Analysis & Sensing

Research Article
doi.org/10.1002/anse.202200094

Chemistry Europe
European Chemical Societies Publishing

www.analysis-sensing.org

ATR-FTIR Spectroscopy for Early Detection of Diabetic Kidney Disease

Zack Richardson,[a, b] Adele Kincses,[a, b] Elif Ekinci,[c, d, e] David Perez-Guaita,*[a, f] Karin Jandeleit-Dahm,*[b, g] and Bayden R. Wood*[a]

Current screening methods for diabetic kidney disease (DKD), characterized by albumin excretion in urine, are expensive or only identify patients in late disease stages. Hence, there is need for a cost-effective, quick, and portable screening tool which identifies patients at DKD onset. Here we report that ultracentrifugation coupled with infrared spectroscopy and machine learning can identify and quantify low level microalbuminuria in urine samples from a cohort of diabetic patients (n = 155) and controls (n = 22). Independent testing of the methods indicated that classification analysis discriminated between normo- and micro/macroalbuminuric samples with sensitivity of >91% and specificity of >99%. Regression methods quantified albumin concentration in the samples with error values of 17 and 44 mg/L for normo- and microalbuminuric patients. Using only 700 μL of sample, this approach identifies patients at an earlier stage of disease than a urinary dipstick, whilst also yielding results cheaper and faster than the albumin to creatinine ratio.

Introduction

Diabetes mellitus is a chronic disease characterised by increased circulating glucose levels. The body is unable to regulate the excess glucose due to insufficient insulin production by the beta cells of the pancreas, or the tissues having a reduced ability to respond to insulin. In 2021, 536.6 million people were living with diabetes,[1] and it was the direct cause of 1.5 million deaths in 2019.[2] Caseloads are forecast to rise to 783.2 million people living with diabetes by 2045,[1] with prevalence similar in developed and developing countries.[3]

Excess circulating glucose causes both micro- and macrovascular complications. Epithelial cells respond to hyperglycaemia by upregulating glucose metabolism and flux, causing reactive oxygen species (ROS) formation and metabolic dysfunction.[4–5] Excess ROS oxidises cellular componentry, including DNA, proteins, and lipids. In the kidneys, this leads to accelerated inflammation and fibrosis, ultimately resulting in diabetic kidney disease (DKD).[6–7] This complication is characterised by increased albumin excretion in the urine and progressive loss of renal function. In 2019, there were 135 million patients, or approximately 30% of the global diabetic population, living with DKD.[8] It is the leading cause of end stage renal disease (ESRD) worldwide, with diabetic patients approximately 20 times more likely to progress to ESRD than non-diabetic patients.[9] All diabetic patients are recommended to undergo yearly kidney function tests, which assess the estimated glomerular filtration rate (eGFR), and albumin excretion, to monitor disease progression.

Both functional and structural changes are observed in the kidney as DKD progresses. There are five stages of eGFR (G1–G5), with patients in G1 considered to have normal kidney function, whilst patients in G5 are developing ESRD.[10] Albumin excretion is assessed using the urinary albumin creatinine ratio (UACR). It is broken down into three categories; normoalbumi-

[a] Z. Richardson,+ A. Kincses,+ Dr. D. Perez-Guaita, Prof. B. R. Wood
Centre for Biospectroscopy, School of Chemistry
Monash University
Clayton, Victoria, 3800 (Australia)
E-mail: david.perez-guaita@uv.es
bayden.wood@monash.edu
[b] Z. Richardson,+ A. Kincses,+ Prof. K. Jandeleit-Dahm
Department of Diabetes
Central Clinical School
Monash University
Melbourne, Victoria, 3004 (Australia)
E-mail: karin.jandeleit-dahm@monash.edu
[c] Prof. E. Ekinci
The Australian Centre for Accelerating Diabetes Innovations (ACADI)
The University of Melbourne
Parkville, Victoria, 3010 (Australia)
[d] Prof. E. Ekinci
Department of Endocrinology
Austin Health
Heidelberg, Victoria, 3084 (Australia)
[e] Prof. E. Ekinci
Department of Medicine, Melbourne Medical School
The University of Melbourne
Parkville, Victoria, 3010 (Australia)
[f] Dr. D. Perez-Guaita
Department of Analytical Chemistry
University of Valencia
Dr Moliner 50, Burjassot, 46100 (Spain)
[g] Prof. K. Jandeleit-Dahm
German Diabetes Centre
Institute for Clinical Diabetology
Heinrich Heine University
Dusseldorf, 40225 (Germany)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/anse.202200094

nuria (UACR of Males: <2.5 mg/mmol; Females: <3.5 mg/mmol or albumin concentration of <20 mg/L), microalbuminuria (UACR of Males: 2.5–25 mg/mmol; Females: 3.5–35 mg/mmol or albumin concentration of 20–200 mg/L) and macroalbuminuria (UACR of Males: >25 mg/mmol; Females: >35 or albumin concentration >200 mg/L).[10–11] Progression to G4 and/or macroalbuminuria is associated with irreversible nephron damage.[10] Testing efforts are focussed on identifying patients before the disease reaches this stage as early intervention reduces the likelihood of progression to ESRD and other cardiovascular complications.[10]

Albumin is the most common biomarker for DKD, as its excretion is most greatly upregulated with kidney damage, compared to baseline levels.[12] At present it is the strongest predictor of progression to ESRD and other complications associated with diabetes, including cardiovascular events.[13–14] Dipsticks are currently the primary screening tool for albuminuria. These rapid tests have a protein limit of detection (LoD) of 150–300 mg/L,[15] making them well suited for detection of macroalbuminuria, but they struggle to detect microalbuminuria. Evaluation of popular dipstick brands found they have sensitivity and specificity of 72% (95% C.I.: 68–77%) and 82% (95% C.I.: 76–89%), respectively, for microalbuminuria detection.[16] Conversely, the gold standard measure of albumin excretion is the urinary albumin creatinine ratio (UACR).[10] This test quantifies albumin content using an enzyme linked immunosorbent assay (ELISA), whereby changes in sample absorbance upon exposure to albumin antibodies is related to the albumin content of the urine; this is then corrected for urinary creatinine[17] This test can detect albumin concentrations in urine down to 2 mg/L, however it is expensive, time consuming, and can only be performed in a pathology laboratory, limiting its point of care use, particularly in rural and remote settings. Thus, there is a need for a cheap and fast test for microalbuminuria identification which can be delivered in a point of care context. Furthermore, albumin excretion in urine can fluctuate with intense exercise, the menstrual cycle and stress, so a test which can be regularly performed whilst minimising patient inconvenience is important.

Fourier transform infrared spectroscopy (FTIR) is a popular analytical chemistry technique which enables non-destructive, rapid analysis of a wide range of sample types. All biological materials absorb IR light at specific wavelengths, to produce a characteristic vibrational fingerprint for each sample which is governed by the specific macromolecules present.[18] Attenuated total reflection (ATR) is a well characterised collection mode of FTIR spectroscopy. A small volume (<5 μL) of sample is directly deposited onto the ATR crystal and dried. Incident radiation totally internally reflects along the crystal, penetrating several microns into the sample to collect sample information.[19] Modern ATR-FTIR spectrometers are compact and light devices, making them ideally suited to point-of-care applications.

ATR-FTIR spectroscopy has been previously applied to quantify creatinine, urea, phosphate, uric acid, pH and sulphate and cystinuria.[20–23] We have recently demonstrated that mid-IR transflection spectroscopy can also quantify 8-isoprostane, a marker of ROS, in urine.[24] Furthermore, this technique has application, in combination with multivariate modelling techniques, for the detection of banned substances in urine.[25] The Amide II band, due to C–N stretching and NH_2 bending, correlates well to albuminuria levels at high concentrations.[26] However, the prediction of proteins from the direct measurement of urine or dried urine is strongly limited by their low concentration and the presence of major interference such as urea and creatinine.[27] Shaw *et al.* evaluated ATR spectroscopy of dried urine films for a wide range of concentrations and concluded that the method had a limit-of-detection of 5000 mg/L,[28] and attempts to differentiate between different DKD disease stages without sample pre-treatment steps have so far been unsuccessful.[29] Sarigal *et al.* recently identified several age and sex-based variations in ATR-FTIR spectra of urine from healthy volunteers which may have confounded previous differentiation attempts.[23]

Centrifugal ultrafiltration is a simple sample preparation technique where samples are subject to high spin speeds to maximise the amount of sample passed through the filter pores, which has greatly enhanced ATR-FTIR metabolite prediction of serum by eliminating proteins from the analytes.[30–31] In a recent study, we introduced the combination of centrifugal ultrafiltration and ATR spectroscopy for the quantification of microproteinuria and macroproteinuria in urine.[27] In urine, the selection of appropriate pore size enriches protein concentration up to 160-fold[32] and dilutes components which normally dominate the spectra, like glucose and urea, allowing more robust quantification of proteins, including microalbuminuria (>200 mg/L) and normoalbuminuria levels (<20 mg/L).

Since the former proof of concept study was performed over artificial urine and spiked urine of healthy samples, the present study aims to evaluate the methodology in real patients suffering from DKD.[27] To that end, we measured the spectra of urines from a cohort of controls and patients with DKD and employed machine learning algorithms to create and validate classification models that can discriminate between normo-, micro- and macroalbuminuric patients as well as regression methods that can predict the concentration of proteins at normal and pathological ranges.

Results and Discussion

Cohort characteristics

A total of 177 urine samples were sourced for this investigation. Of these, 64 were from patients with normoalbuminuria, 61 were from patients with microalbuminuria, 30 were from patients with macroalbuminuria and 22 were from healthy individuals. The characteristics of each sample group have been provided in Table 1. Only samples from males were used, and patient sample groups were age matched and contain a similar proportion of samples with Type 1 vs Type 2 diabetes. The proportion of patients in each group who had reduced kidney filtration capacity, determined by eGFR, increased with

Research Article
doi.org/10.1002/anse.202200094
Chemistry Europe
European Chemical Societies Publishing

Table 1. Characteristics of the diabetic patients and healthy volunteers in the study. Reduced eGFR is defined as eGFR < 60 mL/min/1.73 m2.

	Sample class	N	Mean age (range)	N reduced eGFR [%]	N Type II [%]
Non-diabetic	Normoalbuminuria	22	29.9 (19–48)	N/A	N/A
Diabetic	Normoalbuminuria	64	67.1 (42–85)	26 (48)	58 (91)
	Microalbuminuria	61	69.6 (43–88)	36 (59)	52 (85)
	Macroalbuminuria	30	67.5 (48–87)	21 (70)	26 (87)

worsening disease status. This was anticipated as kidney filtration capacity often decreases with increasing albuminuria in DKD. Urine samples from healthy volunteers were from a much younger cohort, with a mean age of 29.9 years. This cohort served as a "zero-control" population, as young people are less likely to have other health conditions, which alter the urinary excretion profile and thus potentially confound results. At present, this work does not include an age matched control group without a diagnosis of diabetes mellitus. There were more samples sourced from the normoalbuminuric and microalbuminuric groups, as being able to identify patients at the early stages of microalbuminuria is the main clinical focus.

Spectra of filtered urine

Prior to recording the spectra, urine samples underwent a filtration process. The syringe filter with a cellulose acetate membrane was used to remove large scale particulates, whilst retaining the proteins within the sample. The centrifuge filter pre-concentrated analytes of interest, like albumin, whilst removing dominant urine components, like urea and glucose. A spectrum of baseline corrected filtered urine was plotted for a visual inspection of the data. Figure 1a shows a clear trend of increasing sample absorbance with increasing disease severity, with urine samples from healthy individuals (grey) having the least absorbance, followed by samples from patients with normoalbuminuria (red), microalbumuria (green), whilst samples from patients with macroalbuminuria (blue) were the most absorbent. As previously shown,[27] the intensity of the bands

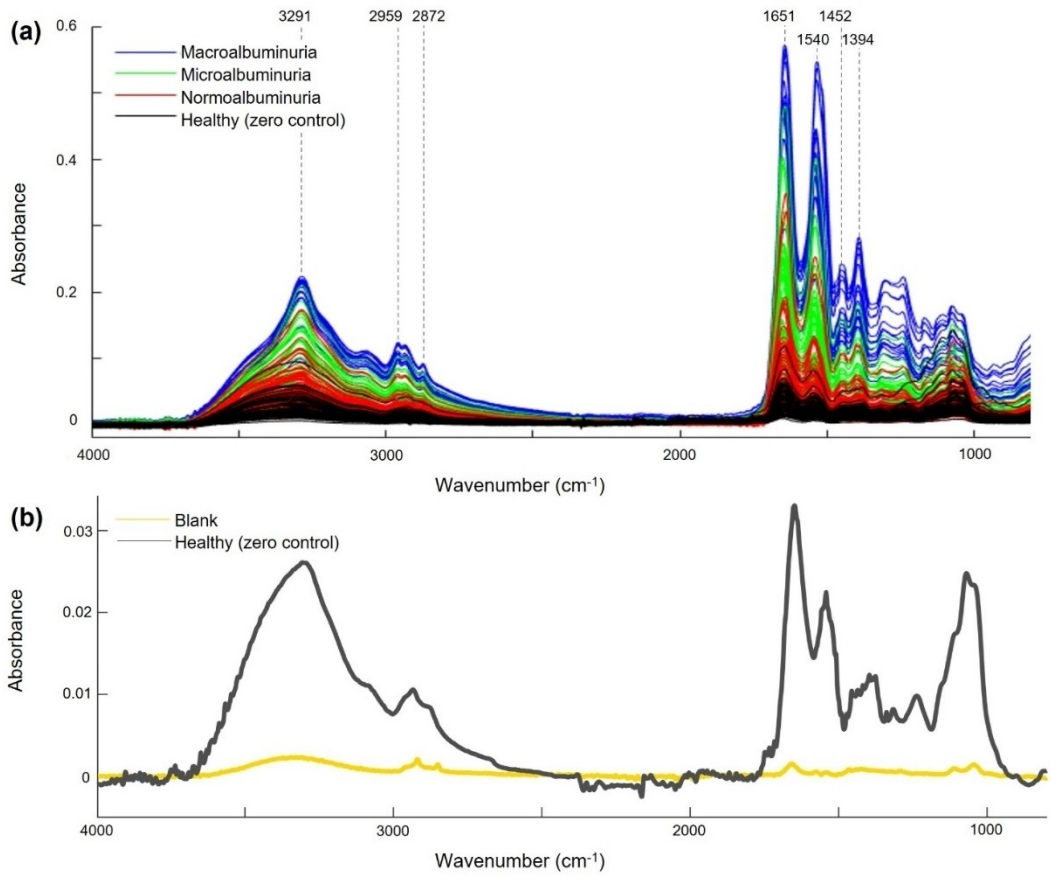

Figure 1. (a) Raw spectra of filtered urine, coded by sample class; healthy (grey), normoalbuminuric (red), microalbuminuric (green) and macroalbuminuric (blue). Absorbance increases with worsening disease status. (b) Mean blank extraction spectra (n = 6) (yellow) contrasted to the mean spectra from the zero control samples (grey).

Research Article
doi.org/10.1002/anse.202200094
Chemistry Europe
European Chemical Societies Publishing

increased with the amount of proteins deposited onto the ATR crystal. Since the volume of dried urine is constant for all the samples, sample absorbance is directly proportional to analyte concentration in each sample. Thus, the increase in sample absorption with worsening disease status indicates that there are analytes present, which increase in concentration as DKD progresses. The increasing absorbance at 1651, 1540, 1452 and 1394 cm^{-1} is associated with albumin (see Supplementary Material 1). This demonstrates that the data is suitable for regression analysis. The OH stretch is observed at 3291 cm^{-1} and correlates with amide I and II modes. The amide I mode would have contributions from the water bending mode at ~1636 cm^{-1}. The increase in absorbance below 1000 cm^{-1} in the high protein content samples is also indicative of bound water and so too the relatively weak bend + libration water combination band (centered at 2130 cm^{-1}). The higher the protein content the more bound water in the sample. It is not possible to remove all water as this is tightly bound by protein, so there is a correlation with protein and bound water.

Principal Component Analysis (PCA) of spectra

Following observations from the raw spectra that absorbance of the amide modes increased with disease progression, a PCA was performed on the dataset. PCA decomposes complex data sets, like spectra, into "principal components" (PCs) that are linear combinations of the original variables that produce a plot where each data point represents a spectrum plotted against the PCs. It is used to look for patterns and clustering in the data set. The corresponding loadings plots can be used to ascertain which bands are responsible for the variance in the data set and determine if there are spectral features that can be used to separate the samples into different classes. This analysis was performed over 3500–2700 cm^{-1} and 1800–800 cm^{-1} on baseline corrected, and mean centred spectra, with the best clustering within groups and separation of different groups achieved on a PC1 vs PC3 scores plot (Figure 2a). PC1 explained 97% of spectral variance. The loadings plot of this PC (Figure 2b) shows a typical spectrum of albumin with the corresponding bands at 1652, 1542, 1453 and 1398 cm^{-1}. The score values of this PC increased with the protein content of the classes considered as evinced by the fact that PC1 represented the albumin content. PC3 (SMX) accounted for less than 1% of the spectral variance. The loading values of PC3 show positive bands in the 1100–1030 cm^{-1} region, which are oppositely correlated to the amide bands from proteins. These positive bands may indicate glycogen (or one of its derivatives) excreted in urine, but further investigation to elucidate the identity of these bands are required. Samples from healthy individuals were clustered very closely together, whilst the tightness of the clustering decreased as disease severity increased, with the macroalbuminuric samples being the most spread out. This is anticipated as

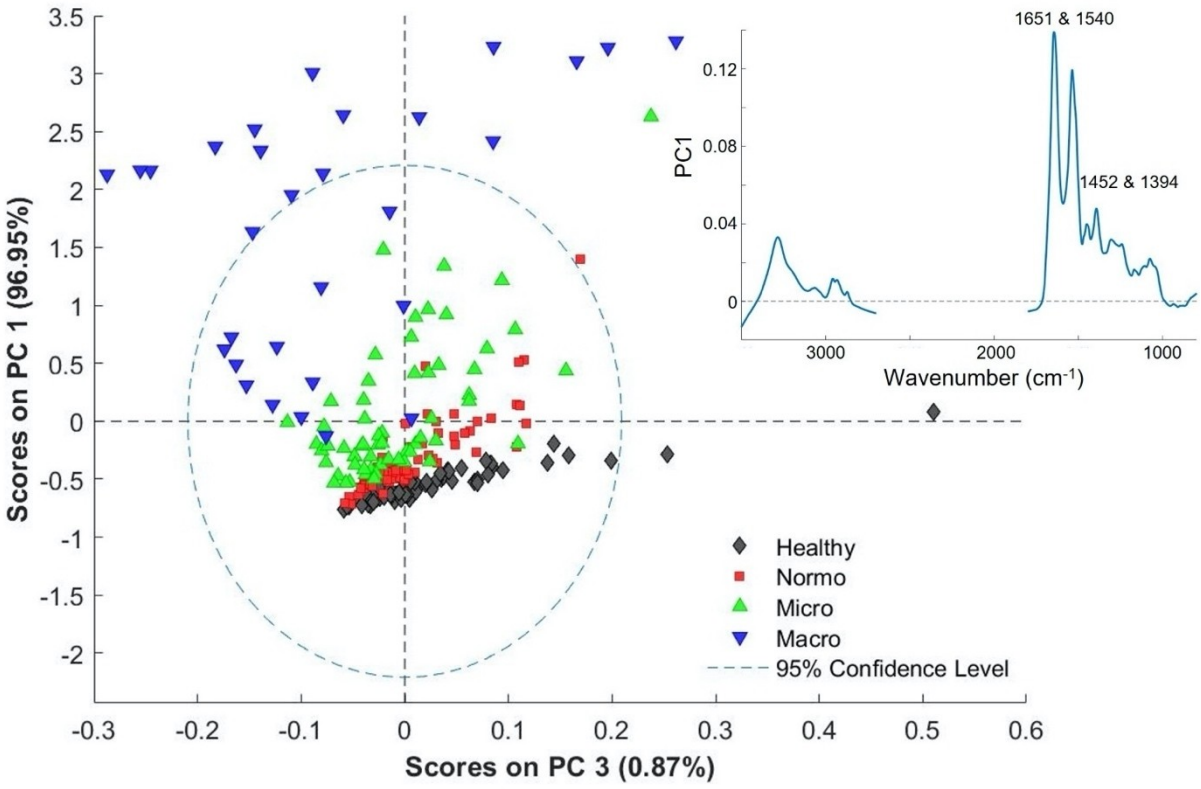

Figure 2. (a) PCA plot for PC1 vs PC3 on baseline corrected, mean centred spectra, with (insert) PC1 loadings plot. This PC accounts for 97% of the total variance and is dominated by peaks related to albumin. Each point is representative of a single spectrum.

Research Article
doi.org/10.1002/anse.202200094
Chemistry Europe
European Chemical Societies Publishing

patients with severe DKD are likely to be taking a wide range of medications and have other concomitant diseases, which alter the metabolite excretion profile. The specific nature of these medications and disease will vary from patient to patient. There is some degree of overlap between the sample groups. This is expected since disease class is based on urinary albumin content, which is a continuous scale. Samples which have albumin content near the disease class cut off points will have very similar spectra, but are in different disease classes, thus leading to overlap. Future work in this area should include an age matched, diabetes negative control group to assess the technique's ability to separate matched samples who do not have diabetes mellitus, but may have other diseases or consume other medication, from the diabetic population. Nonetheless, there is clear clustering of the different sample groups in the PCA plot, which indicates that this data is well suited for classification analysis.

Discriminant analysis

Two types of discrimination models were applied to the data set, with the objective of separating samples from patients with normoalbuminuria from samples from patients with micro- and macroalbuminuria. Modelling parameters of the models along with statistical values can be found in Table 2 and Figure 3, respectively. When partial least squares discriminant analysis (PLSDA) was applied the model returned sensitivity and specificity values of 89.7% (52/58), 97.8% (43/44), respectively, for the cross validation (CV) data and 90.9% (30/33), 100% (19/19), respectively, for the prediction data set. By comparison, the support vector machine discriminant analysis (SVMDA) model was more successful at classifying urine samples based on their albumin excretion level. The prediction of independent data, with sensitivity and specificity of 97% (32/33) and 100% 19/19, respectively, demonstrates excellent diagnostic performance in identifying patients with elevated albuminuria. There is a small misclassification rate, with approximately 3–10% of micro- and macroalbuminuric patients being incorrectly identified as having normal levels of albumin excretion. A degree of misclassification is expected as samples are assigned to a disease class based on a continuous scale. Figure 3 shows the probability of a sample being classed as diseased (micro- or macroalbuminuria) predicted by the model (independent test set) versus the actual content of protein calculated by the reference method. The probability of a sample being classified as micro- or macroalbuminuric increases sharply when albumin content approaches 20 mg/L, featuring a grey zone for the diagnostic. However, compared to urinary dipsticks, which have been evaluated to have sensitivity and specificity values as low

Table 2. Model parameters selected in the optimization of the discrimination models (Using cross validation) and figures of merits obtained for the cross validation and independent prediction. NOTE: FD, first derivative; SD, second derivative; SNV, standard normal variate normalization; MC; Mean Centering; LVs, Number of latent variables selected for the PLSDA model or for the compression performed for the before the SVM-DA.

	PLS-DA CV (n = 102)	PLS-DA Independent Testing (n = 52)	SVM-DA CV (n = 102)	SVM-DA Independent Testing (n = 52)
Spectral Region (cm^{-1})	3140-2770, 1800–1400		3140–2770, 1800–800	
Preprocessing	FD + SNV + MC		SD + SNV + MC	
LVs	4		4	
Sensitivity (%)	89.7	90.9	94.8	97.0
Specificity (%)	97.8	100.0	95.5	100.0
AUROC	0.977	0.992	1	1

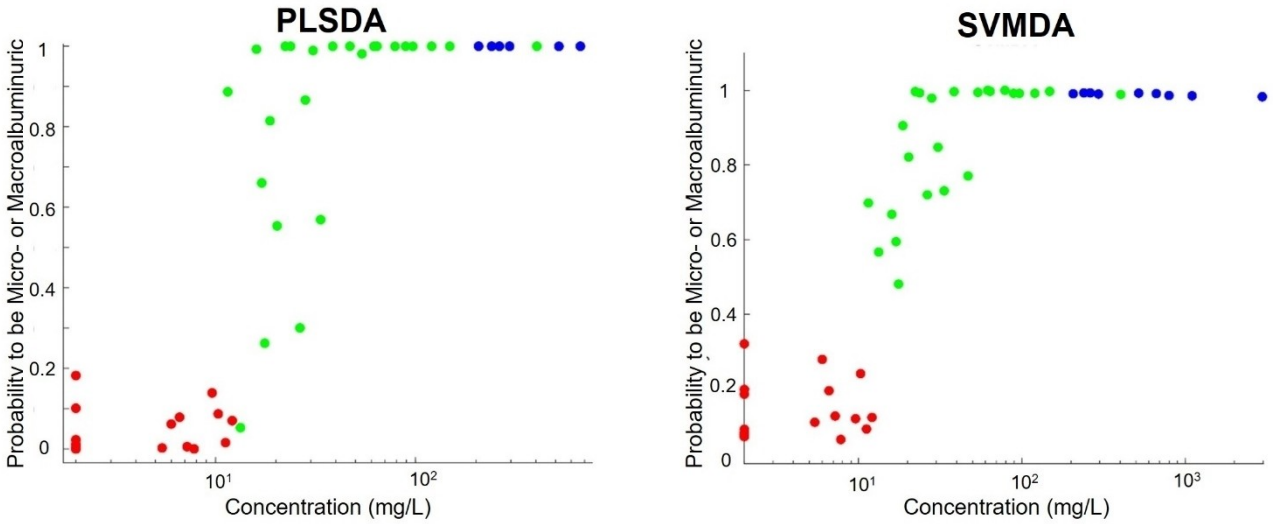

Figure 3. Probability of a sample from the independent test set being classified as diseased (micro- or macroalbuminuria) predicted by the model (PLSDA and SVMDA) versus the known albumin content calculated by the UACR (normoalbuminuric (red), microalbuminuric (green) and macroalbuminuric (blue)).

as 33% and 44%, respectively, for microalbuminuria detection,[33] this test is much more sensitive and specific. In all cases, the AUROC was found to be in the 0.977–1 range (see Supplementary Material 2). Thus, it is a useful screening tool for identifying patients at earlier disease stages than a dipstick, whilst returning faster results than the UACR.

Regression analysis

Next, regression analysis was performed on the data, using the calibration data set to construct the model, whilst the independent test set was used to evaluate how well the model was able to predict the albumin concentration in unseen samples. Regression models were constructed for the entire data range, as well as just for samples with albumin concentrations up to 300 mg/L. This was performed to evaluate how robustly the models can quantify albumin concentrations typically observed in urine samples from diabetic patients, as inclusion of samples with very high albumin values can generate overly optimistic models. Regression models produce root mean square error (RMSE) values, which are in the units of the original data. To allow comparison of the relative error size in this quantification for the cross validation (RMSECV) and the independent prediction (RMSEP) values were normalized to the data range of each sample group. Normalized errors close to zero indicate the magnitude of error in sample concentration prediction is very low, whilst normalized errors close to or above one indicates that the models cannot accurately predict the sample's concentration.

Table 3 shows the pre-processing and spectral regions of the models selected as well as the cross validation and independent testing results. It should be noted that for the classifications the best cross validation results were obtained using normalization, while regression models worked better without it. When a model is computed using the entire data range (0–3000 mg/L) PLS regression (PLSR) modelling produces a more linear fit with higher normalized errors, than the support vector regression (SVR) algorithm, with R^2 (Prediction) values of 0.91 and 0.87, respectively. There is only a small drop in the R^2 value considering the cross validated data, compared to the independent test. This indicates that the models are robust and have not relied on spectral noise during construction. An issue common to both models is the magnitude of error when predicting the albumin concentration of normoalbuminuric samples, with normalized RMSEP values of 1.66 and 0.94 for PLSR and SVR, respectively. This indicates that these models are not suitable for predicting very low concentrations of albumin, and that a separate model should be constructed which only considers the low concentration range as described below. The magnitude of error was more reasonable for microalbuminuric and macroalbuminuric samples, with normalized RMSEP values of 0.22 and 0.19, respectively, for the PLS-R model, and 0.15 and 0.24, respectively for, the SVR model.

When the models are computed using only the samples with lower albumin concentrations (0–300 mg/L) the SVR algorithm is a superior prediction model. This model has an R^2 (CV) of 0.92, with normalized RMSECV of 0.54 and 0.11 for the normo- and microalbuminuric samples, respectively; compared to an R^2 (CV) 0.88 and normalized RMSECV of 0.45 and 0.13, respectively, for the PLSR model. A substantial drop in linearity is observed when the model is evaluated using the independent test set, with R^2 (Prediction) values of 0.77 and 0.71 for the SVR and PLSR, respectively. This indicates that there is a degree of overfitting, meaning that the model relied on noise, rather than characteristic peaks from albumin, when constructing the model. Both the SVR and PLSR algorithms have substantial errors when predicting the concentration of normoalbuminuric samples, with normalized RMSEP values of 0.39 and 0.37, respectively. Conversely, the magnitude of error is much smaller when predicting the concentration of microalbuminuric samples in both models, with RMSEP values of 0.09 and 0.11, respectively. Whilst the linearity of the model did decrease, R^2 (Prediction) values above 0.7 indicate that this model is suitable for approximate quantification of microalbuminuric samples.

It should be noted that the problems on the prediction at the lower range can also be affected by the limitations of the reference method. The instrumentation at the pathology laboratory calculated an albumin concentration of 0 mg/L for 17 samples, indicating that the protein content was below the limit of quantification for ELISA. However, the protein spectra of the extract for these samples (See Supplementary Material 3) shows clear protein features for all the spectra, showing that one of the main limitations of the study is the lack of a highly

Table 3. Parameters of the regression models and figures of merit. NOTE: FD, first derivative; SD, second derivative; MC; Mean Centering; LVs, Number of latent variables selected for the PLSDA model or for the compression performed for the before the SVM-R.

		Parameters	R^2 CV	RMSECV (mg/L)	Normalized RMSECV	R^2 Pred	RMSEP (mg/L)	Normalized RMSEP
PLS-R	NOR	800–1800 & 2770–3140	0.92	35	1.57	0.91	37	1.66
0–3000 mg/L	MIC	SD + MC		60	0.15		88	0.22
	MAC	6 LVs		320	0.11		543	0.19
SVM-R	NOR	800–1800 & 2770–3140	0.94	15	0.67	0.87	21	0.94
0–3000 mg/L	MIC	SD + MC		41	0.10		58	0.15
	MAC	6 LVs		296	0.10		684	0.24
PLS-R	NOR	800–1800	0.88	10	0.45	0.71	8.2	0.37
0–300 mg/L	MIC	SD + MC		53	0.13		44	0.11
		4 LVs						
SVM-R	NOR	800–1800	0.92	12	0.54	0.77	8.7	0.39
0–300 mg/L	MIC	SD + MC		43	0.11		34	0.09
		4 LVs						

accurate reference method. Although these low concentrations maybe not be clinically relevant, the full dynamic range of the proposed technique could only be evaluated with a more sensitive reference method. Previous work has demonstrated that this ultrafiltration technique has an LoD of 6.7 mg/L for albumin,[27] so with refinement, this approach could be used as the new gold standard for albumin quantification in urine.

Passing-Bablok regression[34] was carried for the non-parametric comparison between the results of the proposed methodology based on IR (x) and the reference method (y). Results for all the models considered (See Supplementary material 4) showed slopes and offset values within the ranges [0.85 1.4] and [−10 3.9], respectively. While in most of the cases the slope and offset values indicated that the methods were comparable (Boundaries contained the values of 1 for the slope and 0 for the offset), in some cases there was a small deviation. Furthermore, Bland Altman[35] plots were also computed and shown in supplementary material 4. In general, the differences of the results between both techniques are uniformly distributed, although in some cases high concentrated samples show a large deviation, which skewed the comparison for some of the models.

To summarise, regression analysis has excellent predictive capabilities when considering the entire data range. However, the magnitude of the RMSEP values indicate that the model will struggle with accurate albumin concentration prediction of normoalbuminuric samples. The magnitude of errors decreased when only the lower concentration range was considered, but with this came a substantial decrease in the predictive capability of the model. Considering both the degree of linearity and magnitude of errors, the SVR algorithm appears to be the most appropriate for this dataset.

Predictive capabilities

This test is designed as a screening tool to identify patients with (early stage) albuminuria. Thus, it must be able to handle and appropriately classify samples with unknown urine content. In application to the general population the model will largely encounter urine samples with healthy levels of albumin. To ensure that the model was able to correctly handle and predict the albumin content of healthy samples, the regression model was applied to the cohort of "zero controls". It was anticipated that the model would return albumin concentrations near zero for this sample group. Prediction of the albumin concentration of samples from healthy volunteers involves extrapolation of the models, since these samples have albumin concentrations below that of samples from diabetic patients. PLSR is a linear based regression model, so the effect of extrapolation is minimal, whereas SVR is not linearly based, instead using a radial basis kernel, so extrapolated data can be highly unreliable. Hence, when applying SVR some samples recorded albumin concentrations below zero, which is not possible. The albumin concentration of these samples were manually corrected to zero. PLSR predicted control samples to have a mean and median albumin concentration of 1.5 and 0.8 mg/L, respectively, with inter quartile range (IQR) of 2.1 mg/L, being feasible results for this sample cohort.

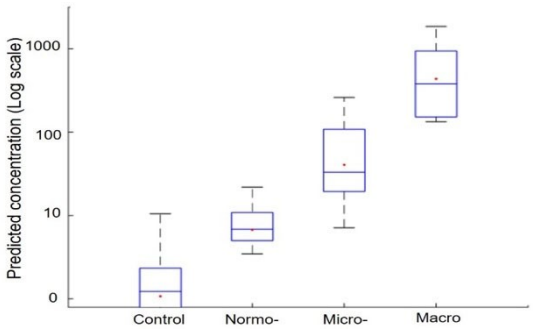

Figure 4. Predicted albumin concentration (mg/L) of urine samples from the independent test set using the PLS-R algorithm. The red dot represents the mean albumin concentration of each sample group.

Due to the abovementioned issues of extrapolation with the SVR algorithm, the predicted albumin concentration values when using the PLS-R algorithm for controls and samples from the validation test set were plotted in a box plot (Figure 4). Tabulation of numerical results is provided in Supplementary Material 5. The mean and median predicted albumin concentrations were approximately in the midpoint of the concentration range for each disease group. As expected, there is significant overlap in predicted albumin concentrations between control and normoalbuminuric samples. There was one normoalbuminuric sample which had a substantially higher predicted concentration than anticipated, indicated by a long whisker on the box plot. This sample was not classed as an outlier, but potentially needs to have its IR spectrum remeasured. In future, the model should be used to predict the albumin content in urine from a cohort of age matched controls without diabetes mellitus, and the necessary adjustments made to the model to account for the presence of other diseases and medication which may influence metabolite excretion in urine. Inclusion of this population in future work will provide an improved indication of how the model will perform when applied to the general population.

Conclusion

The use of ultrafiltration devices for sample pre-treatment increases the concentration of low concentration analytes, like protein, whilst diluting out components like glucose and urea which normally dominate the spectrum. In this paper we demonstrate that albumin concentration prediction is possible using this approach, with an R2 (Prediction) values of 0.939 and normalized RMSEP values below 0.160 for micro- and macro-albuminuric samples when using the SVR algorithm. Applicability of this quantification technique to patients with normoalbuminuria may be limited by the size of the error at low concentrations and limitations with the reference method. More promising, however, is the application of this technique to qualitative purposes. Future work should include an age

matched control group without diabetes mellitus diagnosis to ascertain how the model performs when applied to the general population. Use of an ultrafiltration device enables identification of microalbuminuric urine samples at much higher sensitivity and specificity than can be achieved using a urinary dipstick. The analytical procedure is also substantially simpler and more suited to point of care applications than the UACR.

Experimental Section

Participant selection

This study was approved by the Austin Health Human Research Ethics Committee (Ethics ID: H2013/04937) and the Monash University Human Research Ethics Committee (Ethics ID: 21405). All individuals provided written consent before analysis commenced. Urine samples from patients with diabetes were from males, aged between 42–88, who were diagnosed with diabetes. Urine samples were from a 24-hour collection and were stored at −80 °C until use. UACR values were provided by the pathology department, with albumin content evaluated using ELISA and creatinine content using the enzymatic method. Sample classification (normo-, micro- or macroalbuminuria) was determined by sample's UACR value.[11] Urine samples from healthy volunteers, which served as "zero-controls", were from males, aged 18–50, with no history of diabetes, pre-diabetes, kidney disease, kidney stones or hypertension and who were not currently taking any medications or drugs. Urine samples were collected as a spot urine and stored at −20 °C until use.

Sample preparation

700 μL of sample was drawn into a needleless syringe and dispensed through a 0.2 μM Corning syringe filter with a cellulose acetate membrane. 500 μL of filtrate was transferred into Amicon© centrifuge filter with a pore size of 10 kDa, which had been washed as per the manufacturer's instructions. Samples were centrifuged at 14000 g for 15 minutes and the filtrate discarded. 450 μL of ultrapure water was added to the filter and the sample was spun at 14000 g for 15 minutes and the filtrate discarded. The washing process was repeated three times. After washing the remaining concentrated sample was transferred into a 1 mL Eppendorf tube and stored until use.

ATR FTIR measurements

Spectra were recorded on an Alpha FTIR spectrometer (Bruker Corporation, Billerica, USA) with a diamond crystal ATR accessory. 0.8 μL of concentration sample was pipetted onto the ATR diamond cell and dried under gentle air flow for 2 minutes. Spectra were recorded from 4000 cm^{-1} to 800 cm^{-1} with 40 scans, spectral resolution of 4 cm^{-1} and zero-fill factor of 2. Three technical repeats were recorded for each biological replicate. A background measurement of the clean diamond crystal with 128 scans, spectral resolution 4 cm^{-1} and zero fill factor of 2 was recorded between each biological replicate.

Data pre-processing and analysis

Data visualisation and analysis was carried out in MATLAB 2021b (MathWorks Inc., Natic, USA), using in-house written scripts based on functions available at the statistics and machine learning toolbox from MATLAB and the PLS Toolbox (Eigenvector Research Inc., Manson, USA). PCA was performed on baseline corrected (automatic weighted least squares (AWLS)) and mean centred spectra between 3500–2700 and 1800–800 cm^{-1}.

For modelling, patients were divided into calibration (n = 102) and independent test (n = 52) datasets. The split was performed to create representative datasets of normo-, micro-, and macro-albuminuria. Samples were sorted according to the UACR value, then every third sample was selected for the independent test set, with the remaining samples used for the calibration set.

For the creation and optimization of the models, only the calibration set was used. PLSR and SVR were used to predict the concentration of proteins and PLS discriminant analysis (DA) and SVMDA were used for the discrimination among samples with normo-micro- and macroalbuminuria levels. Models were created using combinations of three regions (2770–3140, 1400–1800 and 800–1400 cm^{-1}) and different pre-processing steps, namely first and second derivative, standard normal variate (SNV) normalization and mean centring). Pre-processing and spectral regions were selected using cross validation (venetian blinds 10 splits), which was also used to optimize the number of latent variables (LVs) in PLSR and PLSDA modelling In short, 49 PLS models combining spectral regions with different pre-processing algorithms were created. For each model, the number of LVs that provided the lowest CV error was considered, providing a list of 49 models containing the corresponding pre-processing, spectral regions, optimal number of LVs and CV error. The model which provided the lowest CV error with a reasonable number of LVs was selected namely model 2. The Root Mean Square Error of cross validation (RMSECV) and the Cross Validation Classification Error was considered as the performance parameters. The selected pre-processing algorithms and spectral regions were also used for SVR and SVMDA. Data was compressed using a PLS algorithm and internal parameters of the SVR and SVMDA were also selected using cross validation (venetian blinds 10 splits).

Then, the models chosen were validated using the independent test set, The root mean square of prediction (RMSEP) was used to evaluate the regressions, while specificity, sensitivity, classification error values and the receiver operating characteristic (ROC) curve were used to assess the performance of the classifications. Passing Bablok regression[34] and Bland-Atman[35] analysis were carried out using Matlab functions obtained from Mathworks repository. Datasets and scripts are available at the zenodo (10.5281/zenodo.7032460) repository.

Acknowledgements

The authors would like to thank Mr Finlay Shanks for instrumental support, Ms Suzanne Golub for assistance with assistance with acquiring patient samples and personal contacts for the supply of urine from healthy adults. Funding: This work was supported by Monash University through the provision of a Monash Institute for Medical Engineering Seed Fund 2020/2021 grant. This work was also supported by grants RYC2019-026556-I and RPID2020-119326RA-I0 funded by MCIN/AEI/ 10.13039/501100011033. The funding sources had no involvement in the conduct of the research or preparation of the article. Open Access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.7032460, reference number 7032460.

Keywords: diabetic kidney disease · IR spectroscopy · urinalysis · ultrafiltration · vibrational spectroscopy

[1] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. N. Chan, J. C. Mbanya, M. E. Pavkov, A. Ramachandaran, S. H. Wild, S. James, W. H. Herman, P. Zhang, C. Bommer, S. Kuo, E. J. Boyko, D. J. Magliano, *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119.
[2] WHO, A. Loke, "Diabetes", can be found under https://www.who.int/news-room/fact-sheets/detail/diabetes, **2021** (accessed 20/01/22).
[3] Z. Xu, D. Yu, X. Yin, F. Zheng, H. Li, *Oncotarget* **2017**, *8*, 44434–44439.
[4] H. Kaneto, N. Katakami, M. Matsuhisa, T. Matsuoka, *Mediators Inflammation* **2010**, *2010*, 453892.
[5] Z. Irshad, M. Xue, A. Ashour, J. R. Larkin, P. J. Thornalley, N. Rabbani, *Sci. Rep.* **2019**, *9*, 7889.
[6] K. Reidy, H. M. Kang, T. Hostetter, K. Susztak, *J. Clin. Invest.* **2014**, *124*, 2333–2340.
[7] G. Pizzino, N. Irrera, M. Cucinotta, G. Pallio, F. Mannino, V. Arcoraci, F. Squadrito, D. Altavilla, A. Bitto, *Oxid. Med. Cell. Longevity* **2017**, *2017*, 8416763–8416763.
[8] Y. Deng, N. Li, Y. Wu, M. Wang, S. Yang, Y. Zheng, X. Deng, D. Xiang, Y. Zhu, P. Xu, Z. Zhai, D. Zhang, Z. Dai, J. Gao, *Front. Endocrinol.* **2021**, *12*, 675320.
[9] O. Pleniceanu, G. Twig, D. Tzur, N. Gruber, M. Stern-Zimmer, A. Afek, T. Erlich, L. Keinan-Boker, K. Skorecki, R. Calderon-Margalit, A. Vivante, *Pediatr. Nephrol.* **2021**, *36*, 333–340.
[10] *Chronic Kidney Disease (CKD) Management in Primary Care* 4ed., Kidney Health Australia, **2020**.
[11] H. J. L. Heerspink, A. H. Brantsma, D. de Zeeuw, S. J. L. Bakker, P. E. de Jong, R. T. Gansevoort, *Am. J. Epidemiol.* **2008**, *168*, 897–905.
[12] D. W. Johnson, G. R. Jones, T. H. Mathew, M. J. Ludlow, S. J. Chadban, T. Usherwood, K. Polkinghorne, S. Colagiuri, G. Jerums, R. Macisaac, H. Martin, *Med. J. Aust.* **2012**, *197*, 224–225.
[13] T. Ninomiya, V. Perkovic, B. E. de Galan, S. Zoungas, A. Pillai, M. Jardine, A. Patel, A. Cass, B. Neal, N. Poulter, C. Mogensen, M. Cooper, M. Marre, B. Williams, P. Hamet, G. Mancia, M. Woodward, S. Macmahon, J. Chalmers, *J. Am. Soc. Nephrol.* **2009**, *20*, 1813–1821.
[14] S. I. Hallan, E. Ritz, S. Lydersen, S. Romundstad, K. Kvenild, S. R. Orth, *J. Am. Soc. Nephrol.* **2009**, *20*, 1069–1077.
[15] 'Uristix: Tests for Protein and Glucose in Urine' Instructions for use by Siemens Healthcare Diagnostics Inc.
[16] J. R. Mejia, J. E. Fernandez-Chinguel, G. Dolores-Maldonado, N. Becerra-Chauca, S. Goicochea-Lugo, P. Herrera-Añazco, J. H. Zafra-Tanaka, A. Taype-Rondan, *Heliyon* **2021**, *7*, e08253.
[17] D. R. Lloyd, E. J. Hindle, J. Marples, J. A. Gatt, *Ann. Clin. Biochem.* **1987**, *24*, 209–210.
[18] M. Diem, in *Modern Vibrational Spectroscopy and Micro-Spectroscopy*, **2015**, p. 71–102.
[19] A. S. Douglas, F. J. Holler, R. C. Stanley, in *Principles of Instrumental Analysis*, Cengage, Mason, OH, United States, **2016**, p. 389–436.
[20] H. M. Heise, G. Voigt, P. Lampen, L. Küpper, S. Rudloff, G. Werner, *Appl. Spectrosc.* **2001**, *55*, 434–443.
[21] G. Hoşafçi, O. Klein, G. Oremek, W. Mäntele, *Anal. Bioanal. Chem.* **2007**, *387*, 1815–1822.
[22] K. V. Oliver, A. Vilasi, A. Maréchal, S. H. Moochhala, R. J. Unwin, P. R. Rich, *Sci. Rep.* **2016**, *6*, 34737.
[23] N. Sarigul, İ. Kurultak, A. Uslu Gökceoğlu, F. Korkmaz, *J. Biophotonics* **2021**, *14*, e202100009.
[24] A. Kincses, K. C. Sourris, M. Mohan, P. Kantharidis, K. Jandeleit-Dahm, B. R. Wood, *Microchem. J.* **2022**, *181*, 107641.
[25] F. K. Algethami, S. M. Eid, K. M. Kelani, M. R. Elghobashy, M. K. Abd El-Rahman, *RSC Adv.* **2020**, *10*, 7146–7154.
[26] M. Yu, P. Rich, L. Foreman, J. Smith, M. Yu, A. Tanna, V. Dibbur, R. Unwin, F. W. K. Tam, *Sci. Rep.* **2017**, *7*, 4601.
[27] D. Perez-Guaita, Z. Richardson, P. Heraud, B. R. Wood, *Anal. Chem.* **2020**, *92*, 2409–2416.
[28] R. A. Shaw, S. Low-Ying, M. Leroux, H. H. Mantsch, *Clin. Chem.* **2000**, *46*, 1493–1495.
[29] K. V. Oliver, F. Matjiu, C. Davey, S. Moochhala, R. J. Unwin, P. R. Rich, *Proc. SPIE* **2015**, *9332*, 10.1117/12.2078971.
[30] F. Bonnier, H. Blasco, C. Wasselet, G. Brachet, R. Respaud, L. F. C. S. Carvalho, D. Bertrand, M. J. Baker, H. J. Byrne, I. Chourpa, *Analyst* **2017**, *142*, 1285–1298.
[31] I. Elsohaby, J. T. McClure, C. B. Riley, J. Bryanton, K. Bigsby, R. A. Shaw, *J. Pharm. Biomed. Anal.* **2018**, *150*, 413–419.
[32] Merck, "Amicon® Ultra 0.5 mL Centrifugal Filters", can be found under https://www.merckmillipore.com/AU/en/life-science-research/protein-sample-preparation/protein-concentration/amicon-ultra-centrifugal-filters/0-5-ml/ytSb.qB.gZIAAAFBQHplvyxn,nav**2022** (accessed 25/08/2022).
[33] A. Nagrebetsky, J. Jin, R. Stevens, T. James, A. Adler, P. Park, A. Craven, B. Shine, A. Farmer, *Fam. Pract.* **2013**, *30*, 142–152.
[34] L. Bilić-Zulle, *Biochem Med (Zagreb)* **2011**, *21*, 49–52.
[35] D. Giavarina, *Biochem Med (Zagreb)* **2015**, *25*, 141–151.

Manuscript received: October 11, 2022
Revised manuscript received: December 21, 2022
Accepted manuscript online: December 27, 2022
Version of record online: January 16, 2023