NMR-based metabolomic profiling identifies inflammation and muscle-related metabolites as predictors of incident type 2 diabetes mellitus beyond glucose: the Di@bet.es study

Short running title: H-NMR metabolomics of diabetes beyond glucose

Enrique Ozcariz[1]*, Montse Guardiola PhD[2,3,4]*, Núria Amigó PhD[1,2,3,7,8], Gemma Rojo-Martínez PhD[2,5,6], Sergio Valdés PhD[2,5,6], Pere Rehues[2,3,4], Lluís Masana MD,PhD[2,3,4], Josep Ribalta PhD[2,3,4].

1. Biosfer Teslab, Plaça del Prim 10, 2on 5a, 43201 Reus, Spain.

2. CIBER de Diabetes y Enfermedades Metabólicas Asociadas, Instituto de Salud Carlos III, Madrid, Spain.

3. Institut d'Investigació Sanitària Pere Virgili (IISPV), Reus, Spain.

4. Universitat Rovira i Virgili, Departament de Medicina i Cirurgia, Unitat de Recerca en Lípids i Arteriosclerosi, Reus, Spain.

5. UGC Endocrinología y Nutrición. Hospital Regional Universitario de Málaga, Málaga, Spain.

6. Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND, Málaga, Spain,

7. Universitat Rovira i Virgili, Departament de Ciències Mèdiques Bàsiques, Reus, Spain.

8. Universitat Rovira i Virgili, Metabolomics Platform, Reus Spain.

Key words: Metabolomics, NMR, Machine Learning, Lipoproteins, Glycoproteins, Glucose

Number of words:3978

Number of tables: 1

Number of figures: 3

ABSTRACT

Objective:

The aim of this study was to combine nuclear magnetic resonance-based metabolomics and machine learning to find a glucose-independent molecular signature associated with future type 2 diabetes mellitus development in a subgroup of individuals from the Di@bet.es study.

Research design and methods:

The study group included 145 individuals developing type 2 diabetes mellitus during the 8-year follow-up, 145 individuals matched by age, sex, BMI not developing diabetes during the follow-up but with equal glucose concentrations than those who did and 145 controls matched by age and sex. A nuclear magnetic resonance-based metabolomic analysis of serum samples was performed. The lipoprotein and glycoprotein profiles and 15 low molecular weight metabolites were obtained. Different machine learning-based models were trained.

Results:

Logistic regression proved to perform the best classification between individuals developing type 2 diabetes during the follow-up and the glucose-matched ones. The area under the curve was 0.628 and its 95% confidence interval was 0.510 – 0.746. Glycoprotein-related variables, creatinine, creatine and small HDL particles presented statistically significant coefficients in the model. Additionally,

the Johnson-Neyman intervals of the interaction of Glyc A and Glyc B were also significant.

Conclusions:

The model highlighted a relevant contribution of inflammation (glycosylation pattern and HDL) and muscle (creatinine and creatine) in the development of type 2 diabetes mellitus as independent factors of hyperglycemia.

INTRODUCTION

Over the past few decades, the prevalence of diabetes mellitus in developed and developing countries has risen substantially, making diabetes a key health public problem (1). The most common is type 2, which is present in 95% of the people affected by this pathology. Type 2 diabetes is characterized by relative insulin deficiency caused by pancreatic β-cell dysfunction and insulin resistance in target organs (2).

The diverse clinical and pathophysiological features of type 2 diabetes contribute to the difficulty for prevention or remission strategies that are universally effective for every patient. Hyperglycemia is the main clinical sign of diabetes (3), (4). Thus, glucose level monitoring via glycated hemoglobin is currently the most popular screening strategy (4). Nonetheless, elevated glucose levels indicate that diabetes is already active. However, not all subjects with mild hyperglycemia or prediabetes will develop clinical diabetes, so new strategies are needed to look for biomarkers associated with prediabetes to diabetes progression, but independent of glycemia.

Indeed, the manifestation of insulin resistance in the prediabetic state is the feature on which early type 2 diabetes prediction is currently based (5). However, many other metabolic imbalances are already occurring in prediabetes before any clinical manifestation (4), (6). A better understanding of these imbalances could help in the earlier detection of biomarkers or metabolic patterns associated with type 2 diabetes, which would result in a more precise type 2 diabetes development risk assessment.

Metabolomics is an emerging approach that is capable of providing measurements of all, or a large number of metabolites in cells, tissues or biological fluids (7). Thus, metabolomics provides the integration of genomic, epigenetic, transcriptomic and proteomic variation, and is also responsive to environmental factors (8). Due to this fact, metabolomics provides direct information about the physiological state of an individual (9). Strategies for performing metabolomic analyses vary greatly in terms of information acquisition (physical-chemical analysis of the sample) and processing.

The current worldwide progress in computational models and statistical analysis has had a great impact in the field of pathology prediction and biomarker discovery (5). The combination of machine learning (ML) techniques with the large amount of biological data offered by metabolomics is a powerful approach to identify not only new biomarkers but also to recognize metabolic patterns differentially associated with pathology (5), (10). Although several ML and metabolomic combination-based approaches have been performed for type 2 diabetes mellitus prediction (5), (6), (11), most of them are based on pathophysiological changes common in diabetes, which are usually triggered by an already developed metabolic disorder associated with hyperglycemia. Therefore, these approaches can hardly identify patients at risk of developing metabolic disorders in the future beyond glucose. The aim of this study was to combine ML and proton nuclear magnetic resonance (H-NMR) metabolomics to identify a glucose independent molecular signature associated with future type 2 diabetes mellitus development in a subgroup of individuals from the Di@bet.es study.

MATERIALS & METHODS

Population. A case-control design nested in a population-based cohort (Di@bet.es study) was followed in this study. The Di@bet.es Study, the first national study in Spain to examine the prevalence of diabetes and impaired glucose regulation is composed of 4700 individuals with metabolomic data available (43% men), with ages ranging from 18 to 93 years old, of which 2181 participated in the follow-up 8 years later. All subjects who developed diabetes during 7.5±0.6 follow-up (n=145) were selected as cases (incident type 2 diabetes group). A set of controls matched by age and sex (reference group (n=145)) and another set matched by age, sex, fasting blood glucose and BMI (glucose-matched group (n=145)) were constructed with subjects who did not develop diabetes during follow-up using greedy nearest neighbor matching (supplemental figure S1).

H-NMR analysis. Before H-NMR analysis, 200 µl of fasting serum collected at the baseline study was diluted with 50 µl of deuterated water and 300 µl of 50 mM phosphate buffer solution (PBS) at pH 7.4. H-NMR spectra were recorded at 306 K on a Bruker Avance III 600 spectrometer operating at a proton frequency of 600.20 MHz.

Lipoprotein analysis. Lipoprotein analysis was performed by using Liposcale® Test, a novel advanced lipoprotein test based on 2D diffusion-ordered H-NMR spectroscopy (12). The methyl signal was deconvoluted by using 9 Lorentzian functions to determine the lipid concentration of the large, medium and small subclasses of the main lipoprotein classes (VLDL, LDL and HDL), and their size associated diffusion coefficients. Then, we combined the lipid concentration with the associated particle volume to quantify the number of particles required to

transport the measured lipid concentration of each lipoprotein subclass. Finally, weighted average VLDL, LDL and HDL particle sizes were calculated from various subclass concentrations by summing the known diameter of each subclass multiplied by its relative percentage of subclass particle number. The variation coefficients for the particle number were between 2% and 4%, and for the particle sizes, they were lower than 0.3%.

Glycoprotein analysis. The region of the H-NMR spectrum where the glycoproteins resonate (2.15-1.90 ppm) using several analytical functions according to a previously published procedure was analyzed (13). For each function, the total area (proportional to concentration) and signal shape (height to bandwidth H/W ratio) were determined. The area of Glyc A provided the concentration of acetyl groups of protein-bound N-acetylglucosamine and N-acetylgalactosamine, and the area of Glyc B provided those of N-acetylneuraminic acid. The Glyc F area arises from the concentration of the acetyl groups of N- acetylglucosamine, N-acetylgalactosamine and N-acetylneuraminic acid unbound to proteins (free fraction) (14). H/W ratios of Glyc A and Glyc B were also reported, a parameter associated with the aggregation state of the sugar-protein bonds. The variation coefficients for the glycoproteins were lower than 3%.

Low molecular weight metabolite (LMWM) analysis. Fully-automated software -developed by the company Biosfer Teslab- was used to perform the deconvolution of the signals associated with 15 different LMWMs as previously reported (15). For each metabolite, the total area was quantified and normalized to obtain the concentration of each metabolite. The variation coefficients for the LMWMs were between 6% and 18%.

Statistical analysis. Non-parametric Kruskall-Wallis test was performed to compare all the numeric variables between the reference group, the glucose-matched group and the incident type 2 diabetes group. Chi-squared test was used to perform the comparison of sex, due to its categorical nature. The Benjamini & Hochberg method was used to control the false discovery rate to an alpha confidence level of 5%. Variable selection was performed before multivariate analysis, to avoid overfitting. Covariance and correlation matrices were created to evaluate the relationship among predictors. The lipoprotein profile was summarized in the triglyceride to cholesterol ratios of each lipoprotein class -VLDL, IDL, LDL and HDL- and in the small-to-total particle ratio of VLDL, LDL and HDL. Due to their high correlation, valine, leucine and isoleucine were summed and included in a unique variable: branched chain amino acids (BCAAs). A genetic algorithm (GA) was applied to supplement the variable selection process (16). After that, machine learning (ML)-based algorithms were created to look for the best multivariate predictive model. Data was scaled to z-scores before modeling. All the models included in this study were based on classification algorithms. The trained models were: logistic regression (LR), extreme gradient boosting (EGB), random forest (RF) and support vector machine (SVM). The training process was performed with 75% of the samples. 10-fold cross-validation was conducted during the training to increase the robustness of the models. Accuracy was the leading metric of the training process. To avoid overfitting, the validation of each model was performed in an independent sample set containing the 25% of the samples not included in the training. LR proved to be the best model. The weights of interactions between predictors in LR were evaluated and for those presenting the highest, Johsonn-

Neyman intervals were built (17). Another LR was trained and tested, including only the interactions that had statistically significant intervals. A new model including those interactions was created. All of the models were created with caret package using version R 4.1 (18).

RESULTS

Comparison of H-NMR-based metabolomic profiles between the incident type 2 diabetes, glucose-matched and reference groups.

Univariate comparison was performed among the reference group, the glucose-matched group and the incident type 2 diabetes group (Table 1). BMI showed strong differences between the two study groups and the reference group, indicating a higher prevalence of obesity in the incident type 2 diabetes group and in the glucose-matched group. Both groups presented higher concentrations of all glycoprotein related variables in comparison with the reference group. However, only the incident type 2 diabetes group showed statistically significant increases relative to the reference group in triglycerides (TG), small VLDL-P (S-VLDL-P), small LDL-P (S-LDL-P), VLDL particle size (VLDL-Z) and LDL particle size (LDL-Z). Additionally, several LMWMs presented statistically significant differences among the three groups. As expected, glucose was increased in the incident type 2 diabetes and glucose-matched groups. Lactate and alanine presented higher concentrations in these two groups in comparison with the reference group, with the observed differences being stronger in the incident type 2 diabetes group. These findings highlight alterations in carbohydrate metabolism. Isoleucine and leucine presented higher concentrations in the matched and incident type 2 diabetes groups, than in the reference group. Nonetheless, only in the incident type 2 diabetes group was valine significantly increased.

Incident type 2 diabetes mellitus associated H-NMR based metabolomic signature.

After univariate analysis, multivariate machine learning (ML)-based prediction models were constructed to find a type 2 diabetes mellitus-associated metabolomic profile beyond glucose.

GA was applied before modeling, to perform variable selection and avoid overfitting. This mathematical algorithm was only performed for the glucose-matched and incident type 2 diabetes groups, because posterior classification models will be applied to them. Most of the metabolites presenting significant differences in the previous analysis were selected by the GA with a frequency higher than 50%. Those predictors that presented a frequency higher than 75% were selected. These were: small to total particle ratios of LDL (sLDLP), small to total particle ratios of HDL (sHDLP), TG to cholesterol ratios of HDL (RatioHDL), LDL (RatioLDL) and IDL (RatioIDL), lactate, Glyc A, Glyc B, Glyc F, creatinine, creatine, acetone, histidine, alanine and glucose.

Once the selection was performed, the data were randomly divided into a training set (75%) and a test or validation set (25%). Only matched and incident type 2 diabetes groups were included in the model training and testing. Several ML-based models were constructed with the selected variables as predictors. Two tree-based models were constructed: RF and an EGB. An area under the curve (AUC) of 0.509 was obtained for the first model, and an AUC of 0.526 was obtained for the second model (Fig. 1). An SVM was also built, and an AUC of 0.505 was obtained (Fig. 1). Finally, a LR was performed, obtaining an AUC of 0.628 (Fig. 1). Only in this case, was the AUC value and its 95% confidence interval higher than 0.5.

LR proved to be the best model, so the direction of the coefficients and the variable importance were analyzed (Fig. 2). Six variables were found to be

significant predictors of incident type 2 diabetes. Glyc A was the strongest predictor of diabetes mellitus, whereas Glyc B, Glyc F and sHDLP were significantly associated with a lower risk of developing type 2 diabetes mellitus.Moreover, creatinine and creatine were significantly associated with lower risk of developing type 2 diabetes. The analysis also rendered several variables suggestive of being associated with the risk of developing type 2 diabetes. Acetone, lactate, histidine, RatioLDL and sLDL towards an increased risk and alanine and RatioHDL towards a decreased risk of developing type 2 diabetes.

After the analysis of the model, possible interactions among the predictors were evaluated. Johnson-Neyman intervals were calculated for those presenting the highest weight. Only the interaction between Glyc A and Glyc B was statistically significant (Fig. 3). The Glyc A and Glyc B interaction Johnson-Neyman plot showed that both predictors had inverse behaviors. This also reflected that at high Glyc B levels, the effect of Glyc A on diabetes development was no longer significant. This interaction was used to construct a new LR, but no improvement in its predictive power was observed in comparison with the previous one.

DISCUSSION

The goal of this study was to find a type 2 diabetes-associated early metabolomic profile beyond glucose, to advance the elucidation of the molecular mechanisms underlying type 2 diabetes development. To achieve that, the Di@bet.es cohort was used, comprising 4700 individuals -with metabolomic data available-, of whom 145 individuals developed type 2 diabetes after 8 years. A matching process was performed to create an age, sex, BMI and glucose adjusted group not developing type 2 diabetes during the follow-up. Our results indicate that variables related to inflammation (glycated proteins and HDL) and to muscle tissue (creatinine and creatine) are significant predictors of type 2 diabetes development regardless of glucose concentrations.

These two groups were compared to a reference normoglycemic group, paired by age and sex. As expected, differences were observed between the two study groups and the reference group. A higher prevalence of obesity was observed in the incident type 2 diabetes group and in the matched group. Indeed, obesity is one of the most relevant risk factors for the development of type 2 diabetes. Additionally, all glycoprotein-related variables were also increased in the two study groups. Glycosylation plays a key role in inflammation (13), (14), so increased levels of these variables indicate a higher systemic inflammatory state in the study groups than in the reference group. Numerous associations between higher protein glycosylation and type 2 diabetes mellitus have been described by several studies (5), (19), (20), (21), (22), (23).

Regarding the lipid panel and lipoprotein profile at basal point, only individuals with type 2 diabetes in the follow-up (incident type 2 diabetes group) presented statistically significant differences in comparison with the reference group.

Subjects who will develop type 2 diabetes mellitus, showed higher TG levels at basal point, as well as smaller VLDL and LDL particles, characteristic of insulin resistance (24). Differences were also observed for alanine, BCAA and lactate between the study groups and the reference group, being stronger in the case of individuals included in the incident type 2 diabetes group. These metabolites are associated with insulin resistance and prediabetes, (25), (26). Therefore, the observation of stronger differences between the incident type 2 diabetes group and the reference group for these metabolites, may reflect a more advanced stage of insulin resistance in this group than in the glucose-matched group. This hypothesis is consistent with statistically significant differences observed in TG and lipoproteins, only between individuals in the incident type 2 diabetes group and subjects from the reference group. Thus, the obtained results suggest an advanced state of insulin resistance in individuals from the incident type 2 diabetes group than in glucose-matched individuals.

To search for a glucose-independent metabolomic profile associated with type 2 diabetes mellitus development, a multivariate classification analysis was performed between individuals in the incident type 2 diabetes group and their best matched individuals for age, sex, BMI and glucose. A previous variable selection was performed by GA.

Most of the variables highlighted in the univariate analysis were selected by the algorithm. For example, sHDLP and sLDLP suggested an important role of small LDL and HDL particles in hyperglycemia (27). Indeed, LDL-P and HDL-P have been found to be robust biomarkers of the cardiovascular disease (CVD)-risk associated with diabetes mellitus (28). RatioHDL, RatioLDL and RatioIDL also presented a high frequency of selection in the genetic algorithm.

Hypertriglyceridemia has been closely linked with diabetes mellitus (29). All glycoprotein-related H-NMR signals were also selected, indicating that inflammation might be an important factor in type 2 diabetes (19). Lactate, alanine and glucose, which were also selected, are the three main metabolites involved in carbohydrate metabolism. Taken together, these results reaffirm the involvement of carbohydrate metabolism suggested by the results of the univariate comparisons. Creatine and creatinine presented a particularly high importance in the genetic algorithm. Both metabolites are mainly found in skeletal muscle and kidneys. Therefore, their disturbances are commonly associated with muscular and renal alterations (30). The last selected predictor was histidine, which has been found to be related to insulin resistance (31).

Creatinine and creatine were found to be important variables in the LR. Lower levels of serum creatinine seem to be associated with a higher risk of developing type 2 diabetes. The serum creatinine concentration accurately reflects skeletal muscle mass, under stable kidney function (32). Thus, lower creatinine and creatine levels may be indicative of a reduced skeletal muscle mass, -also known as sarcopenia-, in individuals developing diabetes in comparison with matched subjects not developing it. Skeletal muscle is the main insulin-targeted tissue (33), so reduced skeletal muscle mass could result in diminished insulin-mediated glucose disposal (34), (35). Glomerular hyperfiltration, can also play a role in the observed lower creatinine serum levels. This phenomenon is recognized as an early renal alteration in subjects with diabetes and hypertension and may accelerate renal function decline in longer-standing diabetes (36), (37), (38). In our study, creatinine and creatine are similarly associated with the incidence of diabetes, with creatine being a clear indicator of muscle mass and

not of kidney damage. Although the exact causes of lower creatinine and creatine levels in individuals developing type 2 diabetes remain unclear, the obtained results suggest a possible role of sarcopenia in type 2 diabetes development in individuals with hyperglycemia. This association was also suggested in a recent study (39).

Glycosylation related variables also presented high importance in the LR. Glyc A arising from the presence of N-acetylglucosamine and N-acetylgalactosamine presented a positive association with type 2 diabetes development. Previous studies reported an association between Glyc A and a higher risk of type 2 diabetes development. Close links have been reported between O-GlcNAc glycosylation and the cellular response to insulin (20),(26). Indeed, increased O-GlcNAc glycosylation of plasmatic proteins can result in enhanced β-cell death (21). However, the most predominant glycosylation in circulating proteins is N-glycosylation (20). Increased complexity in the N-glycome has been previously associated with type 2 diabetes, possibly as a consequence of an altered flux of glucose through the hexosamine pathway (22). This pathway produces uridine diphosphate-N-acetylglucosamine, which is the main substrate for N-linked glycosylation (5). Higher branching of N-glycans in plasmatic proteins implies a higher number of N-acetylglucosamine residues present in plasmatic proteins (22). This could be the cause of the observed positive association between Glyc A and type 2 diabetes development. Interestingly, Glyc B presented the opposite tendency as Glyc A in the LR. Both study groups showed increased sialylation, as observed in previous studies (22),(23). However, they took opposite directions in the model created to distinguish these two groups. The analysis of the interaction between these two variables showed that the glycosylation based on

N-acetylglucosamine and N-acetylgalactosamine ceased to have relevance in the model when Glyc B adopted high values. This effect may reflect an early differential glycosylation pattern between the two study groups. Moreover, mass spectrometry (MS)-based glycomic studies have demonstrated the relevance of the different behaviors of different neuraminic acid residues in type 2 diabetes development (23). A limitation of this study could be that H-NMR is not sensitive enough to distinguish between different antennary structures, but it could be possible that the structurally different glycans containing neuraminic acid contribute differentially to Glyc B. Therefore, although glucose-matched and incident type 2 diabetes groups presented high levels of Glyc B, the sialylation pattern of both may not be the same. Further investigation is needed to better understand the role of glycosylation in type 2 diabetes development.

A strength of our study is the use of H-NMR to perform the metabolomic study. The two main analytical platforms used in metabolic profiling include: NMR and MS. Although MS has a higher sensitivity, NMR spectroscopy offers several advantages. First, it is a nondestructive method that requires minimum sample processing to obtain the exact quantification of the more abundant metabolites present in diverse biological matrices (40). NMR is also advantageous for the identification and quantification of compounds that are difficult to ionize or that have identical masses (40). Additionally, NMR spectroscopy is nonbiased, fast, exceptionally reproducible and highly automatable (14). Finally, the concentrations of several LMWMs and macromolecular complexes, such as lipoprotein subclasses or glycoproteins, can be measured in the same analysis (14).

The main strength of our work is the matched design of cases and controls for the main variables associated with the incidence of diabetes, including glycemia, which allowed us to assess the effect of new variables beyond glucose. The simultaneous study of lipoproteins, glycoproteins and LMWM as a unique metabolomic profile permitted us to evaluate the effects of the different variables on the incidence of type 2 diabetes, independently from glucose. An additional strong point of our study is the fact that all the cases and all the controls were selected from a representative cohort of the general population.

Some limitations must also be pointed out. Firstly, there was not a large amount of new cases of diabetes in follow-up. For this reason, other relevant biomarkers showing more moderate associations with type 2 diabetes development might have not been detected in this study. Furthermore, although the constructed models highlight the relevance of some biomarkers of pathophysiological mechanisms that are acting synergistically with the better-known ones (insulin resistance or obesity, among others) and that could guide the search for new treatment protocols (more focused on inflammation or sarcopenia), its clinical utility is very limited. Finally, we tried to palliate the lack of an independent validation cohort using two different and independent sets of samples to train the models and to validate them.

CONCLUSION

In conclusion, this study highlights the contribution of inflammation (glycosylation pattern and HDL) and muscle metabolism (creatinine and creatine) in the development of type 2 diabetes mellitus as independent factors of hyperglycemia.

Additional studies must be performed to better understand the implications of these biomarkers in the molecular mechanisms underlying diabetes mellitus.

This study also shows that H-NMR is a powerful tool to characterize an early diabetes-mellitus-associated metabolomic profile, going beyond traditional biomarkers and therefore, leading to a better understanding of diabetes mellitus etiopathology.

ARTICLE HIGHLIGHTS

- A more advanced stage of insulin resistance was observed in a group of incident type 2 diabetes subjects, in comparison with a glucose-matched group.
- Individuals in type 2 diabetes group presented smaller LDL and HDL particles than those in glucose-matched group.
- Creatinine and creatine were found to be very relevant predictors of type 2 diabetes mellitus, suggesting an important role of muscle metabolism in its development.
- A differential glycosylation pattern was observed in type 2 diabetes and glucose-matched groups, highlighting the importance of inflammation in type 2 diabetes mellitus development.

ACKNOWLEDGMENTS

The Di@bet.es project is a collaborative study with various phases and subprojects in which a large number of researchers and technicians have collaborated, to whom we are indebted. We are especially grateful to the

Steering Committee of the study together with all collaborators who have made it possible (https://www.sediabetes.org/cientifico-y asistencial/investigacion/proyectos-de-investigacion/estudio-dibet-es/, accessed on 27 June 2022). CIBERDEM Biorepository (IDIBAPS Biobank, Barcelona, Spain) supplied the samples used.

This work was supported by the Spanish Ministerio de Economía y Competitividad (PI21/01294; PI16/00507), Fondo Europeo de Desarrollo Regional (FEDER), and CIBERDEM (CIBER de Diabetes y Enfermedades Metabólicas Asociadas), which are initiatives of ISCIII (Instituto de Salud Carlos III); and by the Cerca Programme, Generalitat de Catalunya. P. R. is a recipient of a predoctoral fellowship from the Ministerio de Universidades (grant number FPU19/04610).

N.A. is stock owner of Biosfer Teslab and has a patent on the method for lipoprotein profiling described in the present manuscript.

J.R., M.G., L.M. and G.R. contributed to the conception and design of the work. E.O., N.A., S.V. and P.R. contributed to the acquisition of the data. E.O. and M.G. performed the analysis and the interpretation of the data. E.O. wrote the draft of the manuscript and all authors revisited it critically, approved and agreed with the final version of the manuscript. E.O. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES

1. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th edn. ed.). Brussels, Belgium.
2. S. Chatterjee, K. Khunti, M,J.Davies. Type 2 diabetes. Lancet. 2017;389(10085):2239–51.
3. M.A. Nauck, J. Wefers, J.J. Meier. Treatment of type 2 diabetes: challenges, hopes, and anticipated successes. Lancet Diabetes Endocrinol. 2021;9(8):525–44.
4. C. Luís, P. Baylina, R. Soares, R. Fernandes. Metabolic dysfunction biomarkers as predictors of early diabetes. Biomolecules. 2021;11(11):1589.
5. A. Cvetko, M. Mangino, M. Tijardović, D. Kifer, M. Falchi, T. Keser, M. Perola, T.D. Spector, G. Lauc, C. Menni, O. Gornik. Plasma N-glycome shows continuous deterioration as the diagnosis of insulin resistance approaches. BMJ Open Diabetes Res Care. 2021;9(1):1–9.
6. Morze J, Wittenbecher C, Schwingshackl L, Danielewicz A, Rynkiewicz A, Hu FB, Guasch-Ferré M. Metabolomics and type 2 diabetes risk: an updated systematic review and meta-analysis of prospective cohort studies. Diabetes Care. 2022;45(4):1013-1024.
7. R.K. Azad, V. Shulaev. Metabolomics technology and bioinformatics for precision medicine. Brief Bioinform. 2019;20(6):1957–71.
8. O. Akinkuolie, D. Aruna, J.E. Buring, P.M. Ridker, and S. Mora. Novel Protein Glycan Side-Chain Biomarker and Risk of Incident Type 2 Diabetes. Physiol Behav. 2017;176(3):139–48.
9. G. Satheesh, S. Ramachandran, A. Jaleel. Metabolomics-Based Prospective Studies and Prediction of Type 2 Diabetes Mellitus Risks. Metab Syndr Relat Disord. 2020;18(1):1–9.
10. FL. Dias-Audibert, LC. Navarro, DN. de Oliveira, J. Delafiori, C.F.O.R. Melo, T.M. Guerreiro, R. F. Troncon, D.L. Petenuci, M.A.E. Watanabe, L. A. Velloso, A.R. Rocha, R.R. Catharino. Combining Machine Learning and Metabolomics to Identify Weight Gain Biomarkers. Front Bioeng Biotechnol. 2020;8(6):1–11.
11. Guasch-Ferré M, Hruby A, Toledo E, Clish CB, Martínez-González MA, Salas-Salvadó J, Hu FB. Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. Diabetes Care. 2016;39(5):833-46.
12. R. Mallol, N. Amigó, M.A. Rodríguez, M. Heras, M. Vinaixa, N. Plana, E. Rock, J. Ribalta, O. Yanes, L. Masana, X. Correig. Liposcale: A novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. J Lipid Res. 2015;56(3):737–46.

13. R. Fuertes-Martín, D. Taverner, J,C. Vallvé, S. Paredes, L. Masana, X. Correig Blanchar, N. Amigó. Characterization of 1H NMR Plasma Glycoproteins as a New Strategy to Identify Inflammatory Patterns in Rheumatoid Arthritis. J Proteome Res. 2018;17(11):3730–9.

14. R. Fuertes-Martín, X. Correig, J. Vallvé, N. Amigó. Human Serum / Plasma Glycoprotein Analysis by 1 H-NMR , an Emerging Method of Inflammatory Assessment. 2020;9(354):1–31.

15. J. Miranda, R.V. Simões, C. Paules, D. Cañueto, M.A. Pardo-Cea, M.L. García-Martín, F. Crovetto, R. Fuertes-Martin, M. Domenech, M.D. Gómez-Roig, E. Eixarch, R. Estruch, S.R. Hansson, N. Amigó, N, Cañellas, F. Crispi, E. Gratacós. Metabolic profiling and targeted lipidomics reveals a disturbed lipid profile in mothers and fetuses with intrauterine growth restriction. Sci Rep. 2018;8(1):1–14.

16. J. McCall. Genetic algorithms for modelling and optimisation. J Comput Appl Math. 2005;184(1):205–22.

17. Bauer, D. J., & Curran, P. J. Probing interactions in fixed and multilevel regression: inferential and graphical techniques. Multivariate Behavioral Research. 2005;40(3): 373-400.

18. Kuhn, M. Building Predictive models in R using the caret package. Journal of Statistical Software. 2008;28(5): 1–26.

19. C. Reily, T.J. Stewart, M.B. Renfrow, J. Novak. Glycosylation in health and disease. Nat Rev Nephrol. 2019;15(6):346–66.

20. J.J. Lyons, J.D. Milner, S.D. Rosenzweig. Glycans Instructing Immunity: The Emerging Role of Altered Glycosylation in Clinical Immunology. Front Pediatr. 2015;3(54):1–10.

21. K. Liu, A.J. Paterson, E. Chin, J.E, Kudlow. Glucose stimulates protein modification by O-linked GlcNAc in pancreatic β cells: Linkage of O-linked GlcNAc to β cell death. Proc Natl Acad Sci USA. 2000;97(6):2820–5.

22. T. Keser, I. Gornik, F. Vučković, N. Selak, T. Pavić, E. Lukić, I. Gudelj, H. Gašparović, B. Biočina, T. Tilin, A. Wennerström, S. Männistö, V. Salomaa, A. Havulinna, W. Wang, J.F. Wilson, N. Chaturvedi, M. Perola, H. Campbell, G. Lauc, O. Gornik. Increased plasma N-glycome complexity is associated with higher risk of type 2 diabetes. Diabetologia. 2018;61(2):2352–60.

23. V. Dotz, F.H.R. Lemmers, R.K. Reiding, A.L.H. Ederveen, A.G. Lieverse, M.T. Mulder, E.J.G. Sijbrands, M. Wuhrer, M. van Hoek. Plasma protein N-glycan signatures of type 2 diabetes. Biochim Biophys Acta. 2018;1862(12):2613–22.

24. H.N. Ginsberg, Y.L. Zhang, A. Hernandez-Ono. Regulation of plasma triglycerides in insulin resistance and diabetes. Arch Med Res. 2005;36(3):232–40.

25. J. Long, Z. Yang, L. Wang, Y. Han, C. Peng, C. Yan, D. Yan. Metabolite biomarkers of type 2 diabetes mellitus and pre-diabetes: a systematic

review and meta-analysis. BMC Endocr Disord. 2020;20(1):1–17.

26. M. Guasch-ferré, J.L. Santos, M.A. Martínez-gonzález, C.B. Clish, C. Razquin, D. Wang. Glycolysis / gluconeogenesis- and tricarboxylic acid cycle – related metabolites , Mediterranean diet , and type 2 diabetes. 2020;(1):835–44.

27. W.T. Garvey, S. Kwon, D. Zheng, S. Shaughnessy, P. Wallace, A. Hutto, K. Pugh, A.J. Jenkins, R.L. Klein, Y. Liao. Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance. Diabetes. 2003;52(2):453–62.

28. D.M. Tehrani, Y. Zhao, M.J. Blaha, S. Mora, R.H. Mackey, E.D. Michos, M.J. Budoff, W. Cromwell, J. Otvos, P.D. Rosenblit, N.D. Wong. Discordance of Low-Density Lipoprotein and High-Density Lipoprotein Cholesterol Particle Versus Cholesterol Concentration for the Prediction of Cardiovascular Disease in Patients With Metabolic Syndrome and Diabetes Mellitus (from the Multi-Ethnic Study of Atherosclerosis [MESA]). Am J Cardiol. 2016;117(12):1921-1927.

29. A. Chait, H.N. Ginsberg, T. Vaisar, J.W. Heinecke, I.J. Goldberg, K.E. Bornfeldt. Remnants of the triglyceride-rich lipoproteins, diabetes, and cardiovascular disease. Diabetes. 2020;69(4):508–16.

30. M. Wyss, R. Kaddurah-Daouk. Creatine and creatinine metabolism. Physiol Rev. 2000;80(3):1107–213.

31. P. Wurtz, P. Soininen, A.J. Kangas, T. Rönnemaa, T. Lehtimäki, M. Kähönen, J.S. Viikari, O.T. Raitakari, M. Ala-Korpela. Branched-chain and aromatic amino acids are predictors of insulinresistance in young adults. Diabetes Care. 2013;36(3):648–55.

32. S.S.Patel, M.Z. Molnar, J.A. Tayek, J.H. Ix, N. Noori, D. Benner, S. Heymsfield, J.D. Kopple, C.P. Kovesdy, K. Kalantar-Zadeh. Serum creatinine as a marker of muscle mass in chronic kidney disease: Results of a cross-sectional study and review of literature. J Cachexia Sarcopenia Muscle. 2013;4(1):19–29.

33. J.R. Zierath, A. Krook, H. Wallberg-Henriksson. Insulin action and insulin resistance in human skeletal muscle. Diabetologia. 2000;43(7):821–35.

34. S. Kashima, K. Inoue, M. Matsumoto, K. Akimoto. Low serum creatinine is a type 2 diabetes risk factor in men and women: The Yuport Health Checkup Center cohort study. Diabetes Metab. 2017;43(5):460–4.

35. P. Srikanthan, A.L. Hevener, A.S. Karlamangla. Sarcopenia exacerbates obesity-associated insulin resistance and dysglycemia: Findings from the national health and nutrition examination survey III. PLoS One. 2010;5(5):1-7.

36. T. Melsom, U.D. Mathisen, O.C. Ingebretsen, T.G. Jenssen, I. Njølstad, M.D. Solbu, I. Toft, B.O. Eriksen. Impaired fasting glucose is associated with renal hyperfiltration in the general population. Diabetes Care. 2011;34(7):1546–51.

37. L. Tonneijck, M.H.A. Muskiet, M.M. Smits, E.J. Van Bommel, H.J.L. Heerspink, D.H. Van Raalte, J.A. Joles. Glomerular hyperfiltration in diabetes: Mechanisms, clinical significance, and treatment. J. Am. Soc. Nephrol. 2017;28(4):1023–39.

38. R. Okada, Y. Yasuda, K. Tsushita, K. Wakai, N. Hamajima, S. Matsuo. Glomerular hyperfiltration in prediabetes and prehypertension. Nephrol Dial Transplant. 2012;27(5):1821–5.

39. Qin P, Lou Y, Cao L, Shi J, Tian G, Liu D, Zhou Q, Guo C, Li Q, Zhao Y, Liu F, Wu X, Qie R, Han M, Huang S, Zhao P, Wang C, Ma J, Peng X, Xu S, Chen H, Zhao D, Zhang M, Hu D, Hu F. Dose-response associations between serum creatinine and type 2 diabetes mellitus risk: A Chinese cohort study and meta-analysis of cohort studies. J Diabetes. 2020;12(8):594-604.

40. P. Soininen, A.J. Kangas, P. Würtz, T. Suna, M. Ala-Korpela. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet. 2015;8(1):192–206.

Table 1. Univariate comparison between the reference group, the glucose-matched group and the incident type 2 diabetes group. The median and interquartile range for each variable and group are included. P value of each comparison are also shown.

	Reference group	Glucose-matched group	Incident type 2 diabetes group	p	Reference vs. Glucose-matched	Reference vs. Incident type 2 diabetes	Glucose-matched vs. Incident type 2 diabetes
	Median±IQR*	Median±IQR*	Median±IQR*				
Age, years	55.0 [42.0;63.0]	57.0 [48.0;66.0]	58.0 [47.0;64.0]	0.163	0.168	0.168	0.770
Sex:				0.258	0.518	0.380	0.518
Men	81 (55.9%)	74 (50.7%)	67 (46.2%)				
Women	64 (44.1%)	72 (49.3%)	78 (53.8%)				
BMI, kg/m^2	27.8 [25.8;31.2]	30.4 [27.2;33.5]	30.5 [28.3;33.9]	<0.001	<0.001	<0.001	0.743
Glyc B, µmol/L	328 [299;355]	350 [321;387]	346 [322;389]	<0.001	<0.001	0.001	0.722
Glyc F, µmol/L	212 [192;247]	232 [206;269]	231 [195;265]	0.003	0.002	0.060	0.252

Glyc A, µmol/L	682 [614;793]	743 [663;850]	744 [659;844]	0.003	0.005	0.005	0.986
H/W-Glyc B	4.13 [3.76;4.47]	4.39 [4.03;4.87]	4.35 [4.06;4.89]	<0.001	0.001	0.001	0.825
H/W Glyc A	15.4 [13.9;18.0]	16.6 [15.1;18.4]	16.7 [15.4;18.2]	<0.001	0.001	<0.001	0.720
VLDL-C, mg/dL	15.7 [10.5;21.5]	17.1 [10.7;22.6]	17.2 [11.7;25.0]	0.245	0.454	0.293	0.454
IDL-C, mg/dL	11.8 [8.49;14.9]	12.8 [9.44;15.7]	12.7 [10.0;15.6]	0.304	0.314	0.314	0.928
LDL-C, mg/dL	139 [125;156]	141 [123;161]	142 [126;160]	0.825	0.878	0.878	0.878
HDL-C, mg/dL	52.5 [47.8;58.4]	51.0 [44.9;59.3]	51.4 [44.5;58.4]	0.427	0.511	0.511	0.710
Total-C, mg/dL	227 [201;249]	226 [203;250]	225 [211;250]	0.689	0.693	0.693	0.693
VLDL-TG, mg/dL	57.3 [44.0;78.9]	64.5 [42.6;91.0]	66.3 [46.9;92.3]	0.076	0.281	0.073	0.281
IDL-TG, mg/dL	12.0 [9.39;14.4]	13.0 [10.1;15.3]	12.7 [10.5;14.9]	0.251	0.245	0.245	0.977
LDL-TG, mg/dL	17.2 [13.7;21.2]	17.9 [14.8;21.8]	18.0 [14.5;21.4]	0.435	0.550	0.550	0.736

HDL-TG, mg/dL	16.1 [12.5;19.0]	15.7 [12.1;19.4]	15.9 [13.1;19.3]	0.682	0.960	0.707	0.707
Total-TG, mg/dL	104 [81.8;131]	114 [84.9;141]	116 [90.1;149]	0.053	0.212	0.047	0.375
VLDL-P, nmol/L	43.2 [31.9;60.8]	50.0 [31.2;69.6]	50.7 [35.7;72.0]	0.076	0.307	0.065	0.307
L-VLDL-P, nmol/L	1.22 [0.94;1.59]	1.30 [0.92;1.78]	1.35 [1.01;1.88]	0.093	0.329	0.078	0.329
M-VLDL-P, nmol/L	4.30 [3.16;5.61]	4.05 [2.93;5.91]	4.38 [3.04;5.43]	0.608	0.707	0.707	0.707
S-VLDL-P, nmol/L	37.3 [28.1;54.1]	42.8 [27.5;62.4]	44.3 [31.0;64.5]	0.049	0.287	0.043	0.257
LDL-P, nmol/L	1420 [1262;1623]	1445 [1263;1662]	1478 [1314;1660]	0.271	0.448	0.298	0.448
L-LDL-P, nmol/L	198 [179;227]	194 [169;218]	194 [169;220]	0.201	0.200	0.200	0.999
M-LDL-P, nmol/L	414 [340;513]	429 [330;518]	403 [331;509]	0.702	0.914	0.738	0.738
S-LDL-P, nmol/L	787 [703;888]	789 [720;945]	841 [765;942]	0.015	0.246	0.010	0.160
HDL-P, µmol/L	27.5 [25.4;30.2]	27.3 [24.9;31.1]	27.6 [24.6;30.0]	0.724	0.755	0.755	0.755

L-HDL-P, μmol/L	0.28 [0.25;0.31]	0.28 [0.25;0.30]	0.28 [0.25;0.31]	0.543	0.541	0.945	0.541
M-HDL-P, μmol/L	9.02 [8.26;9.86]	8.71 [8.04;9.66]	8.96 [8.29;9.81]	0.264	0.349	0.648	0.377
S-HDL-P, μmol/L	18.3 [16.4;20.5]	18.6 [16.6;20.9]	18.0 [15.9;20.4]	0.560	0.714	0.697	0.697
VLDL-Z, nm	42.0 [41.8;42.2]	42.0 [41.7;42.2]	41.9 [41.6;42.1]	0.002	0.095	0.001	0.095
LDL-Z, nm	21.1 [20.9;21.2]	21.0 [20.7;21.2]	20.9 [20.7;21.2]	0.008	0.140	0.005	0.190
HDL-Z, nm	8.25 [8.21;8.31]	8.24 [8.20;8.30]	8.25 [8.21;8.31]	0.200	0.186	0.930	0.186
Acetone, μmol/L	0.01 [0.01;0.02]	0.02 [0.01;0.02]	0.02 [0.01;0.02]	0.293	0.553	0.366	0.497
Alanine, μmol/L	0.40 [0.34;0.46]	0.43 [0.38;0.47]	0.43 [0.38;0.50]	0.001	0.006	0.002	0.431
Creatinine, μmol/L	0.04 [0.04;0.06]	0.05 [0.04;0.06]	0.05 [0.04;0.06]	0.495	0.460	0.982	0.460
Creatine, μmol/L	0.05 [0.04;0.06]	0.06 [0.04;0.07]	0.06 [0.04;0.07]	0.010	0.007	0.099	0.254
Glucose, μmol/L	4.35 [3.96;4.71]	4.79 [4.40;5.25]	4.84 [4.42;5.43]	<0.001	<0.001	<0.001	0.452

Glutamine, μmol/L	0.38 [0.35;0.45]	0.42 [0.37;0.47]	0.39 [0.35;0.47]	0.041	0.053	0.704	0.073
Glutamate, μmol/L	0.20 [0.15;0.24]	0.20 [0.17;0.23]	0.19 [0.17;0.23]	0.819	0.960	0.960	0.960
Lactate, μmol/L	0.71 [0.58;0.97]	0.83 [0.65;1.05]	0.87 [0.66;1.11]	0.002	0.025	0.002	0.297
Valine, μmol/L	0.21 [0.19;0.24]	0.22 [0.20;0.25]	0.23 [0.20;0.26]	0.011	0.053	0.012	0.378
Tyrosine, μmol/L	0.05 [0.05;0.06]	0.06 [0.05;0.07]	0.06 [0.05;0.07]	<0.001	<0.001	<0.001	0.704
Glycine, μmol/L	0.26 [0.22;0.34]	0.28 [0.22;0.34]	0.26 [0.21;0.32]	0.460	0.693	0.713	0.580
Histidine, μmol/L	0.11 [0.10;0.12]	0.11 [0.10;0.12]	0.11 [0.10;0.13]	0.362	0.644	0.507	0.518
Isoleucine, μmol/L	0.05 [0.04;0.06]	0.06 [0.04;0.07]	0.06 [0.05;0.07]	0.004	0.013	0.006	0.615
Leucine, μmol/L	0.11 [0.10;0.13]	0.12 [0.10;0.14]	0.12 [0.10;0.15]	0.002	0.017	0.002	0.390
3-Hydroxibutirate, μmol/L	0.02 [0.01;0.04]	0.07 [0.04;0.09]	0.02 [0.02;0.03]	0.209	0.236	0.637	0.236

* IQR: Interquartile range

Figure legends

Figure 1. ROC curves for different ML-based models. The area under the curve (AUC) and the 95% confidence interval (in brackets) are displayed for LR (in blue), RF (in green), EGB (in orange) and SVM (in red).

Figure 2. Logistic regression coefficients. Statistically significant coefficients are indicated with an asterisk. The coefficients in red are associated with diabetes mellitus, whereas those in green are inversely associated with diabetes mellitus.

Figure 3. Johnson-Neyman plots for the evaluation of the interaction between Glyc A and Glyc B. The plot evaluates how the slope of Glyc A changes depending on the values of Glyc B. The interval in which that change is statistically significant is colored blue, whereas the interval in which it is not is colored pink.