scientific reports

Check for updates

OPEN

Breath biomarkers of insulin resistance in pre-diabetic Hispanic adolescents with obesity

Mohammad S. Khan[1,2], Suzanne Cuda[3,4], Genesio M. Karere[1,2], Laura A. Cox[1,2] & Andrew C. Bishop[1,2]✉

Insulin resistance (IR) affects a quarter of the world's adult population and is a major factor in the pathogenesis of cardio-metabolic disease. In this pilot study, we implemented a non-invasive breathomics approach, combined with random forest machine learning, to investigate metabolic markers from obese pre-diabetic Hispanic adolescents as indicators of abnormal metabolic regulation. Using the ReCIVA breathalyzer device for breath collection, we have identified a signature of 10 breath metabolites (breath-IR model), which correlates with Homeostatic Model Assessment for Insulin Resistance (HOMA-IR) ($R = 0.95$, $p < 0.001$). A strong correlation was also observed between the breath-IR model and the blood glycemic profile (fasting insulin $R = 0.91$, $p < 0.001$ and fasting glucose $R = 0.80$, $p < 0.001$). Among tentatively identified metabolites, limonene, undecane, and 2,7-dimethyl-undecane, significantly cluster individuals based on HOMA-IR ($p = 0.003$, $p = 0.002$, and $p < 0.001$, respectively). Our breath-IR model differentiates between adolescents with and without IR with an AUC-ROC curve of 0.87, after cross-validation. Identification of a breath signature indicative of IR shows utility of exhaled breath metabolomics for assessing systemic metabolic dysregulation. A simple and non-invasive breath-based test has potential as a diagnostic tool for monitoring IR progression, allowing for earlier detection of IR and implementation of early interventions to prevent onset of type 2 diabetes mellitus.

An estimated 1.26 billion adults worldwide have insulin resistance (IR), a complex pathophysiological state connected to an imbalance between insulin and glucose metabolism, which is strongly associated with the development of cardio-metabolic disease[1–3]. Although aging is considered a strong predictor of IR, a meta-analysis of 18 population-based studies from 13 countries indicated that approximately 312 million children globally have IR[4], suggesting that other factors also contribute to IR development earlier in life. Obesity is one other major contributing factor contributing to IR. In the United States, between 2015 and 2016, 18.5% of youth between the ages of 2–19 were obese[5,6]. Severe obesity is also racially and ethnically disproportionate with two-to-four fold higher rates among African American or Hispanic populations compared to their Caucasian counterparts[6]. Therefore, more research is needed to investigate these populations to slow or prevent long term health consequences.

The exact pathophysiology of IR is unclear but involves multiple associations of cellular dysfunction. IR is inversely related to the insulin sensitivity in insulin-dependent cells such as skeletal muscle, adipocytes, and cardiomyocytes[7]. When insulin sensitivity is low, these cells fail to respond to insulin signaling transduction, which is required for glucose uptake[8]. As a result, IR is a common feature in multiple metabolic disorders including obesity, dyslipidemia, hypertension, atherosclerosis, nonalcoholic fatty liver disease (NAFLD), type 2 diabetes mellitus (T2DM), and some cases of type 1 diabetes mellitus (T1DM)[1,8,9]. One potential cause of IR is obesity-induced inflammatory cytokines and inflammatory mediators such as tumor necrosis factor-α (TNF-α), monocyte chemotactic protein-1 (MCP-1), C-reactive protein (CRP), and interleukins are upregulated in individuals with IR[10]. Another factor that plays an important role in IR is reactive oxygen species (ROS). Individuals with IR have reduced antioxidants and increased production of the ROS leading to increased systemic oxidative stress[11]. ROS induced oxidative stress metabolites such as hydrogen peroxide, protein carbonyls, protein oxidative products, carbohydrate metabolites, short chain aldehydes, ketones, and hydrocarbons have been found to strongly associated with IR[11]. Further research into these factors associated with IR development at the earlier stages or

[1]Department of Internal Medicine, Section on Molecular Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA. [2]Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA. [3]Health and Weight Management Clinic, Children's Hospital of San Antonio, San Antonio, TX 78207, USA. [4]Baylor College of Medicine, Houston, TX 77030, USA. ✉email: abishop@wakehealth.edu

prior to the onset of clinical symptoms may uncover early biomarkers of IR. The addition of these biomarkers into the clinical setting would add to the clinical tools available for clinicians to identify at risk individuals prior to onset of disease and initiate earlier interventions.

Current clinical practice for assessing IR utilizes blood-based measures requiring fasting for 8 h and multiple visits to a medical facility for blood sampling and consultation. This is a burden for individuals who may benefit from frequently monitoring IR status, particularly young children and adolescents who rely on parents or guardians to facilitate this medical care. Individuals with IR may not have blood glucose levels that are in a range to be clinically diagnosed as T2DM but may be in a pre-diabetic stage of metabolic dysfunction. This stage is particularly important because it is a window of time when health and weight management strategies can be implemented to improve long-term cardio-metabolic health outcomes. Evidence suggests 70% of IR individuals will develop cardio-metabolic disease within 15–20 years[12,13] without any intervention. This emphasizes the need to develop clinically relevant diagnostics tools to identify at risk individuals for intervention to slow or prevent progression to cardio-metabolic disease.

Non-invasive sampling, such as exhaled breath, provides metabolic information that may be indicative of an individual's health status. Exhaled breath metabolites have been reported in the assessment of infection[14,15], cognition[16], metabolic disease[17], and lifestyle[18]. These studies provide evidence that metabolic markers, which strongly correlate with metabolic dysregulation, can be identified in breath samples.

Previously, we reported an exhaled breath metabolomics study investigating a young cohort of non-human primates (NHP) who developed insulin insensitivity due to developmental programing[19]. Following the development of IR, a volatile organic compound (VOC) breath signature was identified that discriminated between NHPs with IR compared to controls. It is suggested the altered breath signature arose from an altered cardiometabolic state potentially due to increased ROS production, beta-oxidation, inflammation, and lipid peroxidation[19].

Based on these findings, we hypothesized that metabolic markers contained in exhaled breath could be a potential diagnostic resource for defining metabolic status in humans and furthermore could be used to diagnose IR in pre-diabetic adolescents. Our goals were two-folds: (1) To show feasibility of adolescent breath collections using the ReCIVA breath collection device (Owlstone medical, Cambridge, UK). (2) To identify a small set of breath biomarkers in a pre-diabetic population that are important for the detection of IR and the risk factors for cardio-metabolic disease development.

Results

Feasibility of breath sampling from prediabetic adolescents. Twenty-eight adolescents participated in this study with a mean age of 15.46 years old (range 13–17 years old). All participants were obese or severely obese with an average BMI percent of the 95th percentile of 137.96 ± 20.15 (range 95–186), which expresses the extent of obesity (BMI ≥ the 95th percentile are considered obese) among those adolescents[20]. The average exhaled breath sample collection time was 4.10 ± 0.05 min. At the time of collection all participants had not developed T2DM based on clinical blood measurements and clinician assessment.

Increased cardiovascular risk factors in Hispanic adolescents with obesity. As part of the routine clinical assessment, clinical blood tests were performed, and clinical characteristic are included in Table 1. Fasting glucose and insulin measures were used to calculate homeostatic model assessment for insulin resistance (HOMA-IR), a surrogate measure of insulin sensitivity[21]. Based on HOMA-IR measures, three groups could be defined in our cohort, adolescents without IR (normal), adolescents with IR and adolescents with borderline IR with a statistically significant difference in IR value at < 0.001 in a Chi-square test of multiple groups. Adolescents with IR had a statistically significant higher BMI percentile ($p < 0.001$) with an average of 145.20 ± 21.69 compared with the normal adolescents 127.83 ± 16.57. Further emphasizing a prediabetic status of the cohort, measures of cardiovascular disease (CVD) risk factors, were normal for total cholesterol (average 160.67 ± 31.84 mg/dL) and LDL-cholesterol (average 107.42 ± 25.79 mg/dL) and low for HDL-cholesterol (43.30 ± 11.48 mg/dL). However, normal fasting glucose (91.32 ± 14.07 mg/d) and fasting insulin (21.68 ± 12.54 mU/L) levels were significantly different ($p < 0.001$) in adolescents with IR. Detailed clinical measures are presented in Table 1.

Machine learning approach identifies important breathprint for IR. The random forest (RF) regression model identified the most important breath features for identifying individuals with IR based on their HOMA- IR value. Ten features were selected with the highest importance as shown on the RF importance plot (Supplementary Fig. S1). The elbow cut-off > 18 were selected based on the large drop on the important measures on the mean decrease in Gini in the RF model[23].

The ten important analytes include limonene, decane-2,4,6-trimethyl, undecane, undecane-2,7-dimethyl, pentylbenzene, octamethyloctane, eicosane and three unknown analytes (Supplementary Table S1). Identification of these analytes was confirmed by retention indices and mass spectral match score with NIST 2020 library's analytical standard (level 2 and 3 identification)[24]. The unknowns reported are those whose retention indices were consistent, but the mass spectra matching was inconsistent or below 700 similarity score, hence a compound name was not assigned. Chromatographic and mass spectral information on the ten compounds is provided in the Supplementary Table S1. An example of a peak detection and identification used in this study are shown in Supplementary Fig. S2.

The breath-IR model correlates with the standard blood based HOMA-IR. The ten breath VOCs identified in the training dataset, were evaluated against the remaining samples in the test dataset (Fig. 1a). Test dataset comprised of a total of 37 samples that were not used for the training dataset. The test dataset also resulted in a positive correlation with blood based HOMA-IR values with a Pearson correlation, R = 0.87,

	Borderline (N = 7[±]) HOMA-IR 2.60–3.80	Insulin resistance (N = 15) HOMA-IR > 3.80	Normal (N = 6) HOMA-IR < 2.60	Total (N = 28)	*p* value[#]
BMI (% of the 95th Percentile)					< 0.001
	131.14 (12.33)	145.20 (21.69)	127.83 (16.57)	137.96 (20.15)	
Cholesterol					0.951
	162.00 (20.74)	160.66 (28.40)	159.16 (48.11)	160.67 (31.84)	
LDL					0.137
	114.14 (22.38)	107.333 (23.177)	99.83 (33.59)	107.42 (25.79)	
HDL					< 0.001
	36.614 (5.287)	43.82 (6.05)	49.81 (20.25)	43.30 (11.48)	
Triglycerides					0.13
	107.42 (34.26)	115.26 (79.19)	85.00 (21.10)	106.82 (62.07)	
Glucose (mg/dL)					< 0.001
	86.28 (7.27)	96.73 (16.44)	83.66 (5.58)	91.32 (14.07)	
Insulin mU/L					< 0.001
	14.87 (1.56)	29.32 (12.72)	10.56 (1.84)	21.68 (12.54)	
Glucose/insulin ratio					< 0.001
	0.32 (0.05)	0.20 (0.05)	0.45 (0.10)	0.28 (0.12)	
[¥]HOMA-IR					< 0.001
	3.16 (0.39)	7.40 (5.16)	2.16 (0.33)	5.22 (4.46)	
[≠]FIRI					< 0.001
	0.23 (0.03)	0.14 (0.03)	0.32 (0.07)	0.20 (0.08)	
A1C					0.008
	5.00 (0.26)	5.30 (0.59)	5.06 (0.17)	5.17 (0.47)	
AST					< 0.001
	20.00 (9.09)	31.46 (17.38)	21.00 (4.71)	26.59 (14.80)	
ALT					< 0.001
	24.66 (16.27)	50.40 (32.80)	24.83 (9.94)	39.00 (28.90)	
CRP					0.068
	0.65 (0.69)	0.78 (0.94)	0.29 (0.30)	0.65 (0.80)	
Sex					0.559
F	2 (28.6%)	6 (40.0%)	2 (33.3%)	10 (35.7%)	
M	5 (71.4%)	9 (60.0%)	4 (66.7%)	18 (64.3%)	

Table 1. Demographic and clinical characteristics of the study cohort. The HOMA-IR cut-offs, are based on a study in a Hispanic population[22]. Data are presented as mean and standard deviation (SD). [±]Each participant provided 4 technical replicate samples for the analysis. [#]*P*-value is calculated by Chi-square test of multiple groups. [¥]Homeostatic model assessment of insulin resistance (HOMA-IR) is calculated by the formula: fasting insulin (microU/L) x fasting glucose (nmol/L)/22.5. [≠]Fasting insulin resistance index (FIRI) is calculated by the formula: (fasting glucose × fasting insulin)/25.

$p < 0.001$. Combing the training and test datasets in a total dataset resulted in a higher correlation with the blood based HOMA-IR values with a Pearson correlation, R = 0.95, and $p < 0.001$ (Fig. 1b).

Breath-IR model correlation with the glycemic profile but not with the lipid profile. The breath IR model shows a strong correlation with measures of glucose metabolism including fasting glucose (mg/dL) and fasting insulin (mU/L) with R = 0.91, $p < 0.001$ and R = 0.80, $p < 0.001$, respectively (Fig. 1c, d). A weak positive correlation was observed with triglycerides (mg/dL), R = 0.35, $p < 0.001$ and weak negative but non-significant correlation with the LDL levels at R = -0.028, $p < 0.77$ (Supplementary Fig. S3). The breath IR model did not correlate with the total cholesterol (mg/dL) or HDL level mg/dL Supplementary Fig. S3. The breath-IR model also correlated with additional indices of IR; fasting insulin resistance index (FIRI)[25] and the glucose/insulin (G/I) ratio showed in the Supplementary Fig. S3. The HbA1c test positively correlated with the breath IR model, R = 0.63 and $p = < 0.001$ (Supplementary Fig. S3).

Breath-IR model accurately identifies the IR classification. To determine the breath-IR model accuracy for identifying individuals with IR, the ten features identified by RF were further investigated. Three of the ten VOCs, limonene, undecane, and undecane- 2,7-dimethyl were statistically significantly different for individuals with IR ($p = 0.003$, $p = 0.002$, and $p < 0.001$, respectively) compared to those without IR (Fig. 2). A trend was observed for the remaining analytes but was not statistically significant Supplementary Fig. S4. A moderate

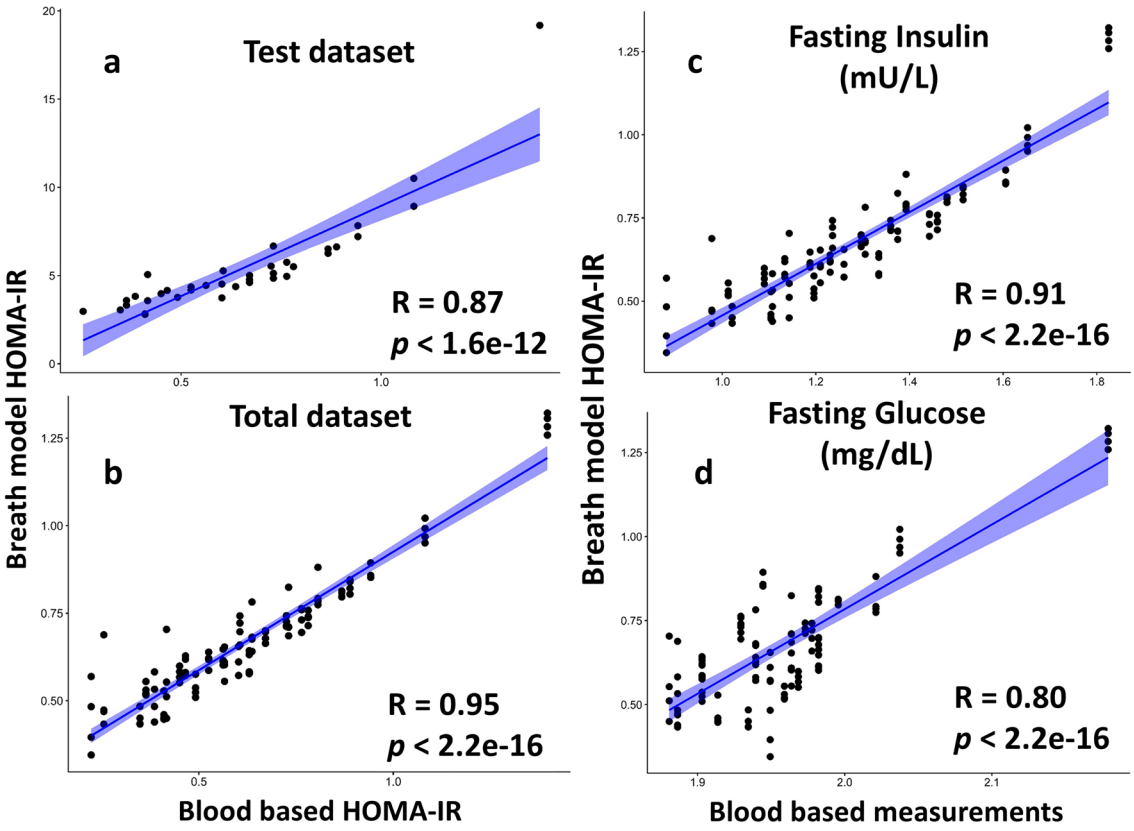

Figure 1. Pearson correlation of the breath-based IR with the blood based HOMA-IR measurement, fasting insulin and glucose. (**a**) Correlation of a test dataset against the blood based HOMA-IR measures testing the Breath IR training set on independent samples. (**b**) Correlation of the total dataset combining training and test datasets for the breath-IR model against blood based HOMA-IR. Correlation of the breath based HOMA-IR model against, (**c**) fasting insulin (mU/L) and (**d**) fasting glucose (mg/dL).

difference was observed for limonene when comparing borderline IR individuals to normal ($p = 0.068$), whereas comparison of borderline individuals to IR group was not significant ($p = 0.82$).

To evaluate the combined performance of the breath-IR model for clustering between the HOMA-IR categories, an orthogonal partial least squares discriminatory analysis (OPLS-DA) was performed[26–28]. The OPLS-DA showed a clear clustering by the HOMA-IR score, shown in Fig. 3. Two extreme HOMA-IR values (24.92, 12.07) were removed from the dataset to show clearer clustering, but even with those extreme values, a similar pattern of clustering of the individuals based on the breath-IR model was observed (Supplementary Fig. S5).

To evaluate the performance of the breath-IR model for classifying individuals with IR, the area under the receiver operating characteristic curve (AUC–ROC) was plotted resulting in a breath IR model value of 0.87 (Fig. 4). This value is generated by an RF iteration based on the selected ten breath compounds. The process was repeated for only two most significantly different VOCs, limonene and undecane reducing to 0.76. The RF model's importance was also assessed by randomly selecting features in the dataset resulting in an AUC of 0.52. The breath-IR model observed sensitivity and specificity as 73.1% (60.0–83.0% in cross-validation) and 81.7% (70.0–89.0% in cross-validation), respectively.

Discussion

Our study shows the feasibility of collecting non-invasive exhaled breath metabolites from an adolescent population, which can then be used to assess an individual's metabolic health status. Utilizing an untargeted metabolomics approach allowed for the identification of all breath metabolites in detectable range of our mass spectrometer. By implementing a machine learning model to assess the normalized data, a predictive model based on the selected features could be built to inform diagnostic potential of exhaled breath signatures. Our previous studies in nonhuman primates identified exhaled breath signatures in animals with IR[19]. The adoption of the HOMA-IR categories was based on a population-based study reported from the adult Mexican Americans population[22], as no clear definition exist for the Hispanic adolescents with IR. Our breath-IR model strongly correlates with all glycemic clinical measurements, but the model did not strongly correlate other clinical measures (i.e., circulating lipids). Lipid levels were not in a range that would have clinically categorized the participants as having T2DM, therefore the breath IR model is specific to measures impacting IR. Given the breath-IR model resulted in a strong correlation with blood based HOMA-IR measurements, information contained in exhaled breath samples has utility for a non-invasive IR diagnostic.

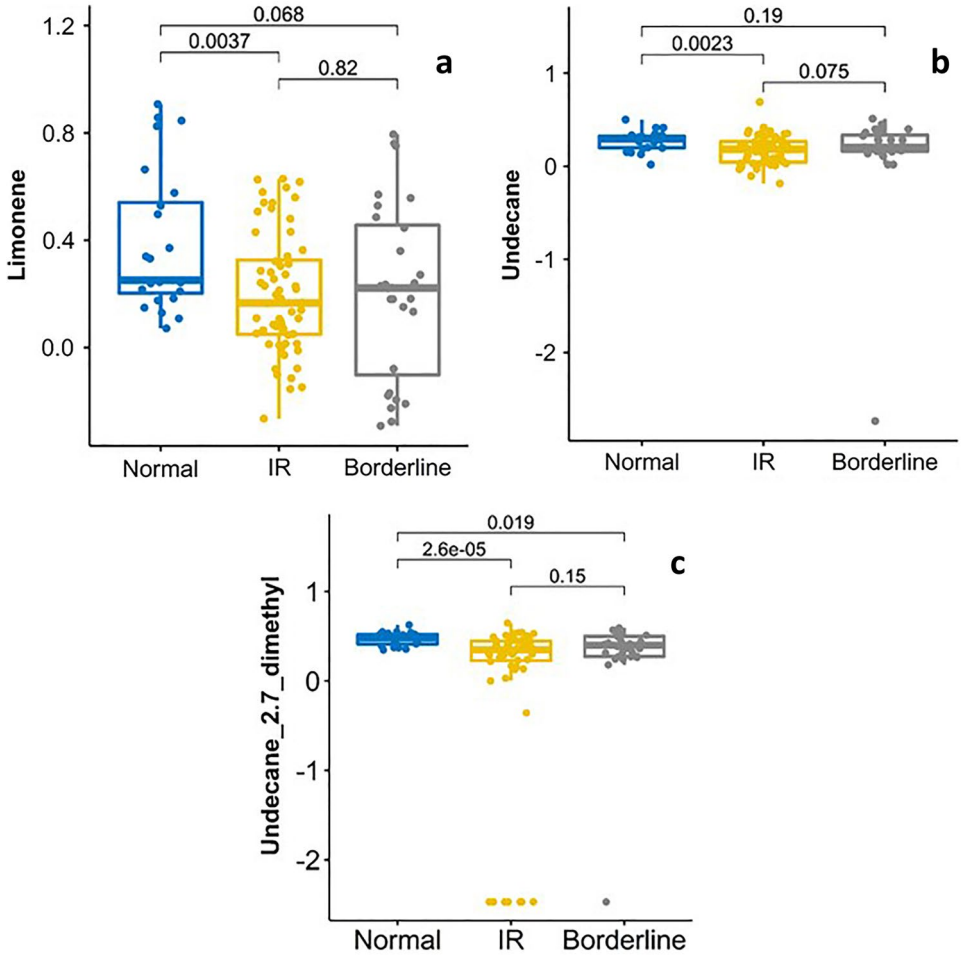

Figure 2. Boxplots of mean centered and normalized peak area for three compounds selected from the breath-IR model across normal, IR and borderline IR adolescents. Each compound's median observed peak area between the groups were different indicating a univariate difference which may be contributing the discrimination of the IR group in the combined 10 compounds model. Boxplots represent the quartiles of the data (first line is the first quartile, midline is the median, third line is the third quartile) where whiskers represent 1.5 × IQR (inter-quartile range).

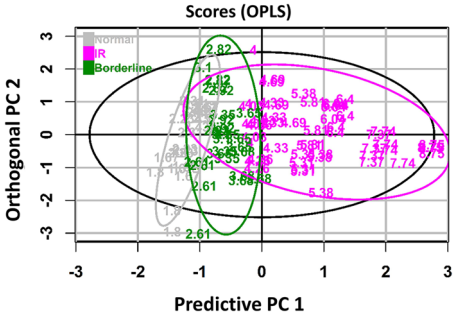

Figure 3. The supervised orthogonal partial least-squares discriminant (OPLS) analysis of breath HOMA-IR model. The score scatter plot of OPLS model showing the clustering of samples in horizontally by predictive principal component (PC) 1 and vertically by orthogonal PC 2. The IR cut-off, is based on a study in Hispanic population[22]. The purple color indicates the IR group with HOMA-IR > 3.80, the green color represents the borderline IR group with HOMA-IR 2.60–3.80 and HOMA-IR < 2.60 indicate normal group which is presented as grey color. The 95% tolerance region corresponds to the ellipse that is defined by the Hotelling's T2 parameter.

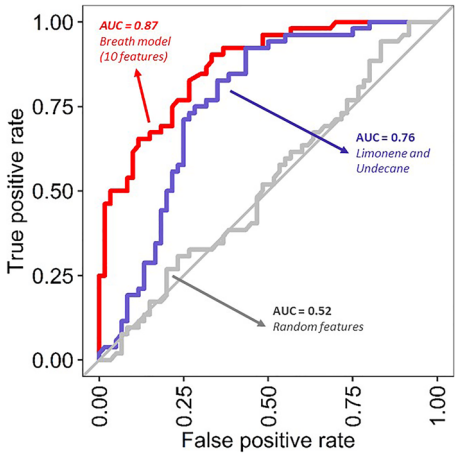

Figure 4. Diagnostics performance of the Breath-IR model. The receiver operating characters (ROC) curves using the random forest model between the IR (HOMA-IR > 3.80) and normal (HOMA-IR ≤ 3.80) adolescents. The 'red line' indicates the model's performance based on 10 compounds in breath IR model, the 'blue line' indicate performance by only two statistically significant compounds, limonene and undecane. The 'gray line' indicates randomly selected 2 breath metabolites from the original dataset. The AUC-ROC across folds after cross validation are 0.87 for 10 features breath model, 0.76 for 2 significant compounds, and 0.52 for random features model.

This pilot study is exploratory in nature and our breath metabolites distribution is based on relative abundance of each VOC. Therefore, metabolite levels are based on the normalized peak area not the actual quantity of metabolites. Three of the ten important features identified in the breath-IR model limonene, undecane, and undecane-2,7-dimethyl showed a statistically significate difference between the normal and IR group. These three metabolites are the major contributors to the accuracy of the model. The OPLS-DA model filtered out noise in the dataset that is not correlated with the outcome variables[29]. Given two of the adolescents had an exceptionally high HOMA-IR (24.92 and 12.07), these two values put a biased weight on the classification problem. Although those values are extreme within the current cohort, two different OPLS-DA models, one without those values and one with those values, as shown in Fig. 3 and Supplementary Fig. S5, respectively. These plots show clear clustering of individuals by their IR levels based on the breath metabolites, as IR individuals with HOMA-IR (> 3.80) were clustered on the right side of the plots whereas the normal HOMA-IR individual (< 2.60) on the left. The border line IR group with HOMA-IR 3.8–2.6 were also clustered in-between of the IR and normal group although there is some overlap. Although the other seven breath metabolites did not reach significance between the groups individually (Supplementary Fig. S4), their contribution to the breath-IR model is evident when assessing the diagnostic capability of the model (Fig. 4) to identify individuals with IR. A combined signature of breath metabolites is clearly more informative than using a single metabolite which may be indicative of some normal variability in metabolites within and between individuals[15,19,30].

Our breath IR model shows good accuracy 77.8% with sensitivity of 73.1% (60–83% within cross-validation) and specificity of 81% (70–89% within cross-validation) (Fig. 4). As a comparison, the American Diabetes Association (ADA) and International Expert Council of the ADA define the diabetes diagnostic criteria for hemoglobin A1c (HbA1c); sensitivity, 73.3%; specificity, 88.2%[31,32]. Although our study needs to be verified in a larger more diverse independent cohort, our results are promising for the breath-based IR diagnostic tool for monitoring IR and metabolic health, which could be developed further.

Among the seven named breath biomarkers, limonene, undecane, pentylbenzene, and eicosane have been previously reported in the human metabolome according to the KEGG[33] and HMDB[34] databases. Undecane was previously detected as a biomarker of nonalcoholic fatty liver disease (NAFLD)[35], asthma[36], and gastrointestinal disease[37]. Pentylbenzen and eicosane are hydrocarbon molecules and subcellular component of membrane epithelium[38,39]. These two analytes were previously reported from human metabolomics studies but were never quantified[34]. Similar hydrocarbon metabolites were reported as oxidative stress markers of disease in exhaled breath related to infectious disease[15], heart disease[40,41], psychiatric disease[42], and cancer[43]. This mechanism involves hydrocarbons being triggered or released in to the breath by ROS, a form of oxygen with a reactive electron from the by-product of oxidation metabolism in the mitochondria. These highly reactive molecules when released to the cytoplasm have the potential to damage DNA, protein, and other cellular metabolites, producing a range of small chain saturated or unsaturated hydrocarbon[44]. These hydrocarbons then have the potential to enter blood circulation and be detected in the breath as non-specific oxidative markers due to normal blood gas exchange in the lungs. IR has been established as a strong contributor to oxidative stress and is known to damage the vascular lining of cells, increasing the risk of atherosclerosis[45]. By identifying oxidative markers in exhaled breath from the obese adolescents that correlate with altered metabolic status within this cohort, suggests the underlying oxidative stress may be specific to the IR development.

Several studies have reported limonene from blood, feces, urine, saliva, and breath[46–50]. Limonene is a monocyclic monoterpenoid, with one isoprene chain. It is naturally abundant in nature and used in the flavor and

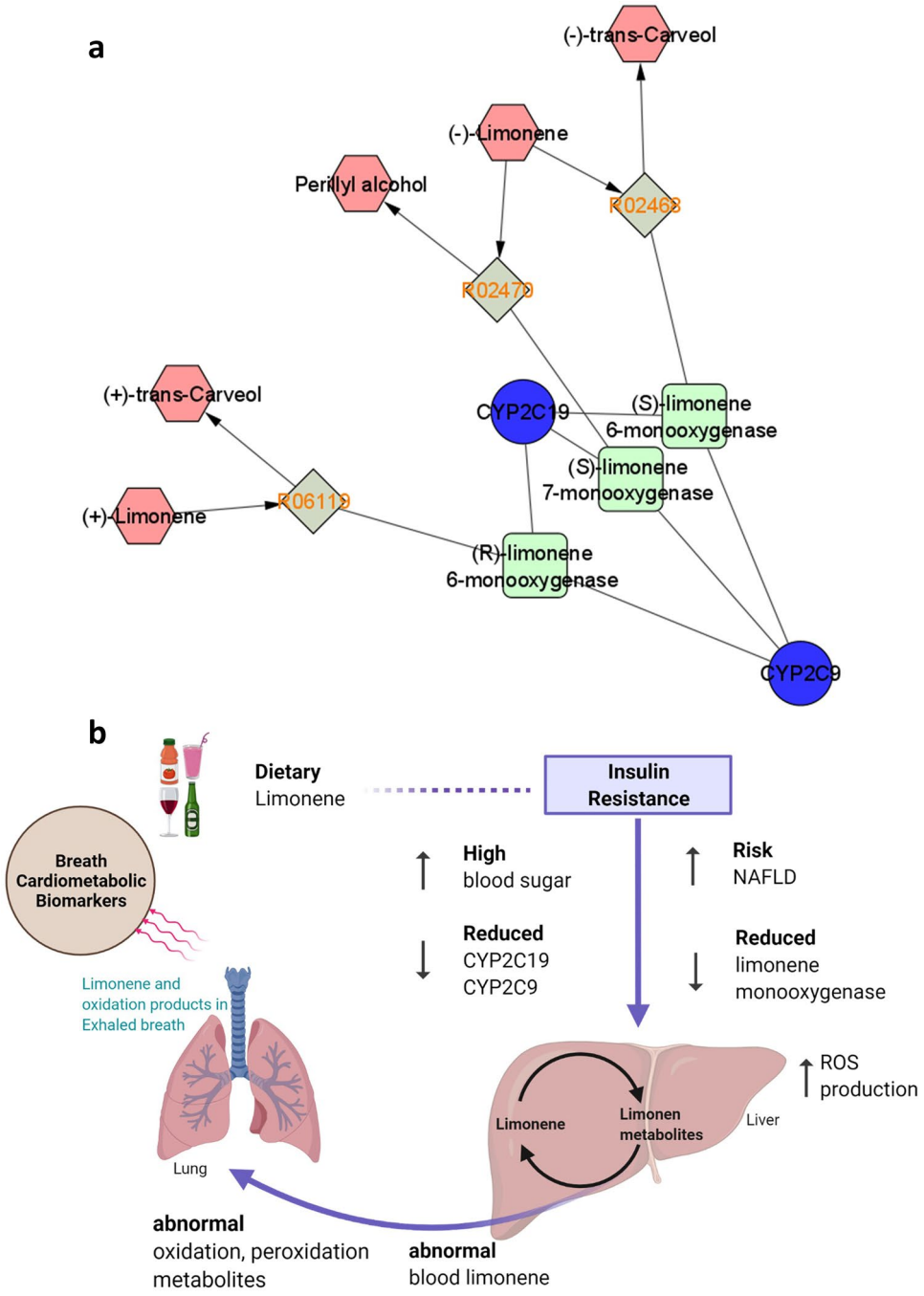

Figure 5. Molecular network analysis of limonene. The complex interaction of the gene, engymes and the metabolites are presented by Cytoscape[66] platform using Metscape 2[65] network anlaysis **(a)**. The reaction R02468, R02470 and R06119 connect the injested (+) limonene and (−) limonene to its oxidized metabolites, (+) trans-carveol, (−) trans-carveol and perillyl alcohol. The enzymes responsible for these reactions are P450 enzymes CYP2C9 and CYP2C19 and their products (S)-limonene 6-monooxygenase, (S)-limonene 7-monooxygenase and (R)-limonene 6-monooxygenase and (**b**) the proposed mechanism of the breath cardio-metabolic biomarkers. Increased ROS production, abnormal limonene and oxidation metabolites due to liver damage may lead to an abnormal level of breath metabolites. The figure is created using biorender[71] platform.

fragrance of foods and drinks. In exhaled breath studies of liver cirrhosis patients elevated level of limonene were observed compare to the controls[51–53]. As an exogenous metabolite, limonene is ingested during dietary intake and then metabolized in the liver by P450 enzymes CYP2C9 and CYP2C19 and their enzyme products (S)-limonene 6-monooxygenase, (S)-limonene 7-monooxygenase and (R)-limonene 6-monooxygenase (Fig. 5a). Limonene can then be metabolized to trans-carveol and perillyl alcohol[54]. It has been noted that in patients with liver disease, particularly NAFLD at any stage, have a reduced capacity to produce both CYP2C and CYP2C19

enzymes[55]. Reduced liver capacity to metabolize limonene results in abnormal levels in circulation and excretion leading to a difference on the limonene which can be identified in exhaled breath (Fig. 5b).

ROS arises from vascular organs including heart cells[56] or kidney cells[57] leading to a higher systemic level of oxidation and peroxidation products. This may include aliphatic or aromatic hydrocarbon, straight-chain mono- or poly-unsaturated aldehyde, straight-chain -mono or -poly unsaturated carboxylic acids, ester, epoxides, MUFAs and PUFAs[58]. Higher levels in the blood may cross the blood–air barrier within the gas exchanging region of the lungs and exhaled through breath. Thus a set of metabolites produced from the multiple organs as a result of cardio-metabolic disease development could potentially lead to a set of breath based cardio-metabolic biomarkers as indicated in Fig. 5b. Our detection of limonene as an important breath metabolite for determining IR status, may be due to early signs of NAFLD in obese adolescents' breath used in this study.

Given that the abnormal levels of limonene are present in individuals with IR suggest potential NAFLD and liver damage, we investigated clinical measures of two liver specific enzymes aspartate aminotransferase (AST) and alanine aminotransferase (ALT) and systemic inflammation marker c-reactive protein (CRP). AST and ALT are both liver specific enzymes that become elevated in blood as a result of liver damage. Our breath-IR model was slightly negative but significantly correlated with the AST/ALT ratio (Supplementary Fig. S6a). The AST/ALT ratio less than < 1 is suggestive of NAFLD/NASH whereas the score > 2 is suggestive of alcoholic liver disease[59]. A negative correlation between breath IR and AST/ALT ratio thus indicates adolescents have early NAFLD/NASH. Evaluation of CRP levels were moderately positive and statistically significant correlated with our breath-IR model (Supplementary Fig. S6b). This confirms that the presence of systemic inflammation that could negatively impact multiple organ systems throughout the body including the liver of those adolescents.

Conclusion

This pilot study demonstrates the feasibility of studying breath in adolescents and the potential of non-invasive breath metabolites as an important tool to detect IR in a unique cohort of Hispanic pre-diabetic adolescents with obesity. The breath model developed by a machine learning-based approach showed a strong correlation with the surrogate blood test for IR that is currently used in the clinical environment. The breath IR model confidently clustered individuals with and without IR, showing a promising diagnostics performance of the breath metabolites as evident by the ROC–AUC = 0.87. The detection breath metabolites as important features for IR status may result from early liver damage or due to increased cellular damage from ROS production as a result of obesity and lifestyle. Future studies will focus on validating the usefulness of breath signatures for detection and monitoring of IR in a larger population of at-risk individuals.

Materials and methods

Ethical approval of study. All study procedures were approved by the Institutional Review Board for Baylor College of Medicine and Affiliated Hospitals (IRB-H-40940). Informed consent was obtained from parents or guardians and assent from participants. This work was performed in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans.

Study design and participants. Participants were recruited from the Pediatric Health and Weight Management Clinic at the Children's Hospital of San Antonio, TX, USA. This cross-sectional study recruited Hispanic adolescents during their initial clinical visit. Criteria for inclusion were male and female ages ranging from 13 to 17 years with a BMI $\geq$ of the 95th percentile. Exclusion criteria of the patients included a diagnosis of T2DM, asthma, or other chronic respiratory condition and any use of neurohormonal medications.

Clinical assessments. Upon arrival to the clinic, a trained medical technologist collected anthropometric measures. Height and weight were collected using a BSM170 digital stadiometer and Scale 570 (InBody, Cerritos, CA, USA). BMI was calculated as kg/m^2 and z score calculated for all patients using their sex, precise age based on the date of birth and date of assessment, height, and weight. BMI percentile is expressed as the percent of the 95th percentile. Participants and parents/guardians verified fasting status prior to blood samples being collected as a standard of clinical care blood work. Blood (3 mL) was collected in red-top vacutainer tubes (BD, Franklin Lakes, NJ, USA). Clinical blood measures included: fasting insulin (mU/L) and fasting glucose (mg/dL), Hemoglobin A1c (HbA1c), alanine aminotransferase test (ALT), aspartate aminotransferase test (AST), fasting lipid profile, and high sensitivity CRP. The Homeostatic Model Assessment of Insulin Resistance, (HOMA-IR) of the participants was calculated based on the formula: fasting insulin (microU/L) $\times$ fasting glucose (nmol/L)/22.5[60]. The HOMA-IR categories are based on an adult Hispanic population based study[22]. Table 1 reports the demographic and clinical measurement of all participants involved in this study.

Exhaled breath sample collection. On the same day of blood collection, participants provided exhaled breath samples through the breath collection protocol described below. The exhaled breath samples were collected using a commercially available breath sampler device, ReCIVA (Owlstone medical, Cambridge, UK). This device safely monitors exhaled air pressure and CO_2 production allowing for consistent exhaled breath collections. The ReCIVA device was supplied medical-grade air (Airgas, Radnor, PA, USA) at a rate of 40 L/min per manufacture recommendations to avoid artifacts and contamination from room air. Exhaled breath metabolites were concentrated onto bio-monitoring thermal desorption tubes (TDT) (CAT# C2-CAXX-5149, Markes International, Sacramento, CA, USA) designed to capture VOCs from breath. Exhaled breath was collected with the following protocol:

1. Device collection parameters were set prior to starting of the collection in ReCIVA Breath Sampling Controller Software (Owlstone medical, Cambridge, UK). Total collection volume was set to 500 mL per tube at a rate of 200 mL/min for lower and upper airways resulting in a mixed fraction sampling.
2. Upon entering the participant room, a trained research assistant opened a new one-time use breath collection mask and assembled the four TDT into the ReCIVA device. The ReCIVA mask was placed over the participant's mouth and nose and they were advised to breathe normally and familiarize themselves prior to the start of collection.
3. Once the participant acknowledged readiness, the breath collection was started by the research assistant.
4. Once the software acknowledged the collection was complete, the ReCIVA mask was removed from the participant and discarded. TDTs were capped, transported to the research laboratory at the Texas Biomedical Research Institute and stored at 4 °C until analysis.

Thermal desorption and analytical instrumentation. The exhaled breath VOCs collected on the TDT were processed within 1 month of collection. TDT was desorbed at 260 °C for 10 min to a cryotrap (general-purpose CAT#U-T11GPC-2S) using a thermal desorption unit, TD100xr (Markes Int., Sacramento, CA, USA). VOCs concentrated on the cryotrap were rapidly heated at 40 °C/min to 280 °C and desorbed to a comprehensive two-dimensional gas chromatography tandem Time-of-Flight mass spectrometer (GC × GC-TOFMS. LECO Corp, St. Joseph MI, USA). The column set was configured as follows: ^{1}D Rxi- 5Sil-MS (95% polydimethylsiloxane, 5% phenyl; 30 m × 0.25 mm dc, 0.25 µm d_f) coupled with ^{2}D Stabilwax column (Crossbond polyethylene glycol; 2.21 m × 0.18 mm dc, 0.18 µm d_f) (Restek, Bellefonte, PA, USA). The inlet and ^{1}D column were connected by a PressFit column connector and the ^{1}D and ^{2}D were connected by an MXT-union connector kits (Restek, Bellefonte, PA, USA). The ^{2}D column end is feed into the transfer line to the MS. The helium carrier gas flow was set to 2 mL/min for the entire run. The GC oven temperature program was set to ramp from 40 to 100 °C by 5 °C/min and then ramp by 8 °C/min to 220 °C with a final 10 min hold for a total run time of 38 min.

Spectral data analysis. Raw data were processed using Chromatof software (version 4.72, LECO Corp, St. Joseph MI, USA). Spectra were collected over the range of *m/z* 40–400 at a rate of 100 Hz. For peak findings, a signal-to-noise (S/N) cutoff was set to 100 and the NIST 2020 library was used for the identification of the analytes with a cut off of 700 match similarity. Analytes were aligned across all participants using Statistical-Compare tool contained within the Chromatof software. A chemical name was assigned if the analytes matched the following four criteria, (1) high mass spectral match, (2) group separation based on the structural formula (3) Retention Index match and (4) the extracted ion chromatogram (EIC) ionization patterns among all observed samples[24]. An analyte was not given a name if it did not match any of the four criteria. Possible contaminants were manually removed before further data analysis (Supplementary Table S2). Aligned data were exported for further data analysis.

Data reduction and normalization. A brief summary of our data cleaning and feature reduction process is shown in Supplementary Fig. S7. All statistical analyses were conducted in R 4.0.3 (R Core Team, Vienna, Austria)[61]. Data cleaning was performed as described[14,15]. In short, a frequency of observation (FOO) cut-offs of 80% was implemented to remove sparse features. The remaining features were normalized using probabilistic quotient normalization (PQN)[62], $\log_{10}$ transformed, and mean-centered. Missing values were imputed using a half-minimum approach[63].

Feature selection, machine learning and statistical analysis. A random forest (RF) regression model was implemented to select features clustering between individuals with different IR status. The important features were selected by considering all four samples collected from each participant. Combining male (n = 18) and female (n = 10) samples, a total of 112 (n = 28 × 4) samples were analyzed. The samples were then divided randomly, 3:1, for a training and validation dataset.

The RF model was built using a tenfold cross-validation (CV) scheme in the 'caret' package[64]. CV split the data into ten equal size pieces, building a model on nine of the ten pieces, and testing on the one remaining piece. The algorithm then leaves a different piece out and repeats this process for all pieces allowing for parameter tuning across the model, as well as estimating the model's generalizability by examining accuracy statistics across the left-out pieces. Features were selected by an elbow cutoff of the RF variables important measure. The variables were given an importance measure by mean decrease accuracy method, which leaves one variable and builds a model with the rest of the variables.

Orthogonal partial least-squares discriminant analysis (OPLS-DA) was performed using the "ropls" package[28]. Network analysis was performed in the MetScape 3[65] for Cytoscape[66] platform using the Kyoto Encyclopedia of Genes and Genomes (KEGG) ID and Human Metabolome Database (HMDB) ID. Additional statistical analysis was generated using Arsenal[67] package of R[61]. For Pearson correlations, data were $\log_{10}$-transformed and correlations were conducted in R[61] using 'ggpubr'[68]. The pairwise comparisons in boxplots were generated using the Wilcoxon non-parametric test and *p*-value were adjusted with the Holm-Bonferroni correction. Figure 3 created in R[61] using 'ggplot2'[69] and 'ggpubr'[68]. The ROC curve is generated in R[61] using the 'MLeval'[70]. Biorender was used for generating Fig. 5[71].

Data availability

Raw spectral data will be made available upon request.

Received: 2 August 2021; Accepted: 15 December 2021
Published online: 10 January 2022

References

1. Ormazabal, V. *et al.* Association between insulin resistance and the development of cardiovascular disease. *Cardiovasc. Diabetol.* **17**(1), 1–14 (2018).
2. Utzschneider, K. M. & Kahn, S. E. The role of insulin resistance in nonalcoholic fatty liver disease. *J. Clin. Endocrinol. Metab.* **91**(12), 4753–4761 (2006).
3. Navaneethan, S. D. *et al.* Adiposity, physical function, and their associations with insulin resistance, inflammation, and adipokines in CKD. *Am. J. Kidney Dis.* **77**(1), 44–55 (2020).
4. van der Aa MP, Fazeli Farsani S, Knibbe CAJ, de Boer A, van der Vorst MMJ. Population-based studies on the epidemiology of insulin resistance in children. *J. Diabetes Res.* **2015**, 362375 (2015).
5. CDC (Centers for Disease Control and Prevention). Prevalence of Obesity Among Adults and Youth: United States. 2017. https://www.cdc.gov/nchs/data/databriefs/db288.pdf.
6. Skinner, A. C., Ravanbakht, S. N., Skelton, J. A., Perrin, E. M. & Armstrong, S. C. Prevalence of obesity and severe obesity in US Children, 1999–2016. *Pediatrics.* **141**(3), e20173459 (2018).
7. Yaribeygi, H., Farrokhi, F. R., Butler, A. E. & Sahebkar, A. Insulin resistance: review of the underlying molecular mechanisms. *J. Cell. Physiol.* **234**(6), 8152–8161 (2019).
8. Samuel, V. T. & Shulman, G. I. The pathogenesis of insulin resistance: integrating signaling pathways and substrate flux. *J. Clin. Investig.* **126**(1), 12–22 (2016).
9. Hirosumi, J. *et al.* A central role for JNK in obesity and insulin resistance. *Nature* **420**(6913), 333–336 (2002).
10. Dandona, P., Aljada, A. & Bandyopadhyay, A. Inflammation: the link between insulin resistance, obesity and diabetes. *Trends Immunol.* **25**(1), 4–7 (2004).
11. Furukawa, S. *et al.* Increased oxidative stress in obesity and its impact on metabolic syndrome. *J. Clin. Investig.* **114**(12), 1752–1761 (2017).
12. ADA (American Diabetes Association) (2003) Standards of medical care for patients with diabetes mellitus. *Diabetes Care.* **26**(Suppl 1), s33–s50.
13. Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J. & Kivimäki, M. Prediabetes: A high-risk state for diabetes development. *The Lancet* **379**(9833), 2279–2290 (2012).
14. Khan, M. S. *et al. The early breath biomarkers of tuberculosis using model macaque monkey* (Breath Summit, Leicestershire, 2019).
15. Bobak, C. A. *et al.* Breath can discriminate tuberculosis from other lower respiratory illness in children. *Sci. Rep.* **11**(1), 1–9 (2020).
16. Tiele, A. *et al.* Breath-based non-invasive diagnosis of Alzheimer's disease: A pilot study. *J. Breath Res.* **14**(2), 026003 (2020).
17. Li, W. *et al.* A cross-sectional study of breath acetone based on diabetic metabolic disorders. *J. Breath Res.* **9**(1), 016005 (2015).
18. Krilaviciute, A. *et al.* Associations of diet and lifestyle factors with common volatile organic compounds in exhaled breath of average-risk individuals. *J. Breath Res.* **13**(2), 026006 (2019).
19. Bishop, A. C. *et al.* Nonhuman primate breath volatile organic compounds associate with developmental programming and cardiometabolic status. *J. Breath Res.* **12**(3), 036016 (2018).
20. Kuczmarski, R. J. *CDC Growth Charts for the United States: methods and development* (Centers for Disease Control and ,Prevention 2002).
20. Gutch, M., Kumar, S., Razi, S. M., Gupta, K. K. & Gupta, A. Assessment of insulin sensitivity/resistance. *Indian J. Endocrinol. Metabol.* **19**(1), 160–164 (2015).
22. Qu, H.-Q., Li, Q., Rentfro, A. R., Fisher-Hoch, S. P. & McCormick, J. B. The definition of insulin resistance using HOMA-IR for Americans of Mexican descent using machine learning. *PloS One.* **6**(6), e21041 (2011).
23. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recogn. Lett.* **31**(14), 2225–2236 (2010).
24. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**(3), 211–221 (2007).
25. Duncan, M. H., Singh, B. M., Wise, P. H., Carter, G. & Alaghband-Zadeh, J. A simple measure of insulin resistance. *Lancet* **346**(8967), 120–121 (1995).
26. Trygg, J. & Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **16**(3), 119–128 (2002).
27. Khan, M. S. *et al.* Multivariate analysis of PRISMA optimized TLC image for predicting antioxidant activity and identification of contributing compounds from Pereskia bleo. *Biomed. Chromatogr.* **29**(12), 1826–1833 (2015).
28. Thévenot, E. A., Roux, A., Xu, Y., Ezan, E. & Junot, C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J. Proteome Res.* **14**(8), 3322–3335 (2015).
29. Wold, S., Antti, H., Lindgren, F. & Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemometr. Intell. Lab. Syst.* **44**(1–2), 175–185 (1998).
30. Beccaria, M. *et al.* Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography—time of flight mass spectrometry and machine learning. *J. Chromatogr. B* **1074**, 46–50 (2018).
31. Cox, M. E. & Edelman, D. Tests for screening and diagnosis of type 2 diabetes. *Clin. Diabetes.* **27**(4), 132–138 (2009).
32. International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care* **32**(7), 1327–1334 (2009).
33. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2007).
34. Wishart, D. S. *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* **35**, D521-526 (2007).
35. Raman, M. *et al.* Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. *Clin. Gastroenterol. Hepatol.* **11**(7), 868-875.e863 (2013).
36. Dallinga, J. W. *et al.* Volatile organic compounds in exhaled breath as a diagnostic tool for asthma in children. *Clin. Exp. Allergy.* **40**(1), 68–76 (2010).
37. Garner, C. E. *et al.* Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. *FASEB J.* **21**(8), 1675–1688 (2007).
38. Wishart, D. S., Tzur, D., Knox, C., et al. Eicosane. *Human Metabolome Database (HMDB),* 2021. https://hmdb.ca/metabolites/HMDB0059909.
39. Wishart, D. S., Tzur, D., Knox, C., et al. Pentylbenzene. *Human Metabolome Database (HMBD),* 2021. https://hmdb.ca/metabolites/HMDB0059834#references.
40. Sobotka, P. A., Gupta, D. K., Lansky, D. M., Costanzo, M. R. & Zarling, E. J. Breath pentane is a marker of acute cardiac allograft rejection. *J. Heart Lung Transplant.* **13**(2), 224–229 (1994).

41. Weitz, Z., Birnbaum, A., Skosey, J., Sobotka, P. & Zarling, E. High breath pentane concentrations during acute myocardial infarction. *Lancet* **337**(8747), 933–935 (1991).
42. Phillips, M., Sabas, M. & Greenberg, J. Increased pentane and carbon disulfide in the breath of patients with schizophrenia. *J. Clin. Pathol.* **46**(9), 861–864 (1993).
43. Hietanen, E. *et al.* Diet and oxidative stress in breast, colon and prostate cancer patients: a case-control study. *Eur. J. Clin. Nutr.* **48**(8), 575–586 (1994).
44. Butterfield, D. A. *et al.* Structural and functional changes in proteins induced by free radical-mediated oxidative stress and protective action of the antioxidants N-tert-butyl-α-phenylnitrone and bitamin E. *Ann. N. Y. Acad. Sci.* **854**(1), 448–462 (1998).
45. Semenkovich, C. F. Insulin resistance and atherosclerosis. *J. Clin. Investig.* **116**(7), 1813–1822 (2006).
46. de Lacy, C. B. *et al.* A review of the volatiles from the healthy human body. *J. Breath Res.* **8**(1), 014001 (2014).
47. Silva, C. L., Passos, M. & Câmara, J. S. Solid phase microextraction, mass spectrometry and metabolomic approaches for detection of potential urinary cancer biomarkers—A powerful strategy for breast cancer diagnosis. *Talanta* **89**, 360–368 (2012).
48. De Preter, V., Van Staeyen, G., Esser, D., Rutgeerts, P. & Verbeke, K. Development of a screening method to determine the pattern of fermentation metabolites in faecal samples using on-line purge-and-trap gas chromatographic-mass spectrometric analysis. *J. Chromatogr. A* **1216**(9), 1476–1483 (2009).
49. Wang, S. *et al.* Gas chromatographic-mass spectrometric analysis of d-limonene in human plasma. *J. Pharm. Biomed. Anal.* **44**(5), 1095–1099 (2007).
50. Gahleitner, F., Guallar-Hoyas, C., Beardsmore, C. S., Pandya, H. C. & Thomas, C. P. Metabolomics pilot study to identify volatile organic compound markers of childhood asthma in exhaled breath. *Bioanalysis* **5**(18), 2239–2247 (2013).
51. Morisco, F. *et al.* Rapid "Breath-Print" of liver cirrhosis by proton transfer reaction time-of-flight mass spectrometry. A pilot study. *PLoS One* **8**(4), e59658 (2013).
52. Dadamio, J. *et al.* Breath biomarkers of liver cirrhosis. *J. Chromatogr. B* **905**, 17–22 (2012).
53. Friedman, M. I. *et al.* Limonene in expired lung air of patients with liver disease. *Digest. Dis. Sci.* **39**(8), 1672–1676 (1994).
54. Miyazawa, M., Shindo, M. & Shimada, T. Metabolism of (+)-and (−)-limonenes to respective carveols and perillyl alcohols by CYP2C9 and CYP2C19 in human liver microsomes. *Drug Metabol. Dispos.* **30**(5), 602–607 (2002).
55. Frye, R. F. *et al.* Liver disease selectively modulates cytochrome P450—mediated metabolism. *Clin. Pharmacol. Therap.* **80**(3), 235–245 (2006).
56. Moris, D. *et al.* The role of reactive oxygen species in myocardial redox signaling and regulation. *Ann. Transl. Med.* **5**(16), 324–324 (2017).
57. Ling, X. C. & Kuo, K.-L. Oxidative stress in chronic kidney disease. *Renal Replacem. Ther.* **4**(1), 53 (2018).
58. Ratcliffe, N. *et al.* A mechanistic study and review of volatile products from peroxidation of unsaturated fatty acids: An aid to understanding the origins of volatile organic compounds from the human body. *J. Breath Res.* **14**(3), 034001 (2020).
59. Cleveland, E., Bandy, A. & VanWagner, L. B. Diagnostic challenges of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. *Clin. Liver Dis.* **11**(4), 98 (2018).
60. Matthews, D. R. *et al.* Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28**(7), 412–419 (1985).
61. R Core Team. R: A Language and Environment for Statistical Computing (2019).
62. Noonan, M. J., Tinnesand, H. V. & Buesching, C. D. Normalizing gas-chromatography–mass spectrometry data: Method choice can alter biological inference. *Bioessays: News Rev. Mol. Cell. Dev. Biol.* **40**(6), e1700210 (2018).
63. Wei, R. *et al.* Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* **8**(1), 663 (2018).
64. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**(5), 1–26 (2008).
65. Karnovsky, A. *et al.* Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **28**(3), 373–380 (2012).
66. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003).
67. Heinzen, E., Sinnwell, J., Atkinson, E., Gunderson, T. & Dougherty, G. Arsenal: An arsenal of 'R' functions for large-scale statistical summaries. *R package version*. 1(0) (2018).
68. Kassambara, H. ggpubr: 'ggplot2' based publication ready plots (2020).
69. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
70. John, C. MLeval: Machine Learning Model Evaluation. *R package version* (2020).
71. BioRender.com. Adapted from "Regulation of Blood Glucose". 2021. https://app.biorender.com/Regulation-of-blood-glucose

Acknowledgements

We would like to acknowledge the Julee Carlton and Veronica Duran from the Health and Weight Management Clinic, Children's Hospital of San Antonio, San Antonio, TX who help collect the clinical measurement and facilitate sample collection. We also acknowledge the Texas Biomedical Research Institute for providing instrumental support of this project.

Author contributions

A.C.B., L.A.C., G.M.K. and S.C. conceptualized the project. A.C.B. and S.C. curated data. M.S.K. performed the formal data analysis. M.S.K. provided the original draft, which was reviewed and edited by A.C.B., L.A.C., G.M.K., S.C.

Funding

This work is funded by Healthy Babies Project, Texas Biomedical Research Institute, San Antonio, TX.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-04072-3.

Correspondence and requests for materials should be addressed to A.C.B.