

# Statistical approaches applicable in managing OMICS data: Urinary proteomics as exemplary case

De-Wei An<sup>1,2</sup>  | Yu-Ling Yu<sup>1,2</sup>  | Dries S. Martens<sup>3</sup>  |  
 Agnieszka Latosinska<sup>4</sup>  | Zhen-Yu Zhang<sup>5</sup>  | Harald Mischak<sup>4</sup>  |  
 Tim S. Nawrot<sup>2,3</sup>  | Jan A. Staessen<sup>1,6</sup>  

<sup>1</sup>Non-Profit Research Association Alliance for the Promotion of Preventive Medicine, Mechelen, Belgium

<sup>2</sup>Research Unit Environment and Health, KU Leuven Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium

<sup>3</sup>Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium

<sup>4</sup>Mosaiques Diagnostics GmbH, Hannover, Germany

<sup>5</sup>Research Unit Hypertension and Cardiovascular Epidemiology, KU Leuven Department of Cardiovascular Sciences, University of Leuven, Leuven, Belgium

<sup>6</sup>Biomedical Research Group, Faculty of Medicine, University of Leuven, Leuven, Belgium

## Correspondence

Jan A. Staessen, Non-Profit Research Association Alliance for the Promotion of Preventive Medicine, Leopoldstraat 59, BE-2800 Mechelen, Belgium.

Email: [jan.staessen@appremed.org](mailto:jan.staessen@appremed.org)

## Abstract

With urinary proteomics profiling (UPP) as exemplary omics technology, this review describes a workflow for the analysis of omics data in large study populations. The proposed workflow includes: (i) planning omics studies and sample size considerations; (ii) preparing the data for analysis; (iii) preprocessing the UPP data; (iv) the basic statistical steps required for data curation; (v) the selection of covariables; (vi) relating continuously distributed or categorical outcomes to a series of single markers (e.g., sequenced urinary peptide fragments identifying the parental proteins); (vii) showing the added diagnostic or prognostic value of the UPP markers over and beyond classical risk factors, and (viii) pathway analysis to identify targets for personalized intervention in disease prevention or treatment. Additionally, two short sections respectively address multiomics studies and machine learning. In conclusion, the analysis of adverse health outcomes in relation to omics biomarkers rests on the same statistical principle as any other data collected in large population or patient cohorts. The large number of biomarkers, which have to be considered simultaneously requires planning ahead how the study database will be structured and curated, imported in statistical software packages, analysis results will be triaged for clinical relevance, and presented.

## KEYWORDS

multidimensional classifiers, proteomics, statistical methods, urinary proteomics

## 1 | INTRODUCTION

Epidemiology is the science that studies the patterns, causes, and effects of health and disease in populations or patients. The fundamental, ethical, and scientific

principles that traditionally inspired epidemiology was to acquire new scientific information that can be used to maintain, enhance and promote people's health. Over the past 20 years, scientists with a wide range of research interests embraced the wave of rapidly evolving novel

**Abbreviations:** AUC, area under the ROC curve; CE-MS, capillary electrophoresis coupled with mass spectrometry; IDI, integrated discrimination improvement; MFA, multifactor analysis; ML, machine learning; NRI, net reclassification improvement; PCA, principal component analysis; PLS-A, partial least squares analysis; PLS-DA, partial least squares discriminant analysis.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Mass Spectrometry Reviews* published by John Wiley & Sons Ltd.

technologies. They are increasingly making use of the opportunities offered by high-throughput omics (genetics, epigenetics, transcriptomics, proteomics, metabolomics, etc.), deep DNA sequencing, greater access to public large databases (genes, proteins, etc.), pooled resources of longitudinal observations in a variety of populations and patients, high-speed information exchange via the internet, interaction with basic scientists in multidisciplinary teams, and natural experiments (e.g., Mendelian randomization).

This review is primarily targeting an audience consisting of physicians, trialists, and scientists of all walks of clinical science with a good working knowledge of applied statistics, but without professional statistical expertise. These researchers are often overwhelmed by the complexity of including omics data in the design, execution, and analyses of their studies. This review describes how omics data can be analyzed, using the urinary proteomic profiling (UPP) as a working example, given the experience of the authors in this particular field and the relatively simple statistical concepts allowing its evaluation. In addition, a PMC search was conducted with as search terms in the title or abstract of publications “*statistical methods*” AND omics. This search generated a list of 140 articles published from 2007 until 2023. After browsing the abstracts, 30 relevant papers were retained and read in detail. Additionally, seven seminal articles were identified from the reference lists of the initial 30 articles. Thus, moving beyond the UPP, the literature search allowed adding references throughout this review from omics fields other than UPP. Given the emerging approaches described in the recent literature, two short sections respectively dealing with the analysis of multiomics data and artificial intelligence/machine learning (ML), are broadening the scope of this review. The analysis of multiomics data fits in the workflow described in the article from preparing preprocessing of the omics data up to the evaluation of the added diagnostic or prognostic accuracy and identification of molecular mechanisms (Tables 1 and 2) and addresses the issue how to handle highly correlated data. In contrast, ML is a completely different approach, however, with a similar finality as the traditional statistical methods, that is, the identification and validation of biomarkers as guide to a personalized approach to prevention and treatment of disease.

## 2 | URINARY PROTEOMICS

In contrast to tissue and blood samples, urine can be easily, repeatedly, and noninvasively collected. A 10-mL midmorning urine sample is all what is needed. The UPP

contains over 20,000 peptide fragments, which can be detected by capillary electrophoresis coupled with mass spectrometry (CE-MS). Each peptide signal is characterized by its migration time and its spectrometric mass (Figure 1). The CE-MS procedure involves normalization of the migration time and peak spectrometric intensity by means of internal polypeptide standards ran along with the samples. These peptides result from normal biological processes and are unaffected by any disease studied so far (Good et al., 2010; Mavrogeorgis et al., 2021; Mischak & Schanstra, 2011). By sequencing the urinary peptide fragments, currently over 5000 parental proteins were identified, which reveal the pathogenic molecular processes, explaining why UPP analyses often focus on peptides with known amino-acid chain (Figure 1).

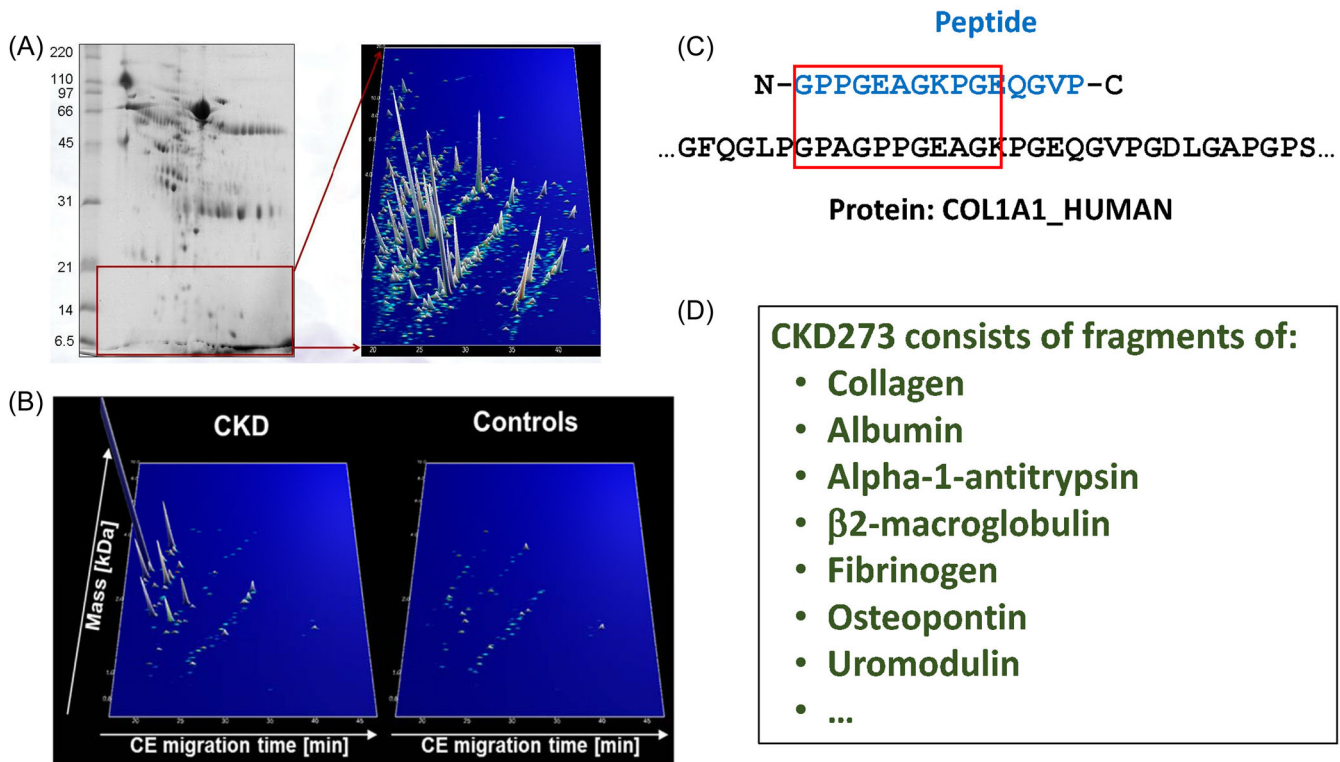
To address the heterogeneity and complexity of diseases and to increase the diagnostic and prognostic information, single urinary peptides can be combined into multidimensional disease-specific classifiers, such as HF1 for a symptomatic left ventricular diastolic dysfunction (Zhang, Ravassa, et al., 2017; Zhang et al., 2019) or CKD273 for imminent decline of glomerular filtration to below 60 mL/min/1.73 m<sup>2</sup> (Tofte et al., 2020). Compared with single biomarkers, such as for instance circulating atrial peptides in the case of left ventricular dysfunction (Zhang, Ravassa, et al., 2017; Zhang et al., 2019) or biomarkers associated with chronic kidney disease (Wasung et al., 2015), multidimensional UPP classifiers not only increase precision but also have less inherent variability.

## 3 | STATISTICAL WORKFLOW

The workflow to analyze the UPP described in the following sections (Table 2) has been applied in multiple peer-reviewed research articles published in the literature dealing with adverse cardiovascular (Htun et al., 2017; Zhang et al., 2015) or noncardiovascular (Staessen et al., 2022; Yang et al., 2022) health outcomes, left ventricular dysfunction (He, Melgarejo, et al., 2021; Zhang et al., 2016, 2019; Zhang, Ravassa, et al., 2017) or chronic kidney disease (Pontillo et al., 2017; Tofte et al., 2020). However, based on the PMC review, references from omics fields other than UPP provide accessible links to further reading.

### 3.1 | Preparing for data collection

Previous publications described in detail the scientific requirements for setting up and reporting on proteomic biomarker data (Latosinska et al., 2019). Sample size



**FIGURE 1** Urinary proteomics by application of capillary electrophoresis coupled with mass spectrometry. Low molecular weight peptides are first separated by capillary electrophoresis (A). Normalized molecular mass (y-axis) is plotted against normalized capillary electrophoresis migration time (x-axis) in a three-dimensional graph representing 230 patients with chronic kidney disease and 379 healthy controls representation (A). (B) Differential signal intensity of 273 urinary peptides separating patients with chronic kidney disease and healthy controls. By sequencing the urinary peptide fragments, the parental proteins can be identified (C). The multidimensional classifier CKD273 includes fragments of multiple parent proteins, which are involved in the pathogenic process (D). Reproduced with permission from Good et al. (2010). [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

determination is a fundamental step in the design of experiments (Mischak et al., 2010; Sui & Zheng, 2016). Over and beyond statistical considerations, sample size of any study depends on the hypothesis to be tested, the presumed effect size of an intervention or the presumed association size in an observational study, the variance and covariance matrices of the variables in an analysis data set. In addition, sample size calculations must be informed by prior knowledge and by the intended statistical methods that will be applied for analysis.

Methods for sample size determination are abundant for univariable analysis methods, but scarce in the multivariable case. Several publications provide guidance in computing sample size (Saccenti & Timmerman, 2016; Sui & Zheng, 2016) and are implemented in most statistical packages, such as R ([https://rpubs.com/mbounthavong/sample\\_size\\_power\\_analysis\\_R](https://rpubs.com/mbounthavong/sample_size_power_analysis_R)) or SAS (Castelloe & Cybrynski, 2017). By definition, omics are multivariable in nature and are commonly investigated using multivariable statistical methods, such as principal component analysis (PCA) and partial least squares analysis (PLS-A) for continuously distributed outcomes

or partial least squares discriminant analysis (PLS-DA) for categorical outcomes. No simple approaches to sample size determination exist for PLS-A and PLS-DA (Saccenti & Timmerman, 2016). Models for the univariable case can be extended to a multivariable design, when multiple omics signals are measured in one or more groups. For the two-group case the Hotelling  $T^2$  test (i.e., the multivariate extension of the classical  $t$ -test) can be used; multigroup cases and complex experimental designs involving repeated measures or different experimental factors can be addressed by multivariable analysis of variance, or variants thereof (Saccenti & Timmerman, 2016).

A crucial consideration related to sample size is that the results of any omics study must be confirmed in at least one independent sample set (Mischak et al., 2010). Bootstrapping or permutation strategies cannot confidently replace validation of the test results in one or more independent replication data sets. The sampling and characteristics of the validation population should be reported, and the analysis should be symmetrical in the test and validation data sets; any deviations should be reported (Table 1). The concept of having a test/discovery

**TABLE 1** Requirements for scientific reporting of proteomic biomarker data.

Describe and justify the clinical question, outcomes, and selection of subjects	Describe the clinical question and justify why it is of interest; describe what outcomes are assessed and comment on their clinical validity, potential for misclassification, and verification bias, if pertinent; clarify what the eligibility criteria for the selected study populations are and justify specific choices.
Describe the assessed subjects	Provide demographic information on ethnicity, sex, age, and concomitant medications at a minimum, and all relevant disease-related and clinical variables.
Describe sampling	Provide an accurate description of the sampling conditions and procedures (including the collection process and any manipulation of the sample before storage, the time between sampling and storage, storage conditions, and the addition of any protease inhibitors and/or preservatives). Justify the sampling choices according to the literature or supporting experimental data.
Describe experimental methodology	The procedure, as well as the observed standard deviation of technical specifications related to the procedure, should be given. To attribute the same identity to a certain feature in several independent analyses, accepted deviations of mass and other parameters (retention time, migration, position on gel, etc.) must be reported. Also, the observed deviation in identifying parameters and (relative) abundance, when the same sample is analyzed repeatedly, must be reported.
Describe the statistical evaluation	Provide details on determination of sample size, statistical analysis plan (for appraising calibration, discrimination, and/or reclassification), any consideration or adjustment for covariates (including treatment, whenever pertinent), methods for adjustment for multiplicity, and parameters used in complex machine-learning approaches, whenever pertinent. Clarify which analyses are predefined and which are post hoc.
Validate results	The results must be confirmed in at least one independent sample set. The sampling and characteristics of the validation population should be reported, and the analysis should be symmetrical in the test and validation data sets; any deviations should be reported.
Acknowledge limitations	No study is perfect; limitations and their potential impact on the results should be clearly acknowledged.
Take responsibility	The contributions of each author should be clearly stated.

data set and one or more replication data sets is applicable to all omics fields. Thinking before starting might also involve reading the required quality standards at the stage of submitting omics results for publication (Mischak et al., 2010).

### 3.2 | Preprocessing of the omics data

Preprocessing is a crucial step in analysis of any omics data. Preprocessing refers to preparing the omics data set for statistical analysis. However, each type of omics data has specific approaches to preprocessing, which is also dependent on the techniques applied to quantify markers in an omics data array (Leek et al., 2010; Voillet et al., 2016). For instance, in UPP, proteomic or metabolomic data sets, a strategy has to address signals below the detection limit of the applied technology, while

in genetic data sets missing genotypes might have to be imputed based on known sequence of the human genome and the recombination rate of the loci of interest (Yang et al., 2022). All omics approaches should include a stringent quality control policy ensuring that the data are reproducible.

There is a need to incorporate adjustment for batch effects as a standard step in the analysis of high-throughput data analysis along with normalization, exploratory analyses, and final significance calculation (Leek et al., 2010). Batch effects can for instance occur, when mass spectrometers are replaced with devices with higher precision. Batch effects, when occurring in large studies with a single omics technology, can be corrected for by using statistical methods, such as simply centering the data from each batch separately before combination or other more complex approaches (Leek et al., 2010). In terms of data integration in multiomics studies, the ideal

**TABLE 2** Schematic representation of the statistical workflow for the molecular analyses.

Analysis step	Methodology and statistical approach
Preprocessing of the omics data	Preprocessing is required to removed biases in the omics data for instance inherent to the technological platform used or originating from batch effects
Preparing for analysis	Checking distributions (Shapiro–Wilk or Kolmogorov–Smirnov test), logarithmic transformation, rank normalization, removal of outliers
Basic statistical approaches	Large-sample $z$ test, $t$ -test or ANOVA (means); $\chi^2$ statistic or Fisher exact test (proportion); log-rank test (survival functions); analyses across quantiles of biomarkers; scatterplots; standardization of rates
Identification of covariables	Stepwise linear or stepwise logistic regression
Analyses with <u>continuous</u> outcome	
Single urinary peptides, one at a time	
Cross-sectional analyses	Multivariable-adjusted linear regression, correction for multiple testing
Longitudinal analyses	Multivariable-adjusted linear regression (including adjustment for the baseline value of the outcome, if available) with correction for multiple testing
All markers	
Cross-sectional analyses	Partial least squares analysis
Longitudinal analyses	Partial least squares analysis
Analyses with <u>categorical</u> outcome	
Single markers, one at a time	
Cross-sectional analyses	Multivariable-adjusted logistic regression with correction for multiple testing
Longitudinal analyses	Multivariable-adjusted Cox regression with correction for multiple testing
All markers	
Cross-sectional analyses	Partial least squares discriminant analysis
Longitudinal analyses	Partial least squares discriminant analysis
Prediction of adverse outcomes	Integrated discrimination improvement, net reclassification improvement, optimized thresholds, $2 \times 2$ classification tables, log-rank test, receiver operating characteristic curve, c-statistic
Molecular pathways	PANTHER, DAVID, IPA, Cytoscape, Proteasix, ...

Abbreviations: ANOVA, analysis of variance.

situation is to have samples originating from the same biological source material. For instance, a piece of tissue may be cut into two sections, and one is used for metabolomics analysis, whilst the other is used to extract RNA.

In UPP and other omics, the handling of non-detectable or missing data is a key issue to be addressed. As mentioned earlier, the UPP includes over 20,000 peptides, of which 25% have been sequenced. UPP analyses often focus on sequenced peptides, because these peptides allow identification of the parent proteins from which they are derived. However, these peptides are seldom detectable in all study participants. There are several ways to deal with explanatory (omics) variables below the measurement threshold (Schisterman et al., 2006; Shaori & Dubé, 2018). In previously published papers relating adverse health outcomes to the UPP, one

requirement in the analysis was that sequenced urinary peptides should have detectable signal in at least 30% of participants (Martens et al., 2021) or in a more conservative approaches in at least 70% (Htun et al., 2017) or even 95% (Zhang et al., 2016) of study participants. Undetectable peptides were set at the distribution minimum (Lazar et al., 2016) or zero (He, Mischak, et al., 2021) before rank normalization. However, if a logarithmic transformation of an omics data set is being considered, missing values cannot be set to zero. In multiomics studies more complex imputation methods are necessary. A multifactor analysis (MFA) approach has been proposed and validated (Voillet et al., 2016). MFA compares and integrates multiple layers of information. Multiple imputation involves filling the missing rows with plausible values, resulting in M-completed data sets. MFA is then applied to each completed data set



to produce  $M$  different configurations (the matrices of coordinates of individuals). Finally, the  $M$  configurations are combined to yield a single consensus solution (Voillet et al., 2016).

### 3.3 | Preparing for analysis

For a long time, the analysis of omics data has been dominated by parametric statistics due to their theoretical soundness, relative ease of use, computational efficiency, and intuitive interpretation (Manduchi et al., 2022). Preparing for analysis refers to checking whether the assumptions underlying parametric statistics are fulfilled. Before statistical analysis, the distribution of continuous variables should be checked for deviation from normality. Most, if not all statistical software packages (e.g., MedCalc, SAS, SPSS, Stata, R, etc.), include the Shapiro–Wilk and Kolmogorov–Smirnov test to check whether the study sample has been generated from a normal distribution. The Shapiro–Wilk test is more appropriate for sample sizes less than 50, although it can also handle larger sample sizes, while Kolmogorov–Smirnov test is applicable when the sample size is 50 or larger. For both tests, the null hypothesis states that data are taken from normal distribution, so that a significant test statistic rejects this assumption.

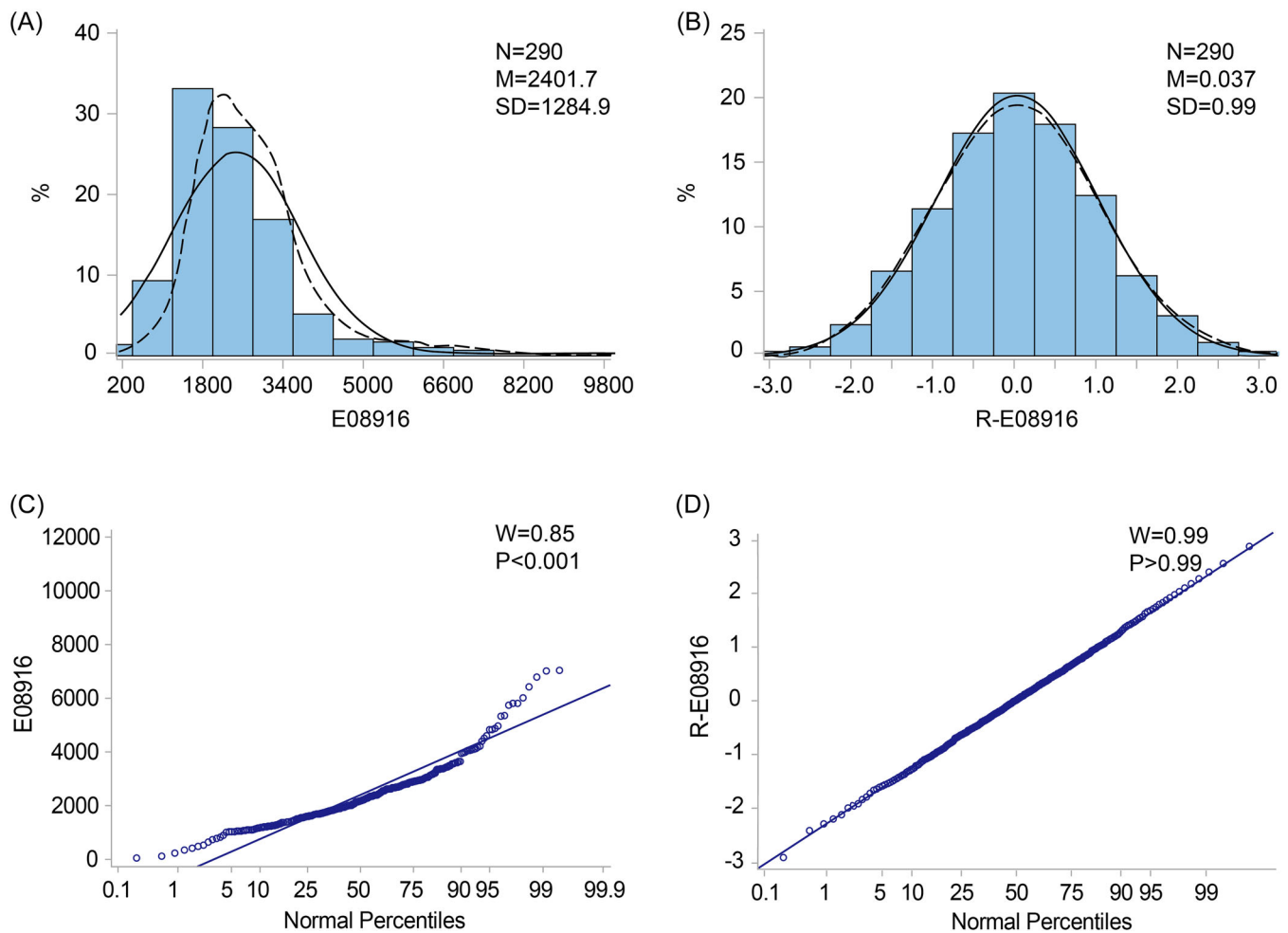
If the statistical methods applied for data analysis assume normally distributed variables, such variables have to be normalized by a logarithmic or other transformation. This is not necessary if nonparametric statistics are used, which often rely on ranking rather than numerical values. Given the normalization applied during the CE-MS procedure, the signals generated by the urinary peptides are dimensionless. Their distributions can be rank normalized (Figure 2) by sorting measurements from the smallest to the highest and then applying the inverse cumulative normal function (Blom, 1958). The distribution of multidimensional UPP classifiers generally do not violate the normality assumption to the extent that a transformation is needed (Tofte et al., 2020; Zhang et al., 2019). Over the past decade, ML has developed into a powerful instrument in the analysis of multiomics data, in which the number of signals to be analyzed far exceeds the number of individuals or experimental units. The assumptions underlying parametric statistics, such as normality of distributions, are not required to be fulfilled in ML algorithms (Cazaly et al., 2019; Manduchi et al., 2022) and for that matter also in exploratory methods, such as PLS-A.

Outliers can be removed if an individual's value is 3-SDs or more distant from the group mean of the

distribution after transformation if so required. However, removing outliers should be accompanied by checking potential experimental or biological reasons for this specific behavior or keying errors in entering clinical variables. Duplicate data entry and comparing the resulting data sets can remove most, albeit not all, of such errors, in particular, if the fault is in the (paper) source from which clinical data or ICD codes for the cause of mortality are entered. Duplicate entry is obviously not needed, when omics markers or clinical data are directly imported in the statistical software from the devices generating the omics signals or e-health records, respectively. However, in some countries, such as for instance Belgium, the General Data Protection Rules (Vlahou et al., 2021) makes direct entry of clinical data in an analysis data set very difficult, often requiring years of discussion with the relevant Ethics Committee, so that often the authors had to resort to paper files.

### 3.4 | Basic statistical approaches

Before any adjustment or multivariable modelling, most analyses will start by showing patient characteristics by categories of the “*exposure*” variable, for example, by quantiles of a multidimensional UPP classifier. For continuously distributed outcome variables (Figure 3), boxplots can be generated showing the association between the outcome of interest and the omics classifier (Staessen et al., 2022). Depending on the sample size and data structure, means can be compared using the large-sample  $z$ -test,  $t$ -test, or analysis of variance (ANOVA), proportions by the  $\chi^2$  statistic or Fisher's exact test, and survival function estimates by the log-rank test. The Fisher test is indicated, when cells in a frequency table are empty or include few study participants. The  $\chi^2$  statistic is typically used in case-control studies. In longitudinal studies, the change in classification variables can be addressed by application of the McNemar test. If prevalence or incidence rates need standardization across subgroups, for example, across cohorts making up the study population, sex, age groups, or any combination thereof, two approaches are commonly applied. One is used when the “*standard*” is the demographic structure of the study population (direct method) and the other is applied when the “*standard*” is a set to specific event rates observed in a reference population (indirect method). The direct standardization is the method of choice for large study samples, while the indirect one is applied to studies of relatively small dimensions (Tripepi et al., 2010). Standardized rates express the absolute risk associated with an exposure variable or categories thereof. The 95% confidence



**FIGURE 2** Rank normalization of a urinary peptide fragment derived from collagen 1. (A, B) Distribution plots before (A) and after (B) rank normalization; (C, D) Normal percentile plots before (C) and after (D) rank normalization. The solid and dotted lines represent the normal and kernel density distributions. *N*, *M*, and *SD* refer to the number of patients, the arithmetic means, and standard deviation. *W* is the Shapiro–Wilk statistic and *P* is the associated significance. A significant Shapiro–Wilk test indicates deviation from the normal distribution. [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

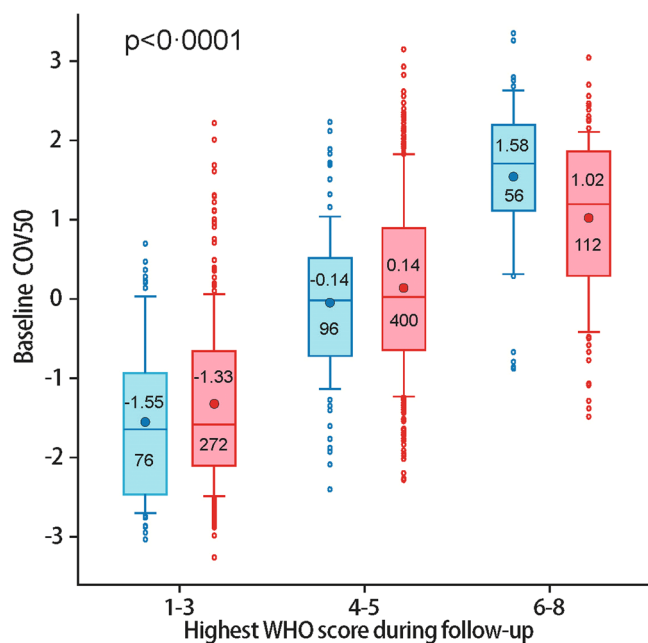
intervals of rates can be computed as  $R \pm 1.96 \times \sqrt{(R \times (100 - R) / T)}$ , where *R* is the rate and *T* is the number of patients at risk of developing an adverse outcome.

### 3.5 | Identification of application of covariables

To identify covariables to be retained in the analyses, continuous outcomes or categorical outcomes will be regressed on covariables of potential relevance, using a stepwise procedure with *p*-values for covariables to enter and stay in the model set at 0.15. A *p*-value of 0.15 allows for retaining covariables in the statistical modelling that are not formally significant, but might still be relevant. Linear regression is applicable for

continuously distributed outcomes, logistic regression for categorical—usually binary—outcomes, and proportional hazard (Cox) regression for modelling the log-linear association of time to a categorical outcome and an explanatory set of variables. Similar procedures, either based on *p*-values or other statistics including Akaike information criterion or the Bayesian information criterion, and so forth, are available in licensable or publicly accessible statistical software packages, such as R. Once a group of covariables is identified, a constant set should be used throughout a given analysis for all related continuous and categorical outcomes.

For continuous outcomes, accounting for covariables can be done by including the covariables plus the omics variable (e.g., the UPP classifier or a urinary peptides) in the same regression model, or alternatively, by standardization of the dependent variable, using the  $\beta$ -coefficients



**FIGURE 3** Boxplots showing the distributions of the urinary biomarker COV50 at baseline by the worst World Health Organization (WHO) score attained during follow-up in the initial (blue) and continued recruitment (pink) cohorts. The central line, the upper and lower lines, and the upper and lower caps represent the median, interquartile range, and the 10<sup>th</sup> to 90<sup>th</sup> percentile interval. The arithmetic means and extreme measurements are represented by circles inside the box and outside the whiskers, respectively. The arithmetic means and the number of data points contributing to each whisker plot is given within the boxes. The *p* value denotes the overall between-WHO category significance derived by analysis of variance (ANOVA). Reproduced from Staessen et al. (2022). [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

of the regression model associating the dependent variable with the covariables. Both approaches yield the same estimates (partial regression coefficient and *p*-value) for the association between the dependent and omics variable. In an exemplary case analysis of aortic pulse wave velocity (PWV) as outcome, stepwise regression identified sex ( $x_1$ ; coded 0,1), age ( $x_2$ ), body mass index ( $x_3$ ), heart rate ( $x_4$ ), mean arterial pressure ( $x_5$ ), smoking ( $x_6$ ; 0,1), daily consumption of alcoholic beverages ( $x_7$ ; 0,1), and the presence of type-2 diabetes ( $x_8$ ; 0,1) as being significantly associated with PWV (Hansen et al., 2006). In this example, PWV can be replaced by any continuously distributed dependent variable and the explanatory and omics variables by any different set of variables.

- Using plain adjustment, the model would be written as:  $y(\text{PWV}) = (\beta_1 \times x_1) + (\beta_2 \times x_2) + \dots + (\beta_8 \times x_8)$ , where  $\beta_1$ – $\beta_8$  are the partial regression coefficients relating the outcome to the covariables. Next the model can be

expanded by including the omics variable. The PWV variability explained by the model is subtracted from the actually observed PWV, thereby reducing the scale of PWV and rendering the interpretation of the model less straightforward.

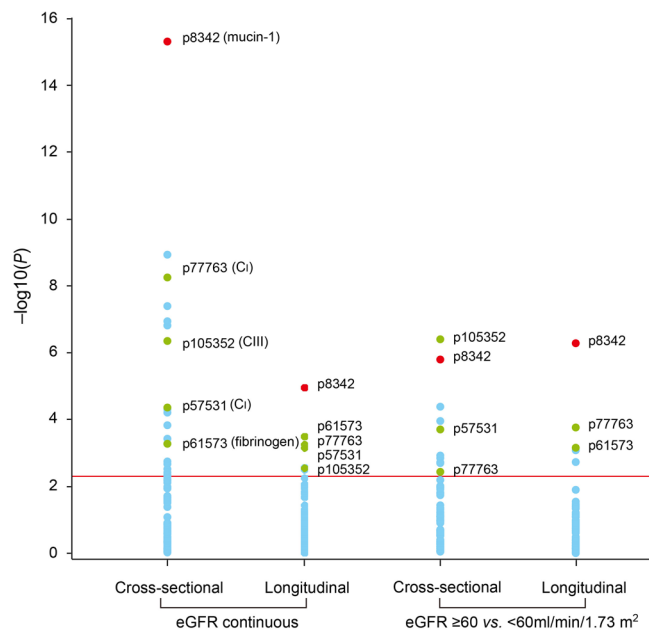
- As an alternative, before entering the UPP biomarker in the model, the continuously distributed outcome can be standardized to the average of the distributions in the study population. The standardized outcome variable can be computed as  $\text{PWV} - (\beta_1 \times (x_{1i} - x_{1m})) - (\beta_2 \times (x_{2i} - x_{2m})) - \dots - (\beta_8 \times (x_{8i} - x_{8m}))$ , where  $\beta_1$ – $\beta_8$  are the signed partial regression coefficients,  $x_{1i}$ – $x_{8i}$  are the values of the covariable in each individual, and  $x_{1m}$ – $x_{8m}$  are the population means of the covariables. With the standardized outcome as dependent variable (*y*), the omics variable can then be entered as independent variable. This approach will not rescale the dependent variable and make interpretation of the association between the dependent variable and the omics variable easily interpretable.

Using logistic and Cox regression, the same approach *mutatis mutandis* can be applied for standardization of the association between a categorical or binary outcome and an independent variable, for example, a multi-dimensional UPP classifier or single peptide fragments. Instead of the standardizing the continuously distributed outcome variable, here, standardization pertains to the predicted risk of each individual of experiencing an adverse health event or the predicted probability that an individual belongs to a category of interest.

### 3.6 | Continuous outcomes in relation to single urinary peptides

The same principles apply to the cross-sectional and longitudinal analyses of a continuous outcome. While accounting for covariables (Zhang et al., 2017b), as outlined in the previous section, the continuous outcome of interest can be regressed on each of the urinary peptides to construct  $-\log_{10}$  probability plots (Figure 4). In line with other omics-wide analyses, this approach can be referred to as a proteome-wide analysis. In this omics-wide approach, based on the number of UPP markers, a correction for multiple testing has to be applied, such as the Bonferroni correction, the Bonferroni step-down (Holm) correction (Holm, 1979), or the Benjamini and Hochberg false discovery rate (Benjamini & Hochberg, 1995). The Bonferroni correction is the most and the false discovery rate least stringent. For the thousands of signals included in any omics data set, the Bonferroni approach is too stringent and the false discovery rate is





**FIGURE 4**  $-\log_{10}(p)$  probability plot of the multivariable-adjusted associations of renal function phenotypes with the urinary peptides. Estimated glomerular filtration rate (eGFR) indicates the glomerular filtration rate derived from serum creatinine. All analyses were adjusted for mean arterial pressure, waist-to-hip ratio, smoking, plasma glucose,  $\gamma$ -glutamyltransferase, total-to-HDL cholesterol ratio, 24-h albuminuria, and use of diuretics, inhibitors of the renin-angiotensin system ( $\beta$ -blockers, angiotensin-converting-enzyme inhibitors, and angiotensin type-1 receptor blockers) and vasodilators (calcium-channel blockers and  $\alpha$ -blockers). The longitudinal analysis of change in eGFR as continuous variable was additionally adjusted for baseline eGFR and follow-up duration. The horizontal line denotes the significance level with Bonferroni correction applied. Red dots represent mucin-1 and green dots the other peptides passing the Bonferroni-corrected significance thresholds. Reproduced from Zhang et al. (2017b). [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

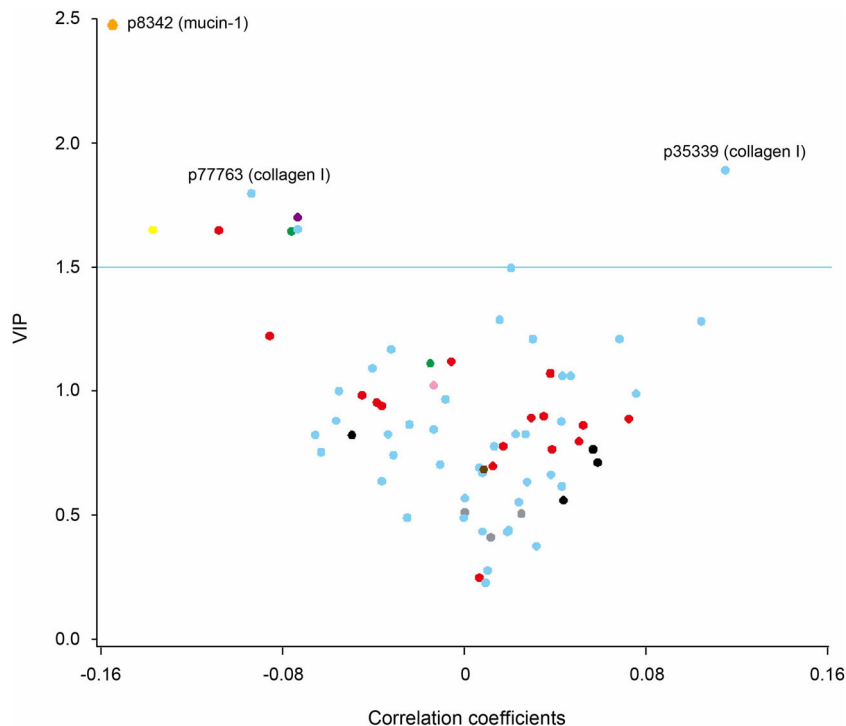
commonly applied. Other methods of dealing with the large number of models to be run—one for each omics signal—are based on the strong correlative nature of the omics features, for instance by accounting for the fact that multiple sequenced urinary peptide fragments are derived from the same parental protein, often a collagen (Martens et al., 2021). The peptide fragments originating from the same protein can be checked for the directionality of their association and ordered according to their multiple testing significance. If there is consistency in the directionality of the regression coefficients relating the outcome of interest to the set of urinary peptides originating from the same protein the fragment with the highest significance can be retained in the analysis. This data reduction step can be implemented by elastic net regression with determination of the  $\lambda_1$  and  $\lambda_2$  by

random cross-validation and a bootstrap procedure to obtain the final estimates of coefficients with 95% confidence interval (Zou & Hastie, 2005).

In longitudinal analyses, in which a continuous outcome (e.g., the glomerular filtration rate or right heart hemodynamic measurements) is predicted from a baseline biomarker, the baseline value of the trait of interest should be accounted for (Zhang et al., 2017b). If outcome, biomarkers and covariables are available at multiple time points, mixed models can be applied to account for clustering of observations within individuals. In such models, the biomarkers and covariables are modelled as fixed effects, whereas the random and unmeasurable variability between individuals (or clusters in a study) is accounted for as a random effect (Littell et al., 1996). Mixed models can also accommodate randomly missing values or for a variable number of time points per individual. All observations are used in the mixed model procedure, whereas in general linear models individuals with missing variables are discarded in the analysis (Littell et al., 1996).

For analyses of single sequenced urinary peptides, in multiple publications only those peptides with a detectable signal in at least 95% of participants might be analyzed (Huang, Trenson, et al., 2018; Zhang et al., 2016; Zhang et al., 2017b). However, ignoring biomarkers with missing values might waste potentially important information, explaining why in other studies of a more exploratory nature, this threshold was relaxed to 70% (Rossing et al., 2016) or even lower (Good et al., 2010). If the goal of the study is to gain deeper insight in pathophysiological pathways leading to adverse health outcomes, a conservative threshold (95% of participants with detectable signal), decreases the risk of false positive findings and reduces the penalty for multiple testing.

In analyses of multiple urinary peptides or metabolites, as applied in several publications by the authors of this review focusing on UPP (Huang, Van Keer, et al., 2018; Huang et al., 2019; Zhang et al., 2017b) or the circulating metabolome (Zhang, Marrachelli, et al., 2017), the supervised dimension reduction method PLS-A can be applied. PLS-A is a statistical technique that constructs models for continuous outcomes in relation to correlated high-dimensional explanatory variables (Bartel et al., 2013; Cavill et al., 2016; Csala et al., 2020; Tobias, 1997; Trygg & Wold, 2002). PLS-A allows identifying a set of independent latent factors that are linear combinations of the urinary peptides and that maximize the covariance between the omics markers (e.g., urinary peptides) and the variable describing the outcome of interest. The smallest number of latent factors can be retained in the analysis, as assessed by the van der Voet T2 statistic. The importance of each urinary



**FIGURE 5** V-plots generated by partial least square analysis. Variable Importance in Projection (VIP) scores indicate the importance of each urinary fragment in the construction of the partial least square factors and are plotted against the centered and rescaled correlation coefficients. The correlation coefficients reflect the associations of the multivariable-adjusted eGFR with the urinary fragments. Fragments associated with reduced eGFR (left side of the V-plot) include, among others, p8342 and p77763. p35339 was associated with higher eGFR (right side of the V-plot). Colors identify fragments derived from collagen I (blue), II (grey), III (red), IV (brown), the mucin-1 subunit  $\alpha$  (orange), fibrinogen (green), protocadherin-12 (purple), retinol-binding protein 4 (pink), stabilin-2 (yellow) and uromodulin (black). For details, refer to Zhang et al. (2017b). [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

marker in the construction of the PLS-A factors will subsequently be assessed from the Variable Importance in Projection (VIP) scores of Wold with the threshold set at approximately 1.5 or an alternative value. PLS-A allows constructing V-plots (Figure 5), in which VIP scores and the rescaled and centered correlation coefficient among variables are plotted along the vertical and horizontal axis, respectively (Zhang et al., 2017b). Plotted biomarkers associated with high VIP score and low correlation coefficients (top left quadrant of the plot) identify predictors of an adverse outcome, whereas those associated with high VIP score but high correlation coefficient (top right quadrant) is inversely associated with an outcome. The PLS approach does not require to adjust for multiple testing, because the minimum set of latent factors is analyzed in relation to the outcome variable in a single run.

### 3.7 | Categorical outcomes in relation to single urinary peptide

Analyses of categorical outcomes will follow the same principles as those with continuous outcomes. Multivariable-adjusted relative risk can be computed by logistic regression or Cox regression. Logistic regression is appropriate for cross-sectional designs or prospective analyses, in which the follow-up duration is approximately similar in all patients. To model time to an

adverse health outcome or until the censoring date, Cox regression will be the approach of choice. Mixed models can also accommodate categorical outcomes. If multiple biomarkers are assessed simultaneously, then constructing  $-\log_{10}$  plots provides a way to present the results graphically and to adjust for multiple testing. Partial least square discriminant analysis (PLS-DA) combines highly correlated biomarkers into a single analysis and allows constructing V-plots for categorical outcomes.

### 3.8 | Evaluation of added diagnostic or predictive accuracy

If the discriminatory threshold of a biomarker is known, computing its diagnostic or predictive value can be simply done from  $2 \times 2$  tables providing sensitivity, specificity, positive and negative predictive value, and the misclassification rate. Optimal discrimination limits for a biomarker can be determined by maximizing the Youden index, that is, the maximum of sensitivity plus specificity minus 1 (Ruopp et al., 2008).

The added value of a biomarker (continuous or categorical), over and beyond a set of covariables, can be assessed from the integrated discrimination improvement (IDI) and the net reclassification improvement (NRI) (Pencina et al., 2008, 2011). IDI is the difference between the discrimination slopes of the basic model and the basic model extended with the biomarker.

The discrimination slope is the difference in predicted probabilities (%) between subjects with and without endpoint. To calculate NRI, the risk of an adverse health outcome in each individual can be derived from a Cox model with and without the omics biomarker. If  $P(\text{up}/\text{event})$  is the percentage of subjects with events whose predicted probability is increased by adding the biomarker to the model and if  $P(\text{up}/\text{nonevent})$  is the percentage of subjects without events whose predicted probability is increased, then NRI equals  $2 \times (P[\text{up}/\text{event}] - P[\text{up}/\text{nonevent}])$ . IDI and NRI provide complementary information. Indeed, if adding a biomarker to a model increases the predicted probability in cases, this is reflected by a significant increase in IDI, while NRI indicates the extent by which a biomarker improves diagnostic accuracy. Although applied frequently, expert statisticians suggested that IDI and NRI have limitations (Kerr et al., 2014). If IDI and NRI are computed, they recommended retaining existing descriptive terms, such as the true-positive and false-positive classification rates, or testing the null hypothesis of no prediction increment from modelled regression coefficients (Kerr et al., 2014). Finally, the capability to discriminate between patients with or without adverse health outcomes can also be assessed by constructing receiver operating characteristic (ROC) curves and by calculating the area under the ROC curve (AUC), as shown in Figure 6. The DeLong method provides a way to compute 95% confidence intervals of the AUC.

### 3.9 | Molecular pathways

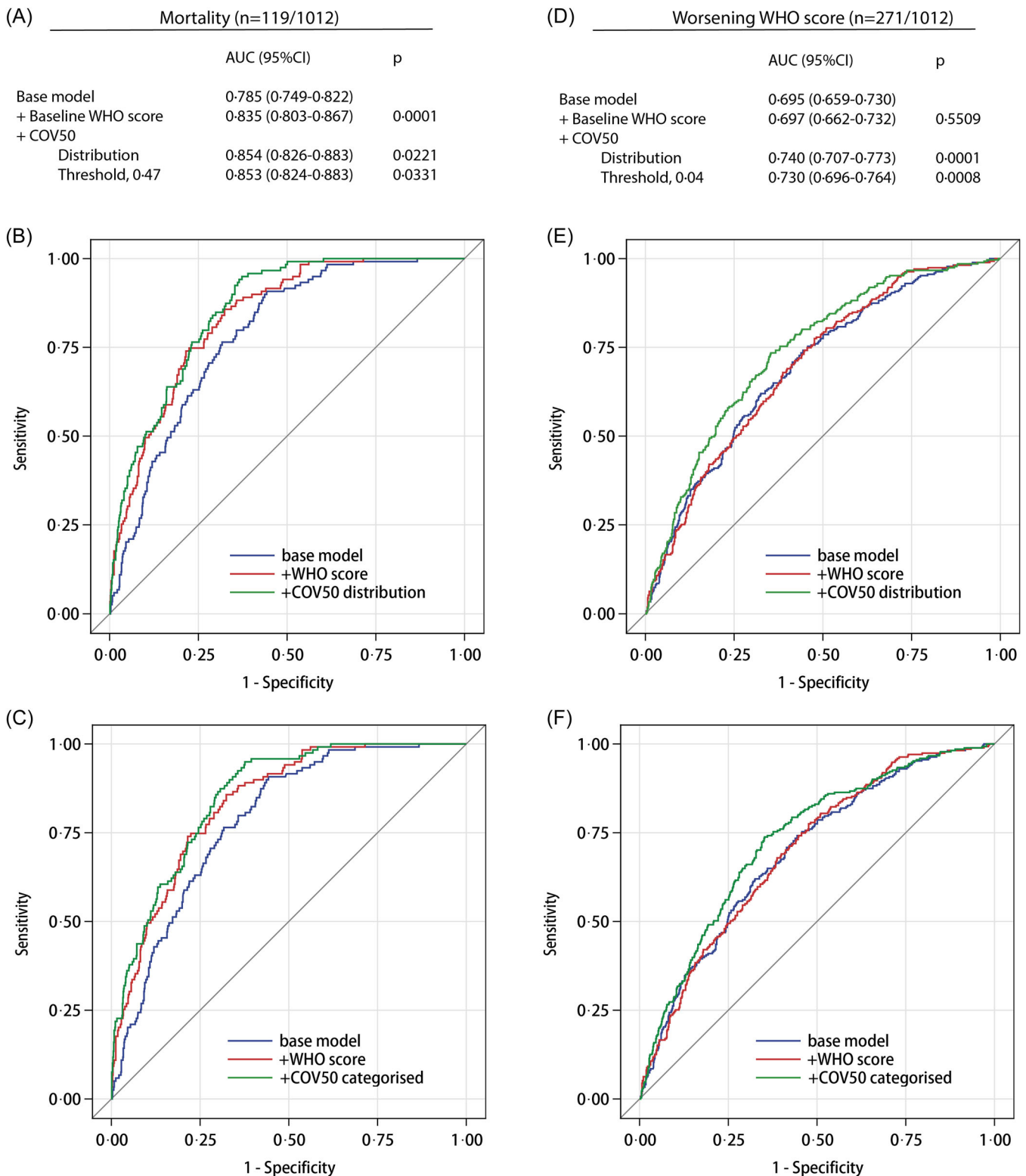
To ensure detection of relevant molecular pathways and to build a network of biologically meaningful interactions, advanced bioinformatics tools should be used in combination with the literature (Bhat et al., 2015; Latosinska et al., 2017). Functional analysis of the features can be performed using open-source tools, such as the Protein Annotation Through Evolutionary Relationship (PANTHER) software (Mi et al., 2013) or the DAVID software suite for pathway and functional annotation (Huang et al., 2009). Additional tools include Ingenuity Pathway Analysis (IPA), Cytoscape's plugins like ClueGO and CluePedia, and the GO, KEGG, and REACTOME databases. In addition, in the case of UPP analyses, proteases responsible for the generation of urinary biomarkers should be investigated *in silico*, using Proteasix (Klein et al., 2013) and the information obtained in the pathway analysis. The hypothesis is that changes in protease activity might be linked to disease pathophysiology. However, proteases active

along the nephron and distal urinary tract might affect the urinary peptide fragments detected by UPP analysis. However, in a placebo-controlled study of a dipeptidyl peptidase-4 inhibitor (Siwy et al., 2019), the UPP included pairs of peptide chains, that is, the substrate for the protease activity (e.g., PPGPPGKNGDDGEAGKPG) and the resulting breakdown product (e.g., GPPGKNGDDGEAGKPG). Thus, protease analyses are useful to check this possibility.

## 4 | MULTIOMICS APPROACHES

State-of-the-art next-generation sequencing, transcriptomics, proteomics, and other high-throughput omics technologies enable the efficient generation of large experimental data sets on the same individual or experimental unit, but require dimension reduction approaches (Csala & Zwinderman, 2019; Meng et al., 2016). Canonical correlation analysis (CCA) and redundancy analysis (RDA) are widespread in the omics data analysis field. Simply stated, CCA is a technique for analyzing the relation between two sets (or groups) of variables (Kuhfield et al., 2017). Each set can contain multiple variables. Given two sets of variables, canonical correlation analysis finds a linear combination from each set, called a canonical variable, such that the correlation between the two canonical variables is maximized. This correlation between the two canonical variables is the first canonical correlation. The first canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. The coefficients of the linear combinations are canonical coefficients or canonical weights. Canonical coefficients can be normalized, such that each canonical variable has a variance of 1. Canonical correlation analysis continues by finding a second set of canonical variables, uncorrelated with the first pair that produces the second-highest correlation coefficient. The process of constructing canonical variables continues until the number of pairs of canonical variables equals the number of variables in the smaller group. Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set (Kuhfield et al., 2017).

From the late 2000s, statisticians developed modified versions of CCA that are better adapted to the high-dimensional structure of multiomics data. Among them, penalized canonical correlation analysis, regularized canonical correlation analysis, sparse canonical correlation analysis, and penalized canonical correlation analysis. These procedures applied a form of penalization to the organic CCA framework, which makes penalized



**FIGURE 6** Performance of the COV50 urinary marker on top of other baseline risk factors in the full data set for contrasting mortality versus survival (A–C) and for progression versus nonprogression in the baseline World Health Organization (WHO) score during follow-up (D–F). The base model included sex, age, body mass index, and the presence of comorbidities: hypertension, heart failure, diabetes, and cancer. In subsequent steps, the baseline WHO score was added and next COV50 as a continuously distributed variable (B, E) or as a categorized variable based on an optimized threshold of 0.47 for mortality (C) or 0.04 for a worsening WHO score (F). At each step, the *p*-values are for the comparison with the preceding model. For details, refer to Staessen et al. (2022). [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



forms of CCA applicable to high-dimensional data and, in most cases, results in models that include only a subset of the original variables from the data sources (i.e., variable selection). Variable selection is a desirable property when the original variables are too numerous to be interpretable in the results of the analysis, which is exactly the case with multiomics data. A description of the exact properties of variable selection, which depends on the type of penalization applied, is beyond the scope of this article and has been reviewed and put into context elsewhere (Csala & Zwinderman, 2019; Wang et al., 2022).

## 5 | ML

ML refers to approaches by which computers learn from data to accomplish certain tasks, without a programmer having to specify every single algorithmic instruction (Manduchi et al., 2022). ML is particularly applicable to the omics data, in which the number of predictors (e.g., single urinary peptides) is much larger than the number of individuals or experimental units. Supervised ML involves generating a predictive model, which through a training step learns a general rule to produce a desired output. Once the general rule is established, the trained model can be applied to new input data of the same type. The input data consist of independent or explanatory data, in which for instance the individual in an omics study represents the observation. The dependent or response variable is the output of interest (Manduchi et al., 2022), for instance coronary heart disease in relation to a multitude of risk factors (Forrest et al., 2023). ML procedures can have parameters and hyperparameters. Parameters are internal configuration variables learned from the data, while hyperparameters refer to values that have been specified before the ML starts. The root mean squared error, that is, the square root of the averaged squared differences between true and predicted values is a metric describing the accuracy of a predictive ML model. However, there are many other choices that can be selected when optimizing algorithms and hyperparameters (Zheng, 2015). In most cases, one has to tune the choice of the ML algorithm and its hyperparameters. This can be done using an independent validation set, with samples drawn from the same population as the training set. A standard approach is the so-called the  $k$ -fold cross-validation. This procedure involves subdividing the input data into  $k$  subsets of equal size. One of the subsets serves to validate the algorithm and the hyperparameters as determined in the remaining  $k-1$  subsets. The procedure is  $k$  times repeated. Finally, the selection yielding the best average performance across

the  $k$ -fold runs is accepted and the corresponding model is fit to the entire training set and evaluated on a hold-out testing set (Manduchi et al., 2022). Multiple permutations are an inspection/interpretation technique that can also be used to interpret any fitted “black-box” ML model and to understand, which features drive the trained estimator (Casalicchio et al., 2019).

Any ML application should be carefully configured and tuned with as finality to provide reproducible results, whenever it is applied (Manduchi et al., 2022). Fortunately, Automated ML is available to biomedical researchers, in particular the intended readers of this article, who have little experience in writing ML procedures. Automated ML refers to automating single steps in a ML algorithm, such as feature engineering or hyperparameter optimization of an algorithm for a specific scientific objective. Moreover, auto ML code able to handle multiple tasks can be downloaded from several software sites. Such applications are particularly appealing to nonexpert users as they provide tailor-made solutions (Waring et al., 2020). A detailed description of the multitude of approaches to auto ML falls beyond the scope of this article and has been reviewed in detail elsewhere (Manduchi et al., 2022; Waring et al., 2020). Needless to state that over the past decade open-source software has stimulated the exponential growth of supervised and non-supervised ML applications. Open source specifically refers to making the source code for the software publicly available, either by distributing the software directly as source code at no cost, or by maintaining a source code repository, which end-users can change and improve. Names of the open-source algorithms and download sites have been listed in previously published reviews focused on ML (Manduchi et al., 2022; Waring et al., 2020).

## 6 | CONCLUSIONS

Basically, the analysis of adverse health outcomes in relation to omics data rests on the same statistical principle as any other data collected within large population or patient cohorts. The only difference resides in the large number of biomarkers (“exposure” data points), which all have to be considered simultaneously. This requires planning ahead how data will be structured, imported in statistical software packages, results will be triaged on relevance, and how markers will be presented to the readers. From a more general viewpoint, the introduction and use of omics markers in clinical and population science revolutionized thinking and almost unlimitedly expanded the working horizon of scientists. Epidemiological studies in population or well-defined



patient cohorts with long follow-up provide unbiased estimates of the prevalence and prognostic significance of risk factors and health-related events and generate the ultimate validation of potential disease-causing mechanisms identified in experimental studies. Conversely, population and patient studies also generate hypotheses for mechanisms to be tested in experimental studies. Thus, omics studies will increasingly contribute to personalized evidence-based medicine and translating experimental findings into clinically applicable strategies for prevention and disease management.

## ACKNOWLEDGMENTS

OMRON Healthcare, Co., Ltd., Kyoto, Japan provided a nonbinding grant to the Alliance for the Promotion of Preventive Medicine (APREMED), Mechelen, Belgium. Internal Funds KU Leuven (STG-18-00379) currently support the Research Unit Hypertension and Cardiovascular Epidemiology, KU Leuven Department of Cardiovascular Sciences, University of Leuven, Leuven, Belgium.

## CONFLICT OF INTEREST STATEMENT

Agnieszka Latosinska is an employee of Mosaiques-Diagnostics GmbH, Hannover, Germany. Harald Mischak is the cofounder and co-owner of Mosaiques-Diagnostics GmbH. The other authors declare no conflict of interest.

## ORCID

De-Wei An  <http://orcid.org/0000-0002-1395-7050>

Yu-Ling Yu  <http://orcid.org/0000-0002-8255-3770>

Dries S. Martens  <http://orcid.org/0000-0001-7893-3642>

Agnieszka Latosinska  <http://orcid.org/0000-0001-8917-2412>


Zhen-Yu Zhang  <https://orcid.org/0000-0002-3785-7417>

Harald Mischak  <http://orcid.org/0000-0003-0323-0306>

Tim S. Nawrot  <http://orcid.org/0000-0002-3583-3593>

Jan A. Staessen  <http://orcid.org/0000-0002-3026-1637>

## TWITTER

Jan A. Staessen  @jasta49

## REFERENCES

- Bartel J, Krumsiek J, Theis FJ. 2013. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* 4, e201301009.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B* 57, 289–300.
- Bhat A, Heinzl A, Mayer B, Perco P, Mühlberger I, Husi H, Merseburger AS, Zoidakis J, Vlahou A, Schanstra JP, Mischak H, Jankowski V. 2015. Protein interactome of muscle-invasive bladder cancer. *PLoS One* 10, e0116404.
- Blom G. 1958. *Statistical estimates and transformed beta-variables*. 1st ed. New York/Stockholm: Wiley/Almquist and Wiksell.
- Casalicchio G, Molnar C, Bischl B. 2019. Visualizing the feature importance for black box models. In: *Machine Learning and Knowledge Discovery in Databases* (Berlingerio M, Bonchi F, Gärtner T, eds.). Cham, Switzerland: Springer International Publishing, 665–670.
- Castelloe J, Cybrynski M. 2017. Chapter 91. The POWER Procedure. In: *SAS/STAT® 14.3 User's Guide* (Baxter A, Huddleston E, eds.). Cary, North Carolina, USA: SAS Institute Inc., 7353–7602.
- Cavill R, Jennen D, Kleinjans J, Briede JJ. 2016. Transcriptomic and metabolomic data integration. *Brief Bioinform* 17, 891–901.
- Cazaly E, Saad J, Wang W, Heckman C, Ollikainen M, Tang J. 2019. Making sense of the epigenome using data integration approaches. *Front Pharmacol* 10, 126.
- Csala A, Zwinderman AH. 2019. Multivariate statistical methods for high-dimensional multiset omics data analysis. In: *Computational Biology* (Husi H, ed.). Brisbane, Australia: Codon Publications, 71–83.
- Csala, A, Zwinderman AH, Hof MH. 2020. Multiset sparse partial least squares path modeling for high dimensional omics data analysis. *BMC Bioinformatics* 21, 9.
- Forrest IS, Petrazzini BO, Duffy Á, Park JK, Marquez-Luna C, Jordan DM, Rocheleau G, Cho JH, Rosenson RS, Narula J, Nadkarni GN, Do R. 2023. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *Lancet* 21, 215–225.
- Good DM, Zürgbig P, Argilés A, Bauer HW, Behrens G, Coon JJ, Dakna M, Decramer S, Delles C, Dominiczak AF, Ehrlich JH, Eitner F, Fliser D, Frommberger M, Ganser A, Girolami MA, Golovko I, Gwinner W, Haubitz M, Herget-Rosenthal S, Jankowski J, Jahn H, Jerums G, Julian BA, Kellmann M, Kliem V, Kolch W, Krolewski AS, Luppi M, Massy Z, Melter M, Neusüss C, Novak J, Peter K, Rossing K, Rupperecht H, Schanstra JP, Schiffer E, Stolzenburg JU, Tarnow L, Theodorescu D, Thongboonkerd V, Vanholder R, Weissinger EM, Mischak H, Schmitt-Kopplin P. 2010. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Moll Cell Proteomics* 9, 2424–2437.
- Hansen TW, Staessen JA, Torp-Pedersen C, Rasmussen S, Thijs L, Ibsen H, Jeppesen J. 2006. Prognostic value of aortic pulse wave velocity as index of arterial stiffness in the general population. *Circulation* 113, 664–670.
- He T, Mischak M, Clark AL, Campbell RT, Delles C, Díez J, Filippatos G, Mebazaa A, McMurray JJV, González A, Raad J, Stroggilos R, Bosselmann HS, Campbell A, Kerr SM, Jackson CE, Cannon JA, Schou M, Girerd N, Rossignol P, McConnachie A, Rossing K, Schanstra JP, Zannad F, Vlahou A, Mullen W, Jankowski V, Mischak H, Zhang Z, Staessen JA, Latosinska A. 2021. Urinary peptides in heart failure: a link to molecular pathophysiology. *Eur J Heart Fail* 23, 1875–1887.
- He T, Melgarejo JD, Clark AL, Yu YL, Thijs L, Díez J, López B, González A, Cleland JG, Schanstra JP, Vlahou A, Latosinska A, Mischak H, Staessen JA, Zhang ZY, Jankowski V. 2021b. Serum and urinary biomarkers of

- collagen type-I turnover predict prognosis in patients with heart failure. *Clin Transl Med* 11, e267.
- Holm S. 1979. A simple sequentially rejective Benferroni test. *Scand J Stat* 6, 65–70.
- Htun NM, Magliano DJ, Zhang ZY, Lyons J, Petit T, Nkuipou-Kenfack E, Ramirez-Torres A, von Zur Muhlen C, Maahs D, Schanstra JP, Pontillo C, Pejchinovski M, Snell-Bergeon JK, Delles C, Mischak H, Staessen JA, Shaw JE, Koeck T, Peter K. 2017. Prediction of acute coronary syndromes by urinary proteome analysis. *PLoS One* 12, e0172036.
- Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44–57.
- Huang QF, Trenson S, Zhang ZY, Van Keer J, Van Aelst LNL, Yang WY, Nkuipou-Kenfack E, Thijs L, Wei FF, Mujaj B, Ciarka A, Droogné W, Vanhaecke J, Janssens S, Van Cleemput J, Mischak H, Staessen JA. 2018. Biomarkers to assess right heart pressures in recipients of a heart transplant: a proof-of-concept study. *Transplant Direct* 4, e346.
- Huang QF, Van Keer J, Zhang ZY, Trenson S, Nkuipou-Kenfack E, Van Aelst LNL, Yang WY, Thijs L, Wei FF, Ciarka A, Vanhaecke J, Janssens S, Van Cleemput J, Mischak H, Staessen JA. 2018. Urinary proteomic signatures associated with  $\beta$ -blockade and heart rate in heart transplant recipients. *PLoS One* 13, e0204439.
- Huang QF, Zhang ZY, Van Keer J, Trenson S, Nkuipou-Kenfack E, Yang WY, Thijs L, Vanhaecke J, Van Aelst LNL, Van Cleemput J, Janssens S, Verhamme P, Mischak H, Staessen JA. 2019. Urinary peptidomic biomarkers of renal function in heart transplant recipients. *Nephrol Dial Transplant* 34, 1336–1343.
- Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. 2014. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology* 25, 114–121.
- Klein J, Lacroix C, Caubet C, Siwy J, Zürgbig P, Dakna M, Muller F, Breuil B, Stalmach A, Mullen W, Mischak H, Bandin F, Monsarrat B, Bascands JL, Decramer S, Schanstra JP. 2013. Fetal urinary peptides to predict postnatal outcome of renal disease in fetuses with posterior urethral valves (PUV). *Sci Transl Med* 5, 198ra106.
- Kuhfield WF, Kuo A, Sarle WS, Watts DL. 2017. Chapter 30. The CANCORR Procedure. In: *SAS/STAT® 14.3 User's Guide* (Baxter A, Huddleston E, eds.). Cary, North Carolina, USA: SAS Intitute Inc., 1891–1920.
- Latosinska A, Mokou M, Makridakis M, Mullen W, Zoidakis J, Lygirou V, Frantzi M, Katafigiotis I, Stravodimos K, Hupe MC, Dobrzynski M, Kolch W, Merseburger AS, Mischak H, Roubelakis MG, Vlahou A. 2017. Proteomics analysis of bladder cancer invasion: targeting EIF3D for therapeutic intervention. *Oncotarget* 8, 69435–69455.
- Latosinska A, Siwy J, Mischak H, Frantzi M. 2019. Peptidomics and proteomics based on CE-MS as a robust tool in clinical application: the past, the present, and the future. *Electrophoresis* 40, 2294–2308.
- Lazar C, Gatto L, Ferro M, Bruley C, Burger T. 2016. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J Proteome Res*, 15, 1116–1125.
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev Genet* 11, 733–739.
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD. 1996. Chapter 3. Analysis of Repeated Measures Data. In: *SAS System for Mixed Models*. Cary, North Carolina, USA: SAS Institute Inc., 87–134.
- Manduchi E, Romano JD, Moore JH. 2022. The promise of automated machine learning for the genetic analysis of complex traits. *Hum Genet* 141, 1529–1544.
- Martens DS, Thijs L, Latosinska A, Trenson S, Siwy J, Zhang ZY, Wang C, Beige J, Vlahou A, Janssens S, Mischak H, Nawrot TS, Staessen JA; FLEMENGHO Investigators. 2021. Urinary peptidomics to address age-related disabilities: a prospective population study with replication inpatients. *Lancet Healthy Longevity* 2, e690–e703.
- Mavrogeorgis E, Mischak H, Latosinska A, Siwy J, Jankowski V, Jankowski J. 2021. Reproducibility evaluation of urinary peptide detection using CE-MS. *Molecules* 26, 7260.
- Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. 2016. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 17, 628–641.
- Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8, 1551–1566.
- Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, Bennett SE, Bischoff R, Bongcam-Rudloff E, Capasso G, Coon JJ, D'Haese P, Dominiczak AF, Dakna M, Dihazi H, Ehrlich JH, Fernandez-Llama P, Fliser D, Frokiaer J, Garin J, Girolami M, Hancock WS, Haubitz M, Hochstrasser D, Holman RR, Ioannidis JP, Jankowski J, Julian BA, Klein JB, Kolch W, Luider T, Masy Z, Mattes WB, Molina F, Monsarrat B, Novak J, Peter K, Rossing P, Sánchez-Carbayo M, Schanstra JP, Semmes OJ, Spasovski G, Theodorescu D, Thongboonkerd V, Vanholder R, Veenstra TD, Weissinger E, Yamamoto T, Vlahou A. 2010. Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* 2, 46ps42.
- Mischak H, Schanstra JP. 2011. CE-MS in biomarker discovery, validation, and clinical application. *Proteomics Clin Appl* 5, 9–23.
- Pencina MJ, D'Agostino, Sr. RB, D'Agostino, Jr. RB, Vasan RS. 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 27, 157–172.
- Pencina MJ, D'Agostino, Sr. RB, Steyerberg EW. 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 30, 11–21.
- Pontillo C, Zhang ZY, Schanstra JP, Jacobs L, Zürgbig P, Thijs L, Ramirez-Torres A, Heerspink HJL, Lindhardt M, Klein R, Orchard T, Porta M, Bilous RW, Charturvedi N, Rossing P, Vlahou A, Schepers E, Glorieux G, Mullen W, Delles C, Verhamme P, Vanholder R, Staessen JA, Mischak H, Jankowski J. 2017. Prediction of chronic kidney disease stage 3 by CKD273, a urinary proteomic biomarker. *KI Reports* 2, 1066–1075.

- Rossing K, Bosselmann HS, Gustafsson F, Zhang ZY, Gu YM, Kuznetsova T, Nkuipou-Kenfack E, Mischak H, Staessen JA, Koeck T, Schou M. 2016. Urinary proteomics pilot study for biomarker discovery and diagnosis in heart failure with reduced ejection fraction. *PLOS ONE* 11(6), e0157167.
- Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. 2008. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* 50, 419–430.
- Saccetti E, Timmerman ME. 2016. Approaches to sample size determination for multivariate data: applications to PCA and PLS-DA of omics data. *J Proteome Res* 15, 2379–2393.
- Schisterman EF, Vexler A, Whitcomb BW, Liu A. 2006. The limitations due to exposure detection limits for regression models. *Am J Epidemiol* 163, 374–383.
- Shaori N, Dubé JS. 2018. Toward improved analysis of concentration data: embracing nondetects. *Environ Toxicol Chem* 37, 643–656.
- Siwy J, Klein T, Rosler M, von Eynatten M. 2019. Urinary proteomics as a tool to identify kidney responders to dipeptidyl peptidase-4 inhibition: a hypothesis-generating analysis of the MERLINA-T2D trial. *Proteomics* 13, 1800144.
- Staessen JA, Wendt R, Yu YL, Kalbitz S, Thijs L, Siwy J, Raad J, Metzger J, Neuhaus B, Papkalla A, von der Leyen H, Mebazaa A, Dudoignon E, Spasovski G, Milenkova M, Canevska-Taneska A, Salgueira Lazo M, Psychogiou M, Rajzer MW, Fuławka Ł, Dzitkowska-Zabielska M, Weiss G, Feldt T, Stegemann M, Normark J, Zoufaly A, Schmiedel S, Seilmaier M, Rumpf B, Banasik M, Krajewska M, Catanese L, Rupprecht HD, Czerwieńska B, Peters B, Nilsson Å, Rothfuss K, Lübbert C, Mischak H, Beige J; CRIT-CoV-U Investigators. 2022. Predictive performance and clinical application of COV50, a urinary proteomic biomarker in early COVID-19 infection: a cohort study. *Lancet Digital Health* 4, e727–e737.
- Sui I, Zheng L. 2016. Topics in study design and analysis for multistage clinical proteomic studies. In: *Statistical Analysis in Proteomics* (Jung K, ed.). Methods in Molecular Biology (MIMB, 1362) New York, NY: Humana Press, 29–61.
- Tobias RD. 1997. *An introduction to partial least squares regression*. Cary, NC: SAS Institute Inc., 1250–1257.
- Toft N, Lindhardt M, Adamova K, Bakker SJL, Beige J, Beulens JWJ, Birkenfeld AL, Currie G, Delles C, Dimos I, Francová L, Frimodt-Møller M, Girman P, Göke R, Havrdova T, Heerspink HJL, Kooy A, Laverman GD, Mischak H, Navis G, Nijpels G, Noutsou M, Ortiz A, Parvanova A, Persson F, Petrie JR, Ruggenenti PL, Rutters F, Rychlík I, Siwy J, Spasovski G, Speeckaert M, Trillini M, Zürlbig P, von der Leyen H, Rossing P; PRIORITY Investigators. 2020. Early detection of diabetic kidney disease by urinary proteomics and subsequent intervention with spironolactone to delay progression (PRIORITY): a prospective observational study and embedded randomised placebo-controlled trial. *Lancet Diabetes Endocrinol* 8, 301–312.
- Tripepi G, Jager KJ, Dekker FW, Zoccali C. 2010. Stratification for confounding—Part 2: direct and indirect standardization. *Nephron Clin Pract* 116, c322–c325.
- Trygg J, Wold S. 2002. Orthogonal projections to latent structures (O-PLS). *J Chemom* 16, 119–128.
- Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A, Bischoff R, Black PC, Boehm F, Céraline J, Chrousos GP, Delles C, Evenepoel P, Fridolin I, Glorieux G, van Gool AJ, Heidegger I, Ioannidis JPA, Jankowski J, Jankowski V, Jeronimo C, Kamat AM, Masereeuw R, Mayer G, Mischak H, Ortiz A, Remuzzi G, Rossing P, Schanstra JP, Schmitz-Dräger BJ, Spasovski G, Staessen JA, Stamatialis D, Stenvinkel P, Wanner C, Williams SB, Zannad F, Zoccali C, Vanholder R. 2021. Data sharing under the general data protection regulation: time to organize law and research ethics? *Hypertension* 77, 1029–1035.
- Voillet V, Besse P, Liaubet L, San Cristobal M, Gonzalez I. 2016. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics* 17, 402.
- Wang T, Renteria ME, Peng J. 2022. Editorial: data mining and statistical methods for knowledge discovery in diseases based on multimodal omics. *Front Genet* 13, 895796.
- Waring J, Lindvall C, Umeton R. 2020. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Art Intell Med* 101, 101822.
- Wasung ME, Chawla LS, Madero M. 2015. Biomarkers of renal function, which and when? *Clin Chim Acta* 438, 350–357.
- Yang WY, Izzi B, Bress AP, Thijs L, Citterio L, Wei FF, Salvi E, Delli Carpini S, Manunta P, Cusi D, Hoylaerts MF, Lutun A, Verhamme P, Hardikar S, Nawrot TS, Staessen JA, Zhang ZY. 2022. Association of colorectal cancer with genetic and epigenetic variation in PEAR1—a population-based cohort study. *PLoS One* 17, e0266481.
- Zhang ZY, Thijs L, Petit T, Gu YM, Jacobs L, Yang WY, Liu YP, Koeck T, Zürlbig P, Jin Y, Verhamme P, Voigt JU, Kuznetsova T, Mischak H, Staessen JA. 2015. Urinary proteome and systolic blood pressure as predictors of 5-year cardiovascular and cardiac outcomes in a general population. *Hypertension* 66, 52–60.
- Zhang ZY, Ravassa S, Yang WY, Petit T, Pejchinovski M, Zürlbig P, López B, Wei FF, Pontillo C, Thijs L, Jacobs L, González A, Koeck T, Delles C, Voigt JU, Verhamme P, Kuznetsova T, Díez J, Mischak H, Staessen JA. 2016. Diastolic left ventricular function in relation to urinary and serum collagen biomarkers in a general population. *PLoS One* 11, e0167582.
- Zhang ZY, Marrachelli VG, Yang WY, Trenson S, Huang QF, Wei FF, Thijs L, Van Keer J, Monleon D, Verhamme P, Voigt JU, Kuznetsova T, Redón J, Staessen JA. 2017. Left ventricular function in relation to circulating metabolic biomarkers: cross-sectional and longitudinal observations in a general population. In: *OMICS as Tool to Address the Burden of Non-Communicable Age-Related Disease in Populations in Epidemiological Transition*. Acta Biomedica Lovaniensa Leuven University Press, 459–495.
- Zhang ZY, Nkuipou-Kenfack E, Staessen JA. 2019. Urinary peptidomic biomarker for personalized prevention and treatment of diastolic left ventricular dysfunction. *Proteomics Clin Appl* 13, e1800174.
- Zhang ZY, Ravassa S, Pejchinovski M, Yang WY, Zürlbig P, López B, Wei FF, Thijs L, Jacobs L, González A, Voigt JU, Verhamme P, Kuznetsova T, Díez J, Mischak H, Staessen JA. 2017b. A



urinary fragment of mucin-1 subunit  $\alpha$  is a novel biomarker associated with renal dysfunction in the general population. *KI Reports* 2, 811–820.

Zheng A. 2015. *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. Sebastopol, CA: O'Reilly Media Inc.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J Royal Stat Soc B* 37, 301–320.

## AUTHOR BIOGRAPHIES



**De-Wei An** is a specialization student in KU Leuven, Belgium. He received his PhD degree in internal medicine from Shanghai Jiao Tong University School of Medicine. His research interests lie in cardiovascular health, specifically in exploring the relationship between hypertension and arterial stiffness, and how this can be measured through noninvasive techniques such as pulse wave velocity, pulse wave analysis, and urinary proteomics.



**Yu-Ling Yu** is a PhD candidate in biomedical science in KU Leuven, Belgium. She received her bachelor's degree of clinical medicine in Southern Medical University and her master's degree of internal medicine in Guangdong Cardiovascular Institute, China. Her research interests are mainly focused on clinical research in cardiovascular disease, urinary proteomics, and environmental toxicity.



**Dries S. Martens** obtained his degree in bioscience engineering, and molecular biology at the University of Leuven, Belgium, prior to completing a PhD at the Centre for Environmental Science at Hasselt University, Belgium. Currently, he holds a postdoctoral fellowship at Hasselt University where his research primarily revolves around studying and comprehending the variations in early-life molecular ageing markers. His research interests lie in the gene-environmental determinants of telomere length in newborns and the possible health consequences of telomere length at birth. Apart from this, he also investigates the impact of air pollution exposure on molecular ageing markers, including omics-derived signatures, throughout the lifespan.



**Agnieszka Latosinska** received her PhD from the Charité Universitätsmedizin Berlin (Germany) in 2016. Her research was advanced by two consecutive EU-funded Marie Skłodowska-Curie programs (ITN-BCMolMed, IF-PCaProTreat), focusing on translational research. As a part of these projects, she had the opportunity to share time between academic (Biomedical Research Foundation of the Academy of Athens, Greece) and industrial laboratories (Mosaiques Diagnostics, Germany). Currently, she directs cardiovascular and drug discovery research at Mosaiques Diagnostics. She has published more than 60 manuscripts, cited >1100 times (H-index: 20). Her research interest includes the application of proteomics and system biology approaches for investigating pathophysiology and defining biomarkers and potential drugs.



**Harald Mischak** received his PhD from the Technical University Vienna and subsequently worked on kinases at the NCI/NIH, and at GSF in Munich. He was involved in proteomics research for over 30 years, with emphasis on identifying biomarkers and therapeutic targets. He founded Mosaiques Diagnostics and initiated the use of urinary proteomics and CE-MS for clinical application, focusing on kidney and cardiovascular disease. With over 400 scientific articles on signaling and proteomics that have been cited over 30000 times, he is a leading expert in proteome research, personalized medicine, and applied systems biology.



**Tim S. Nawrot** heads the environmental and molecular epidemiology research line at Hasselt University and is part-time professor at the Environment and Health Unit of Leuven University (Belgium). He has initiated the ENVIRONAGE (ENVIRONMENTAL influence ON AGEing in early life) birth cohort. He is specially interested in early life exposures and how this might impact telomere length to understand the health disease continuum over the life course. He is active on several advisory bodies including the EU summit on air pollution (Vilnius declaration), Dutch Health Council, World Health Organisation, and Health Effect Institute. He is associate editor of Environmental Health, editor in chief for children studies for Frontiers in Public Health.



**Jan A. Staessen** is Emeritus Professor of Medicine at the University of Leuven and former Head of Clinic at the University Hospitals Leuven. Dr Staessen's current research interests focus on the population science, clinical genetics, the clinical application of omics biomarkers, environmental medicine, and the treatment of cardiovascular disease, in particular resistant hypertension. Dr Staessen founded the not-for-profit Research Institute Alliance for the Promotion of Preventive Medicine (APPREMED; <https://www.appremed.org>). The mission of APPREMED is the development of new strategies to prevent chronic age-related disorders in people, who do not yet have any symptoms, but who

based on their clinical characteristics and biomarkers (e.g., the urinary proteome), have a high risk of developing target organ damage and thus becoming a “patient.” The focus of APPREMED is therefore on primary prevention.

**How to cite this article:** An, D.-W., Yu, Y.-L., Martens, D. S., Latosinska, A., Zhang, Z.-Y., Mischak, H., Nawrot, T. S., & Staessen, J. A. Statistical approaches applicable in managing OMICS data: Urinary proteomics as exemplary case. *Mass Spectrometry Reviews*, (2023);1-18. <https://doi.org/10.1002/mas.21849>