts on most outcomes, but the direction of age and sex effects differed across outcomes. The majority of the two-way interactions assessed with the univariate regression models were not marked as significant using either of the univariate significance criteria. Only hemoglobin was significantly predicted by multiple interaction effects (age:sex, age:smoker, sex:smoker); neutrophil count was found to be significantly associated with the interaction of age and sex, and the sex-by-smoking interaction was marked as a significant predictor of lymphocyte count at α_f^u. Figure 2 illustrates the nature of these five interaction effects on their corresponding univariate outcome. None of the other hematological variables were found to be significantly predicted by any interaction effects.

Table 1. Descriptive statistics for the numeric predictors and the hematological outcome variables.

Variable	Mean	SD	Min	First quantile	Median	Third quantile	Max	# Missing
Predictor variables								
Age	42.279	15.075	13.000	30.000	39.000	55.000	90.000	0
BMI	24.951	4.140	14.906	22.018	24.403	27.166	49.071	0
Hematological outcome variables								
#neut (10^9/l)	3.473	1.268	0.300	2.600	3.200	4.100	9.700	10
#lymp (10^9/l)	2.230	0.682	0.300	1.798	2.100	2.600	5.900	10
#mono (10^9/l)	0.534	0.171	0.000	0.400	0.500	0.600	1.400	10
#eos (10^9/l)	0.200	0.128	0.000	0.100	0.200	0.300	0.900	10
hemo (mmol/l)	8.798	0.769	6.100	8.300	8.700	9.400	11.100	1
MCV (fl)	91.536	4.534	69.300	88.800	91.600	94.400	113.500	2
MCHC (g/dl)	20.721	0.549	16.600	20.400	20.700	21.000	23.000	2
RDW (%)	12.364	0.743	10.700	11.900	12.200	12.700	16.600	360
#plt (10^9/l)	253.807	59.702	51.000	212.000	248.000	287.000	537.000	3
MPV (fl)	8.889	1.069	6.300	8.200	8.700	9.400	14.000	162

The first two rows (Age, BMI) correspond to the numeric predictors, and the remaining rows correspond to the hematological outcome variables.

Table 2. p-values from the linear regression models predicting individual hematological indices

Effect	*#neut*	*#lymp*	*#mono*	*#eos*	*hemo*	*MCV*	*MCHC*	*RDW%*	*#plt*	*MPV*
Full model	**<1e-16**	<1e-16	<1e-16	<1e-16	<1e-16	<1e-16	7.10e-13	<1e-16	<1e-16	4.10e-15
Main effects										
Age	**3.40e-05**	**1.20e-10**	0.014	*0.003*	0.91	**5.20e-62**	**8.20e-11**	**1.80e-19**	**3.10e-05**	**7.70e-14**
Sex	**2.80e-05**	**2.20e-09**	**5.60e-24**	**9.70e-08**	**<1e-70**	0.10	*0.0015*	0.0077	**1.80e-41**	0.025
BMI	**4.00e-31**	**1.80e-13**	**4.00e-08**	*0.00059*	**1.10e-11**	**2.10e-06**	0.55	*0.0017*	**9.30e-11**	0.90
Smoker	**3.00e-66**	**2.50e-58**	**4.90e-43**	**2.50e-14**	**1.30e-18**	**1.80e-45**	0.016	0.20	0.039	0.89
Interaction effects										
Age:sex	**1.50e-09**	0.41	0.92	0.16	**9.80e-15**	0.15	0.13	0.01	0.49	0.11
Age:BMI	0.79	0.054	0.73	0.15	0.85	0.039	0.011	0.41	0.034	0.018
Age:smoker	0.30	0.96	0.21	0.45	**1.90e-05**	0.087	0.053	0.036	0.57	0.0093
Sex:BMI	0.032	0.14	0.37	0.67	0.028	0.34	0.48	0.25	0.0087	0.012
Sex:smoker	0.77	*0.0036*	0.48	0.96	**1.90e-05**	0.16	0.13	0.23	0.94	0.38
BMI:smoker	0.54	0.84	0.41	0.0094	0.36	0.51	0.58	0.61	0.39	0.51

Each column corresponds to a model fit to one of the outcome variables. The first row corresponds to the p-value of the model as a whole, rows 2–5 correspond to the p-value of a main effect and rows 6–11 report the p-values of each interaction effect. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e., $\alpha_f^u = 0.05/10$) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e., $\alpha_{pc}^u = 0.05/100$) are emphasized with bold font.

Multivariate association tests

Both Euclidean and Manhattan MDMR resulted in extremely small p-values for all four main effects. However, Euclidean MDMR also detected three interactions with age (age:sex, age:BMI, age:smoker) that were significant at α_{pc}^m, and two more that were significant at α_f^m (sex:BMI, sex:smoker). Manhattan MDMR found the same three highly significant interaction effects involving age, and one more that was significant at α_f^m (sex:BMI). BMI and smoking were the only two predictors that did not combine to yield a statistically significant interaction effect from at least one of the MDMR models. See Table 4 for all MDMR p-values.

Significant effects imply substantial differences between participants in their hematological profiles based on the ten observed blood variables. To visualize these interacting effects across different strata, we conducted a median split on each of the two predictors comprising a significant interaction (with the resulting groups labeled young/old, thin/heavy, male/female and nonsmoker/smoker) and plotted the average blood profiles in each of the resulting four groups (high/low × high/low on each pair of predictors). Figure 3 displays these 'prototypical' or 'average' hematological profiles for each resulting group. Note that the decision to bin the continuous covariates according to a median-split was largely arbitrary; the 'average profile' in each group involving age or BMI depends on the cutoffs used to define the groups.

The five subplots comprising Figure 3 elucidate the five significant two-way interaction effects and the main effects of each predictor by illustrating how differences in the predictor variables relate to differences in the blood profiles. For example, the top-left subplot illustrates the effects of age, sex and their interaction on the multivariate outcome. There are clear differences in hematological profiles among younger females, older females, younger males and older males, but the differences between groups are not constant among the ten indices defining the profile. That is, it is not the case that one group tends to score uniformly higher or lower than another on all ten outcomes. The multivariate effects have complex patterns that allow for potentially different effects on each variable comprising the outcome. For example, both age and sex seem to have comparatively small effects on mean platelet volume (all four groups tend to score similarly), age seems to have a main effect on mean corpuscular hemoglobin concentration (young people tend to score higher regardless of sex), sex seems to have a strong main effect on hemoglobin level (males tend to score higher regardless of age) and the interaction between these two predictors is important in predicting eosinophil count (younger females tend to score lower than the other three groups). The differential effects of age and sex on the remaining six hematological variables are also illustrated in the top-left subplot of Figure 3, and the other four subplots illustrate the differential effects of the other pairs of predictors that comprise a significant interaction effect.

Discussion

In line with previous research, our univariate analyses confirmed that age, sex, BMI and smoking are related to individual hematological parameters. This set of analyses, however, did not provide strong evidence for interactive effects of these predictors. On the other hand, focusing on differences among subjects' hematological profiles rather than differences in individual hematological indices was shown to yield sufficient power to detect interactions among predictors in the model. By using information from all of the outcomes jointly, the multivariate approach can detect smaller, but still meaningful, effects more efficiently than the traditional univariate approach, which can only consider each effect in isolation. The 'average profiles' illustrated in Figure 3 highlight differences between typical hematological profiles among patient groups that might be deemed indistinguishable by more traditional univariate methodology.

For example, the interaction of age and BMI was not marked as significant in any of the univariate analyses, but this interaction was found to be significantly related to the hematological profiles as a whole. This phenomenon can be understood by examining the upper-rightmost plot of Figure 3, which illustrates the effects of age and BMI on the hematological profiles. While no single outcome variable is characterized by a large interaction effect, the top-right panel of Figure 3 illustrates the finding that the interaction of age and BMI has a modest effect on many of the blood variables. These small interaction effects are most visible when inspecting eosinophil count, hemoglobin level and red blood cell distribution. Participants who were both younger and thinner than average were found to have lower eosinophil counts than the rest of the sample (the black circle is lower than all other points). Age was not predictive of hemoglobin levels for patients with above-average BMI, but among the lower-BMI subgroup, older participants had lower hemoglobin levels than younger participants (squares overlap, but the gray dot is lower than the black dot). Similarly, BMI was not predictive of red blood cell distribution among the younger participants, but it increased as a function of BMI among the older participants (black points overlap, but the gray square is higher than the gray circle).

Beyond facilitating higher statistical power than the standard univariate approach, using the multivariate approach to identify 'prototypical profiles', such as those illustrated in Figure 3, may be useful for clinicians in the future. These profiles could be used to formulate expectations about patient groups in a more fine-grained manner than could be achieved based on analyses of individual outcome variables.

Furthermore, the benefits of this multivariate approach invites future research on personalized treatment that directly utilizes multivariate association tests. Jointly modeling several outcomes facilitates the simultaneous study of multiple biological responses to a treatment. The multivariate approach

Table 3. R^2 and standardized regression coefficients from the linear regression models predicting individual hematological indices.

Effect	*#neut*	*#lymph*	*#mono*	*#eos*	*hemo*	*MCV*	*MCHC*	*RDW%*	*#plt*	*MPV*
R^2	**0.133**	**0.102**	**0.105**	**0.044**	**0.467**	**0.138**	**0.024**	**0.048**	**0.07**	**0.029**
Main effects										
Age	**-0.073**	**-0.115**	0.044	*0.055*	0.002	**0.296**	**-0.121**	**0.183**	**-0.076**	**-0.147**
Sex	**0.069**	**0.100**	**-0.17**	**-0.092**	**-0.645**	-0.027	*-0.055*	0.05	**0.233**	0.04
BMI	**0.208**	**0.134**	**0.099**	*0.064*	**0.095**	**-0.084**	0.011	*0.062*	**0.119**	-0.002
Smoker	**0.29**	**0.276**	**0.234**	**0.133**	**0.114**	**0.236**	0.042	0.023	0.035	0.002
Interaction effects										
Age:sex	**-0.105**	0.014	0.002	-0.026	**0.105**	-0.025	0.028	-0.051	-0.012	-0.031
Age:BMI	0.004	0.034	0.006	0.026	0.003	-0.035	0.046	0.017	-0.038	0.047
Age:smoker	0.019	-0.001	0.023	-0.014	**0.060**	0.031	0.037	0.043	-0.01	0.052
Sex:BMI	0.040	-0.028	-0.017	0.008	-0.032	-0.018	-0.014	0.023	0.05	-0.05
Sex:smoker	0.005	*0.048*	-0.012	0.001	**0.054**	0.022	0.026	-0.021	0.001	0.015
BMI:smoker	-0.011	-0.004	-0.015	-0.048	-0.013	-0.012	-0.011	-0.01	0.016	-0.013

R^2 is shown in the first row, standardized regression coefficients of each main effect in rows 2–5, and all two-way interaction effects in rows 6–11. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e., α_r^u = 0.05/10) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e., α_{pc}^u = 0.05/100) are emphasized with bold font.

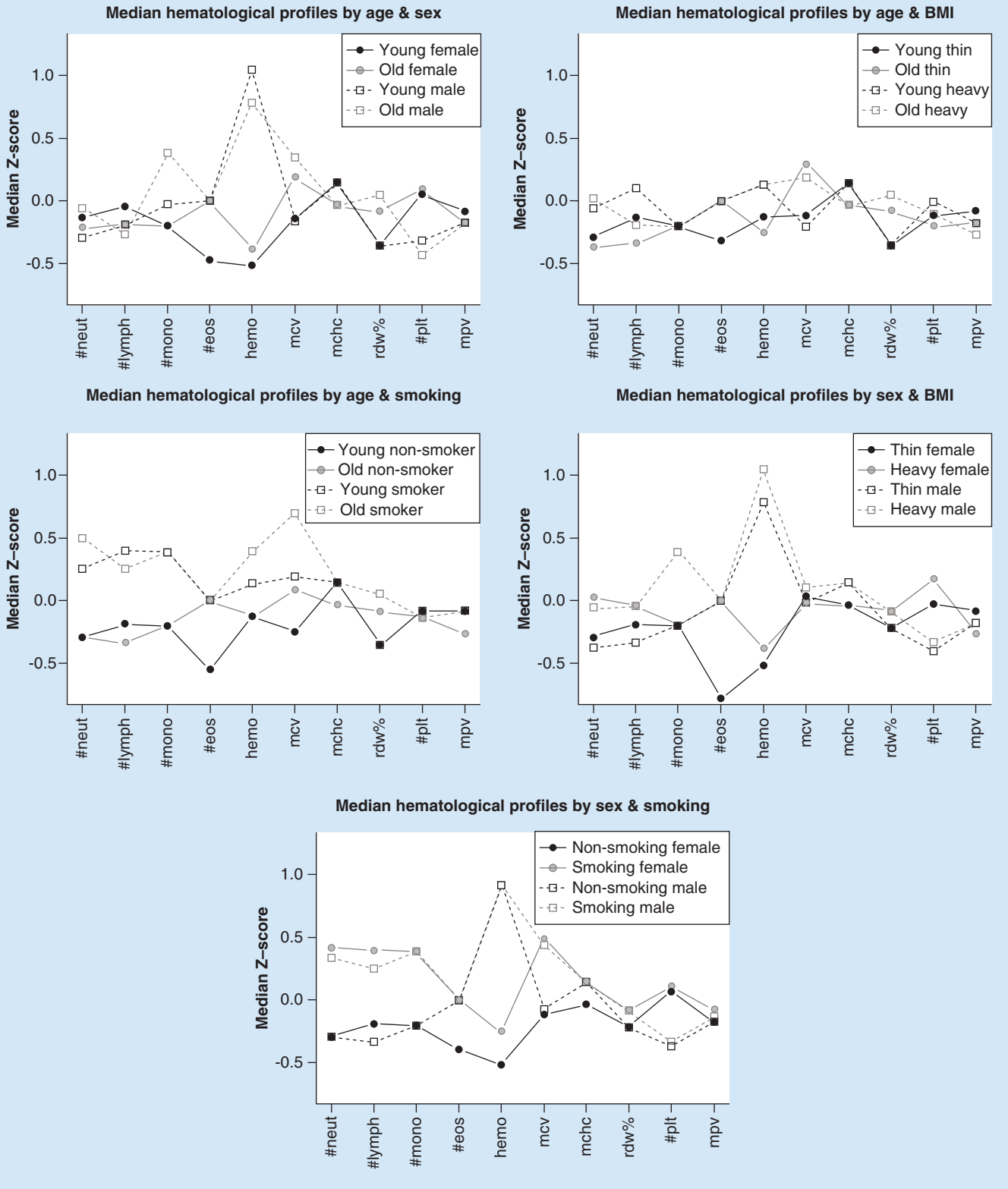

Figure 2. Illustration of significant interaction effects found in the univariate linear regression models. Each panel illustrates predicted values (vertical axis) across two variables (horizontal axis and line type) that interacted to predict a hematological variable. All predictors not involved in an illustrated interaction were kept fixed at their sample means.

can, therefore, be used to test the effectiveness of a treatment on several target variables while also considering potential treatment interactions with demographic variables. For example, the multivariate approach could be used to model phenotypes that are known to be impacted as a side effect of a treatment in conjunction with the variable(s) that are targeted by the treatment. This allows the identification of subgroups of individuals within a population who respond well to the treatment while also uncovering subgroups who are particularly susceptible to its side effects.

Importantly, the multivariate approach may also be useful in the context of genetic association studies. The effects of individual genetic variants on complex human traits are usually small [24]. Genome wide association studies for hematological parameters have now implicated several loci in the regulation of hematological indices, but the power is currently insufficient to detect all loci involved [25–28]. To attain sufficient power to detect these effects, consortia currently focus on increasing sample sizes. However, an alternative approach to increasing power involves improving the way that the phenotypes are operationalized and analyzed. When a researcher has multiple variables that measure a trait of interest, the multivariate approach can be used to test their joint association with individual genetic variants. The results presented here suggest that this approach could lead to increased power relative to analyzing each variable on its own, and this approach can also yield higher power than analyzing an aggregate-score computed from all of the variables measuring the trait [29]. MDMR facilitates the inclusion of an arbitrary number of outcome variables. It can even be used when there are more outcomes than observations, so these benefits can still be capitalized upon when the outcome is extremely high-dimensional.

In addition to the statistical strengths of our study discussed in the precedent paragraphs, some limitations of the current design should be mentioned. Our results indicated that age is an important moderator of the effects of sex, BMI and smoking behavior on differences in hematological profiles and, given the apparent impact of age on these other predictors, this cross-sectional study should be followed-up using a longitudinal design. Furthermore, our study was based on a large population-based sample, which was not selected on the basis of disease or other characteristics related to health. This design facilitates learning about the population at large and represents a proof of principle supporting the use of a multivariate approach for modeling biological profiles, but this approach needs to be utilized on samples including clinical groups to determine its usefulness in a clinical setting. Further research focusing on clinical populations is necessary to quantify the extent to which the multivariate approach can facilitate more clinically actionable insights than more traditional analysis techniques.

Table 4. p-values from the MDMR models fit to all hematological indices jointly

Effect	*Euclidean*	*Manhattan*
Full model	**<1e-16**	**<1e-16**
Main effects		
Age	**<1e-16**	**<1e-16**
Sex	**<1e-16**	**<1e-16**
BMI	**<1e-16**	**<1e-16**
Smoker	**<1e-16**	**<1e-16**
Interaction effects		
Age:sex	**4.3e-10**	**4.4e-09**
Age:BMI	**0.00034**	**0.0044**
Age:smoker	**0.0022**	**0.00034**
Sex:BMI	*0.0099*	*0.0094*
Sex:smoker	*0.0075*	0.33
BMI:smoker	0.33	0.062

Rows correspond to predictors, columns correspond to metrics used to define the dissimilarity between pairs of hematological profiles. Values smaller than the Bonferroni-adjusted significance criterion to ensure that each predictor has a Type-I error rate of 0.05 (i.e., $\alpha_{Pr}^{m} = 0.05$) are emphasized with italic font. Values smaller than the Bonferroni-adjusted significance criterion to ensure that the probability of any Type-I error is 0.05 (i.e., $\alpha_{Pc}^{m} = 0.05/10$) are emphasized with bold font. The p-value corresponding to the joint effect of all predictors is found in the first row, the main effects of each predictor in the next four rows, and all two-way interaction effects in the final six rows.

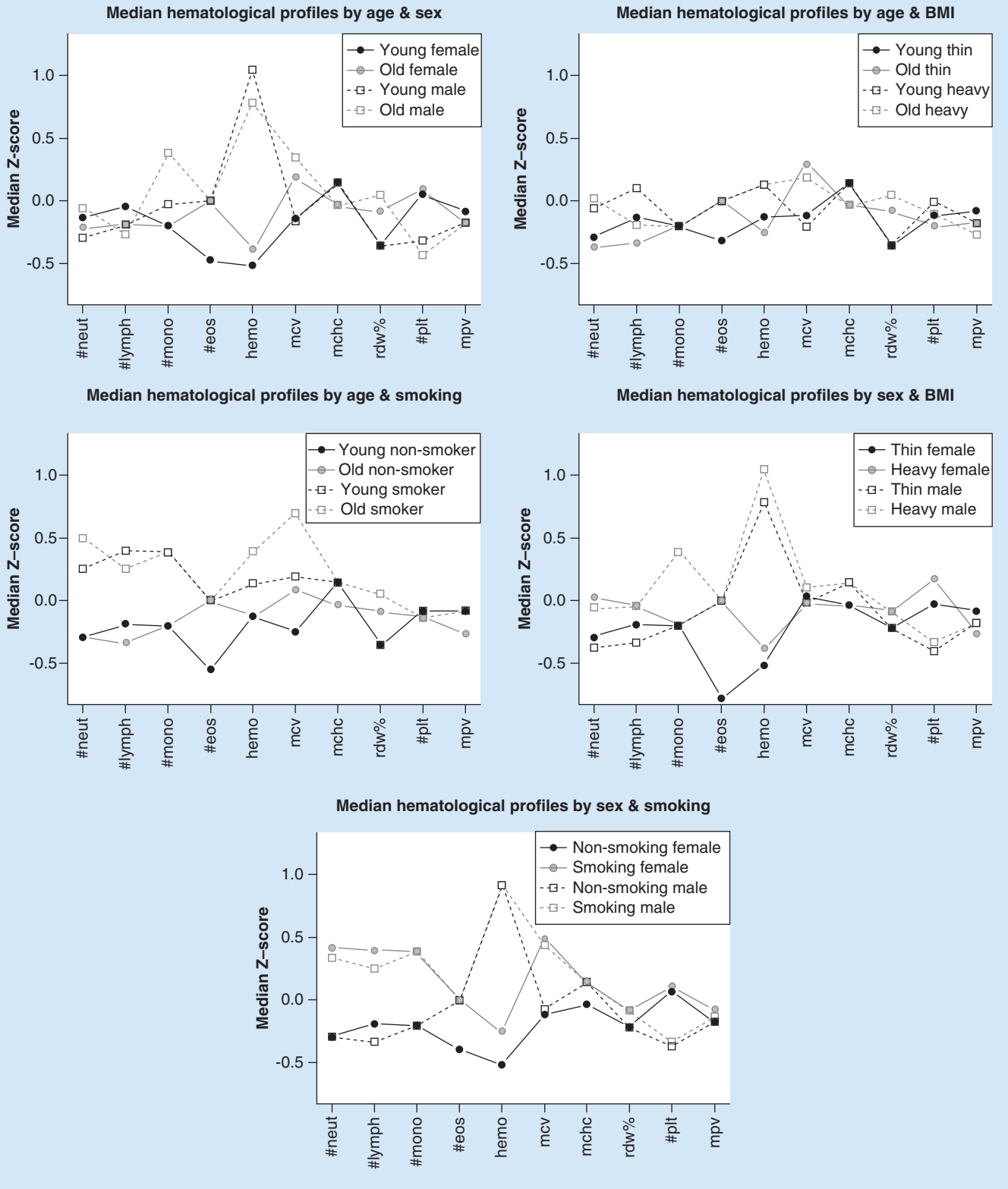

Figure 3. Median standardized scores (vertical axes) on each hematological outcome variable (horizontal axes) for five sets of subgroups (each plot) to illustrate the effects of the five interactions identified as significantly associated with subjects' hematological profiles. Each interaction is illustrated by plotting the 'average hematological profile' of four subgroups that characterize the two-way interaction.

Figure 3. Median standardized scores (vertical axes) on each hematological outcome variable (horizontal axes) for five sets of subgroups (each plot) to illustrate the effects of the five interactions identified as significantly associated with subjects' hematological profiles (cont. from facing page). These groups are defined by **(A)** a median-split on age and by sex, **(B)** a median split on age and on BMI, **(C)** a median-split on age and by smoking, **(D)** sex and a median split on BMI and **(E)** sex and smoking. The average profiles of each subgroup within each plot are differentiated by color, point type and line type, as indicated in each figure legend. Connecting lines were added to allow for an easier visual comparison of the groups' profiles. These visualizations illustrate the differential covariate effects on the hematological profiles as a whole. For example, the subplot concerning the effects of age and sex illustrates the comparatively small effects of both predictors on mean platelet volume (all four groups tend to score similarly), the main effect of age on mean corpuscular hemoglobin concentration (young tending to score higher regardless of sex), the main effect of sex on on hemoglobin level (males tend to score higher regardless of age), and the effect of the interaction between these two predictors on eosinophil count (younger females tend to score lower than the other three groups).

In the analyses presented here, both Euclidean and Manhattan MDMR marked the interactions of age with sex, BMI and smoking, as well as the interaction of sex and BMI, as significantly related to the hematological profiles. The use of Euclidean and Manhattan distances, however, yielded inconsistent results with respect to the interaction of sex and smoking. The use of Euclidean distances to define the dissimilarity between pairs of response profiles resulted in a significant sex by smoking interaction, but the use of Manhattan distances did not. Manhattan distances are less sensitive to outlying observations and are, therefore, preferable if analyses are conducted in small samples in order to avoid potentially spurious results. This robustness, however, comes at the expense of potentially suboptimal power to detect genuine effects in larger samples. In future studies, researchers should therefore consider their sample size in addition to the relative cost of false positives and false negatives when choosing between Euclidean and Manhattan distances.

In conclusion, a multivariate approach to hematological analysis increases the power to detect important interactions within predictors relative to standard univariate analyses. In the future, multivariate methods, including MDMR, have the potential to help identify subgroups of patients who benefit from different treatment or prevention measures.

Acknowledgements

The authors thank the individuals who participated in The Netherlands Twin Register biobank projects.

Author contribution

DB McArtor and BD Lin carried out the quality control of the data, performed the statistical analyses and wrote the paper and were supervised by JJ Hottenga and GH Lubke. DI Boomsma and G Willemsen are responsible for designing and setting up the study. All authors contributed to the writing.

Financial & competing interests disclosure

This work was supported by: Genotype/phenotype database for genetic studies (ZonMW Middelgroot [911-09-032]); Database Twin register (NWO 575-25-006); Twin family database for behavior genetics and genomics studies (NWO 480-04-004); Genome-wide analyses of European twin and population cohorts (EU/QLRT-2001-01254); a collaborative study of the genetics of DZ twinning (NIH R01 HD042157-01A1); EMGO+ Institute for Health and Care Research, Neuroscience Campus Amsterdam, Center for Medical Systems Biology (CMSB), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL) 184.021.007; GENOMEUTWIN/EU (QLG2- CT-2002-01254); NIH (NIH-HEALTHF4-2007-201413); European Research Council (230374-GMI). B Lin received a PhD grant (201206180099) from the China Scholarship Council. G Lubke is supported by NIDA R37 DA-018673. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Summary points

- Multivariate distance matrix regression resulted in higher power to detect effects on the hematological profiles than did the use of separate linear regression models.
- This increased power can be partially attributed to: the ability to leverage the shared information among the multiple hematological indices in a single test, and a less stringent correction for multiple testing.
- When studying the indices in isolation, neutrophil count and hemoglobin level were the only two indices found to be affected by interactions among the predictors, but the multivariate approach provided stronger evidence for interaction effects on the hematological profiles as a whole.
- The additional information provided by jointly modeling interrelated biomarkers with a multivariate model can provide more fine-grained results for clinicians due to increases in power.
- The multivariate approach may also prove clinically useful by virtue of its ability to provide more detailed and personalized predictions for biomarkers of different subpopulations.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Evans DM, Frazer IH, Martin NG. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res.* 2, 250–257 (1999).

2 Garner C, Tatu T, Reittie JE *et al.* Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* 95, 342–346 (2000).

3 Hall MA, Ahmadi KR, Norman P *et al.* Genetic influence on peripheral blood T lymphocyte levels. *Genes Immun.* 1, 423–427 (2000).

4 Okada Y, Kamatani Y. Common genetic factors for hematological traits in humans. *J. Hum. Genet.* 57, 161–169 (2012).

5 Parichehreh V, Estrada R, Kumar SS *et al.* Exploiting osmosis for blood cell sorting. *Biomed. Microdevices.* 13, 453–462 (2011).

6 Karpman D, Ståhl AL, Arvidsson I *et al.* Complement interactions with blood cells, endothelial cells andmicrovesicles in thrombotic and inflammatory conditions. *Adv. Exp. Med. Biol.* 865, 19–42 (2015).

7 Ambayya A, Su AT, Osman NH *et al.* Haematological reference intervals in a multiethnic population. *PLoS ONE* 9, e91968 (2014).

8 Barazzoni R, Cappellari GG, Semolic A *et al.* The association between hematological parameters and insulin resistance is modified by body mass index – results from the north-east Italy MoMa Population Study. *PLoS ONE* 9, e101590 (2014).

9 Biino G, Santimone I, Minelli C *et al.* Age- and sex-related variations in platelet count in Italy: a proposal of reference ranges based on 40987 subjects' data. *PLoS ONE* 8, e54289 (2013).

10 Farhangi MA, Keshavarz AS, Eshraghian M *et al.* White blood cell count in women: relation to inflammatory biomarkers, haematological profiles, visceral adiposity, and other cardiovascular risk factors. *J. Health Popul. Nutr.* 31, 58–64 (2013).

11 Vuong J, Qiu YL, La M *et al.* Reference intervals of complete blood count constituents are highly correlated to waist circumference: should obese patients have their own "normal values?" *Am. J. Hematol.* 89, 671–677 (2014).

12 Sunyer J, Munoz A, Peng Y *et al.* Longitudinal relation between smoking and white blood cells. *Am. J. Epidemiol.* 144, 734–741 (1996).

13 Bland J, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 310, 170 (1995).

14 Anderson MJ. A new method for non-parametric multivariateanalysis of variance. *Austral. Ecol.* 26, 32–46 (2001).

15 McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancyanalysis. *Ecology* 82, 290–297 (2001).

16 Sirota M, Willemsen G, Sundar P *et al.* Effect of genome and environment on metabolic and inflammatory profiles. *PLoS ONE* 10, e0120898 (2015).

• **An alternative examination of the complex interaction between hematological indices.**

17 Willemsen G, deGeus EJ, Bartels M *et al.* The Netherlands Twin Register Biobank: a resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* 13, 231–245 (2010).

18 Willemsen G, Vink JM, Abdellaoui A *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res. Hum. Genet.* 16, 271–281 (2013).

19 Boomsma DI, Willemsen G, Sullivan PF *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA Biobank Projects. *Eur.J. Hum. Genet.* 16, 335–342 (2008).

20 Wang X, Duren Z, Zhang C, Chen L, Wang Y. Clinical dataanalysis reveals threesubytpes of gastric cancer. In: *Proceedings of the 6th International IEEE Conference on Systems Biology (ISB 2012),* 316–321 (2012).

21 R Core Team. R: a language and environment for statistical computing. Vienna, Austria (2014).

22 The Comprehensive R Archive Network. https://cran.r-project.org/

23 McArtor DB, Lubke GH, Bergeman CS. Extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic. *Psychometrika* doi:10.1007/s11336-016-9527-8 (2016) (Epub ahead of print).

•• **Discusses multivariate distance matrix regression in much more mathematical and theoretical detail.**

24 Lee JJ, Vattikuti S, Chow CC. Uncovering the genetic architectures of quantitative traits. *Comput. Struct. Biotechnol. J.* 14, 28–34 (2015).

25 Ferreira MA, Hottenga JJ, Warrington NM *et al.* Sequence variants in three lociinfluencemonocyte counts and erythrocyte volume. *Am. J. Hum. Genet.* 85, 745–749 (2009).

26 van der Harst P, Zhang WH, Leach IM *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* 492, 369–375 (2012).

• **An illustration of the progress being made in understanding the genetic causes of variation in blood cell counts.**

27 Keller MF, Reiner AP, Okada Y *et al.* Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum. Mol. Genet.* 23, 6944–6960 (2014).

28 Soranzo N, Spector TD, Mangino M *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in theHaemGen consortium. *Nat. Genet.* 41, 1182–1190 (2009).

• **An overview of gene finding efforts concerning the total of hematological parameters.**

29 Lubke GH, McArtor DB. Multivariate genetic analyses in heterogeneous populations. *Behav. Genet.* 44, 232–239 (2014).

•• **A simulation study illustrating the benefits of multivariate distance matrix regression relative to other approaches to analyzing multivariate phenotypes.**