Detection of diabetes from whole-body MRI using deep learning

Benedikt Dietz,[1] Jürgen Machann,[2,3,4] Vaibhav Agrawal,[5,6] Martin Heni,[3,4,7,8] Patrick Schwab,[9] Julia Dienes,[10] Steffen Reichert,[3,4,8] Andreas L. Birkenfeld,[3,4,8] Hans-Ulrich Häring,[3,4] Fritz Schick,[2,3,4] Norbert Stefan,[3,4,8] Andreas Fritsche,[3,4,8] Hubert Preissl,[3,4,8] Bernhard Schölkopf,[6] Stefan Bauer,[6,11] and Robert Wagner[3,4,8]

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. [2]Department of Radiology, Section on Experimental Radiology, Eberhard-Karls University Tübingen, Tübingen, Germany. [3]Institute for Diabetes Research and Metabolic Diseases, Helmholtz Center Munich, University of Tübingen, Tübingen, Germany. [4]German Center for Diabetes Research, Neuherberg, Germany. [5]Werner Siemens Imaging Center, Tübingen, Germany. [6]Max Planck Institute for Intelligent Systems, Department of Empirical Inference, Tübingen, Germany. [7]Department for Diagnostic Laboratory Medicine, Institute for Clinical Chemistry and Pathobiochemistry, University Hospital Tübingen, Tübingen, Germany. [8]Department of Internal Medicine, Division of Diabetology, Endocrinology and Nephrology, Eberhard-Karls University Tübingen, Tübingen, Germany. [9]Institute of Robotics and Intelligent Systems, ETH Zürich, Zürich, Switzerland. [10]Department of Gynecology and Obstetrics, University Hospital Tübingen, Tübingen, Germany. [11]Department of Intelligent Systems, KTH Stockholm, Stockholm, Sweden.

Obesity is one of the main drivers of type 2 diabetes, but it is not uniformly associated with the disease. The location of fat accumulation is critical for metabolic health. Specific patterns of body fat distribution, such as visceral fat, are closely related to insulin resistance. There might be further, hitherto unknown, features of body fat distribution that could additionally contribute to the disease. We used machine learning with dense convolutional neural networks to detect diabetes-related variables from 2371 T1-weighted whole-body MRI data sets. MRI was performed in participants undergoing metabolic screening with oral glucose tolerance tests. Models were trained for sex, age, BMI, insulin sensitivity, HbA1c, and prediabetes or incident diabetes. The results were compared with those of conventional models. The area under the receiver operating characteristic curve was 87% for the type 2 diabetes discrimination and 68% for prediabetes, both superior to conventional models. Mean absolute regression errors were comparable to those of conventional models. Heatmaps showed that lower visceral abdominal regions were critical in diabetes classification. Subphenotyping revealed a group with high future diabetes and microalbuminuria risk. Our results show that diabetes is detectable from whole-body MRI without additional data. Our technique of heatmap visualization identifies plausible anatomical regions and highlights the leading role of fat accumulation in the lower abdomen in diabetes pathogenesis.

Authorship note: BD and JM contributed equally to this work and are co–first authors. SB and RW contributed equally to this work and are co–senior authors.

Conflict of interest: The authors have declared that no conflict of interest exists

Submitted: December 30, 2020
Accepted: September 29, 2021
Published: September 30, 2021

Reference information: *JCI Insight*. 2021;6(21):e146999. https://doi.org/10.1172/jci.insight.146999.

Introduction

Currently around 500 million individuals worldwide have diabetes, which has a dramatically rising prevalence (1). Most diabetes cases are type 2 diabetes, which is a condition determined by a combination of reduced insulin action in the insulin target tissues, i.e., insulin resistance, and an insufficient compensation for this insulin resistance due to an impaired insulin secretion. Epidemiologically obesity is the main driver of insulin resistance (2), but excess body fat mass is neither a prerequisite nor a guarantee for insulin resistance. There are individuals who remain metabolically healthy despite being obese (3), while others develop insulin resistance despite normal body weight (4). This is because the distribution of fat within the human body crucially determines its metabolic role. Individuals with mostly deep abdominal and visceral fat accumulation are more prone to develop insulin resistance compared with individuals with mostly subcutaneous fat deposition (5). On the other hand, subcutaneous abdominal and thigh fat seem to act as protective triglyceride dumps in the body, which preserve insulin sensitivity by confining fat to metabolically inert body regions (6, 7). Specific fat compartments, such as fat depots near arteries, seem to play distinctive roles in the pathophysiology of insulin resistance, insulin secretion, and probably also the manifestation of metabolic complications (8). Some of these perivascular fat depots, such as those near the brachial artery, have been shown to associate with insulin resistance (9). Fat tissue in the renal sinus could contribute to nephropathy (10, 11). Furthermore, pancreatic fat deposition associates with reduced insulin secretion and may be involved in the decompensation of insulin secretion and thus in the pathogenesis of diabetes (12–14). However, it is challenging to assess the aggregate effect of fat distribution on diabetes. Simple anthropometric variables of fat distribution, such as waist and hip circumference, are not very accurate and provide only limited information on the distribution of fat over the body. A more accurate measurement of fat distribution can be achieved by whole-body T1-weighted MRI (15), which by design contrasts fat and water signals as a distribution of gray scale voxels over the body. It is possible to perform a segmentation of MR images to quantify specific predetermined regions, but this approach is laborious and could be biased by predefined areas of interest. We therefore investigated if 3-dimensional whole-body MR tomograms could be applied in an unbiased way to determine if the represented individual had diabetes at the time of the scan. Adequately trained machine-learning models have recently been very successful in associating high-dimensional data with medical labels (16). Although it is notoriously challenging to derive human-readable information on key patterns of machine-learning classifiers (17), we also aimed to extract information on the anatomical regions decisive in establishing these associations.

Results

Model training for diabetes and related labels. Sex classification converged to approximately 99% area under the receiver operating characteristic curve (AUROC) within the first approximately 25 epochs on the training set. All other labels tended to take considerably longer to converge, and individual performances varied with different network parameters. While sex classification seemed to be easily feasible for the network, the smoothed AUROC scores for the diabetes labels mostly peaked at approximately 85% for diabetes and approximately 70% for prediabetes and the extended diabetes label. As for the regression tasks, all of the predictive performances slightly varied. However, with different network parameters they generally converged to approximately 5%–15% mean absolute error (MAE) on the normalized training labels before starting to overfit. In all models, the lowest MAE has been reached for the estimation of BMI. The results of classifications and regressions achieved by the potentially novel dense convolutional neural networks in the optimal model are shown in Table 1.

The AUROC for classification of diabetes was 0.87. Receiver operator characteristics curves with and without stratification for sex are shown in Figure 1. As a comparison for the dense convolutional neural networks, conventional models were trained using body fat volumes determined by fat compartment segmentation.

A normalized MAE of 0.17 for age is equivalent to ±10 years average error. Similarly, a normalized BMI MAE of 0.07 represents an average error of ± 2kg/m^2, and 0.13 normalized HbA1c MAE equals ±0.4%. Finally, the insulin sensitivity error of 0.26 corresponds to an average error of ±10.2 AU, which represents the weakest regression performance among the continuous outcome variables.

Sensitivity analyses with different model setups. We also tested the diagnostic precision of different model setups for the labels sex, prediabetes, and diabetes and the diabetes label extended by impaired fasting glucose and impaired glucose tolerance (IFG+IGT) (Supplemental Table 2; supplemental material available online with this article; https://doi.org/10.1172/jci.insight.146999DS1). Part of these alternative models used images cropped to torso only or abdomen only (Supplemental Figure 3). As additional augmentation technique, we tested random zooming on the images. Detection of sex was not affected by the restricted images, but the diabetes and prediabetes labels reached lower AUROC values. Interestingly, diabetes detection was only slightly affected when using abdominal images, and these techniques had no relevant effect on the detection of diabetes with IFG+IGT. With model training rerun using only the first scan from each participant, we yielded lower AUROC for diabetes, but the prediabetes and the extended diabetes labels showed comparable diagnostic precision to the original data set.

Target-specific gradient maps. We computed attention heatmaps to acquire information about the reasoning behind predictions and to provide visualizations for further analyses. Comparison plots of heatmaps for 50 randomly selected samples for the detection of diabetes and insulin sensitivity are shown in Figure 2A The highlighted areas were assigned to prespecified anatomic regions by 3 clinicians who had expertise in the interpretation of medical imaging. For each of the 8 traits, the human experts rated 100 images presented in 3 dimensions (see example in Figure 2B). Interrater agreement was 76%.

Mean percentages for the predefined anatomical regions appearing in the heatmaps are shown in Table 2. The deep lower abdominal (visceral) region was associated with most cases of diabetes classification

Table 1. Model performance metrics

Model	Classification (AUROC)				Regression (MAE)			
	Sex	Diabetes	Prediabetes	Diabetes with IFG+IGT	Age	BMI	Matsuda-index	HbA1c
DCN	0.99	0.87	0.68	0.72	0.17	0.07	0.26	0.13
LR/KN		0.54	0.51	0.63	0.193	0.075	0.143	0.135
RF		0.51	0.59	0.56	0.19	0.07	0.142	0.14
SVM		0.60	0.51	0.42	0.192	0.07	0.141	0.134

Summary of model performance metrics for the final dense convolutional network (DCN) compared, with conventional classifiers (LR, linear regression; KN, K-neighbors classifier; RF, random forest; SVM, support vector machine) used as benchmarks. Model performance is shown as area under the receiver operator characteristics curve (AUROC) for classification and mean absolute error (MAE) for regression.

(89%). This region also seemed to be important, however, less prominent, for classifying diabetes with IFG+IGT cases (84%) and prediabetes cases (69%). For the classification of sex, the upper thorax region, including the breasts, played the major role (73% highlighted). Arms and upper legs were also important (67 and 61%, respectively). The upper leg region was also often highlighted in the heatmaps of the regression on BMI (64%) and insulin sensitivity (70%).

Clusters based on the embedding layer of MRI scans. Of the highly complex MRI scans, the training process generated vectors with 128 values per image. These are called embeddings and contain all relevant information from the scans. To investigate whether this information contained in whole-body MRI scans can be used for the prediction of metabolic features, we performed sex-stratified data-driven clustering on the embeddings. The clusters solely based on data from the embeddings delineated groups with different anthropometric and glycemic features (Figure 3, A and B, and Table 3). Furthermore, they also predicted future diabetes (Figure 3C*,* n = 586 with follow-up data, mean follow-up 4 ± 3.7 years, number of events = 48, $P < 0.0001$) and the development of microalbuminuria (Figure 3D*,* n = 550 with follow-up data, mean follow-up 4.3 ± 3.6 years, number of events = 95, $P = 0.004$). Anthropometric variables were different across clusters, but the association of cluster 4 with increased risk of diabetes and microalbuminuria was still significant after adjustment for sex, age and BMI ($P = 0.01$ and $P = 0.03$, respectively). In addition, the association of cluster 4 with these outcomes was not explained by differences in baseline glycemia ($P = 0.02$ for future diabetes after adjusting for baseline glycated hemoglobin) or baseline urinary albumin-to-creatinine ratio (uACR) ($P = 0.04$ for future microalbuminuria after adjusting for baseline uACR, n = 441, events = 76).

Discussion

Here, we tested if presence of diabetes can be identified from specific patterns of body fat distribution assessed by MRI. With a machine-learning approach on more than 2000 whole-body MRI data sets, we produced excellent classification results that were superior to those from state-of-the-art statistical modeling of body fat compartment volumes. These results prove that diabetes is detectable with deep learning from imaging data. Accordingly, 3-dimensional MRI images harbor patterns for a sufficiently good discrimination of patients with and without diabetes. Of note, the images were normalized for body length to target a classification based on fat distribution rather than body height.

In an empirical approach to look into the black box of machine learning, we applied human expert rating of heatmaps representing regions important for classification and regression. The interpretation of these heatmaps suggests that deep lower abdominal fat was most critical for the detection of diabetes by the machine (89% of diabetes heatmaps contained these areas). Furthermore, we also detected diabetes-related signals in the upper legs (66%), the arms (51%), and the neck region (51%).

Visceral fat, in contrast to subcutaneous fat, has been previously identified as an important predictor of insulin resistance, the failure to respond to lifestyle intervention and the future manifestation of diabetes (18, 19). Interestingly and somewhat unexpectedly, structures in the lower rather than upper abdomen turned out to be the most important topographic areas in our analyses. These results suggest that not all visceral adipocytes have the same impact on metabolism and point toward a heterogeneity with metabolically unfavorable fat enriched in the lower part. Highlights in the neck region could be linked to insulin resistance.

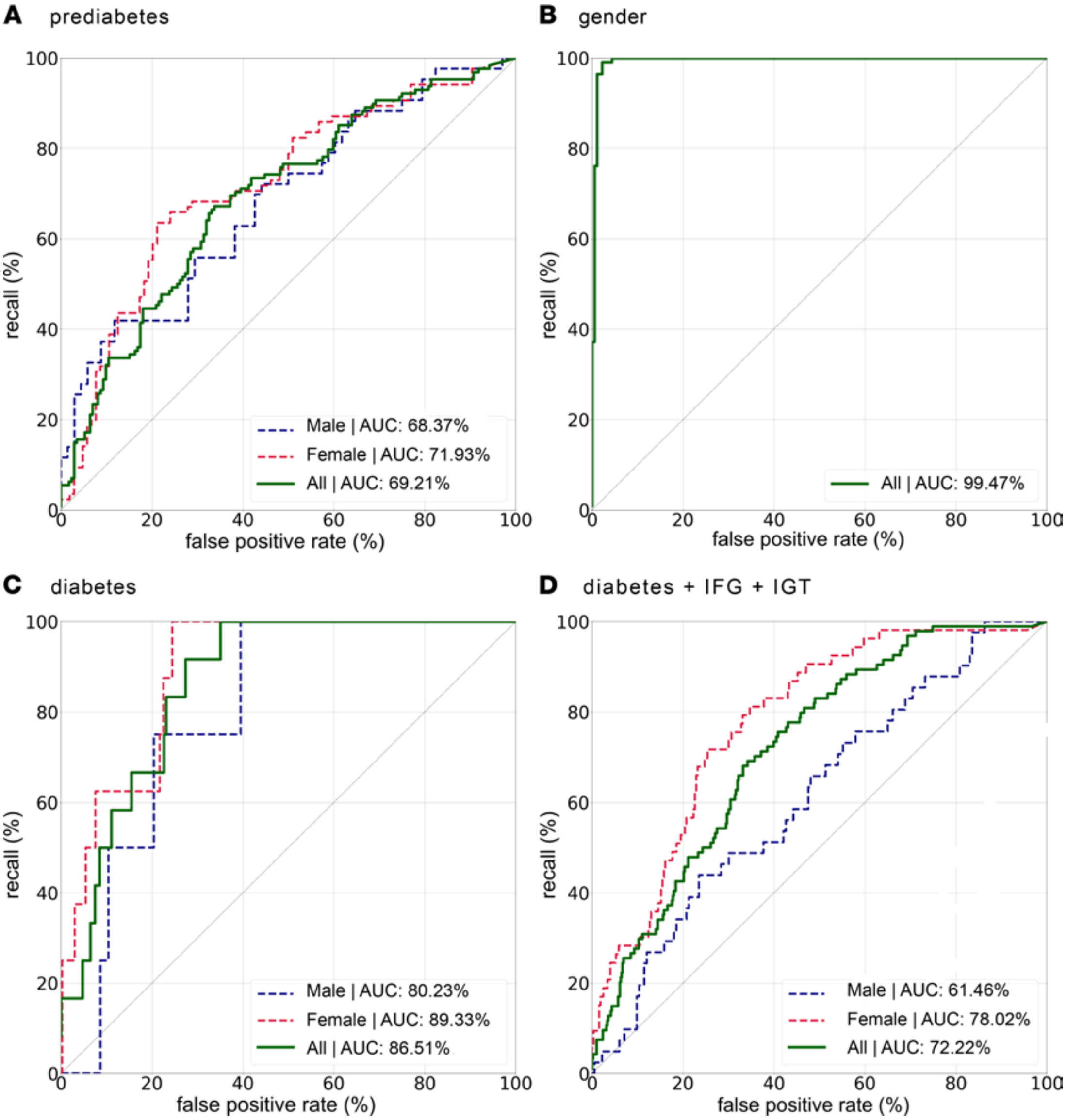

Figure 1. Diagnostic accuracy of the machine-learning classifiers. Receiver operating characteristic curves for the detection of prediabetes (**A**), sex (**B**), diabetes (**C**), and diabetes with impaired fasting glucose and impaired glucose tolerance (diabetes+ IFG + IGT, **D**) by the dense convolutional neural network.

Indeed, interscapular fat had been shown as an important independent marker of insulin resistance (20). Its importance in diabetes pathology is corroborated by our current hypothesis-free approach. The arms and upper leg regions are unforeseen hot spots, because they mostly comprise subcutaneous, metabolically inert fat depots. For estimation of BMI, arms and upper legs were the leading anatomic regions and might therefore predominantly represent general obesity. Of note, the upper leg region was also leading in the regression for insulin sensitivity (featured in 70% of heatmaps). Insulin resistance is the major body fat–derived metabolic factor in the pathogenesis of diabetes. However, insulin resistance is by itself not sufficient to cause diabetes (21). Diabetes only manifests if there is an additional disruption of pancreatic insulin secretion. Accordingly, we see a clear dissociation of the diabetes- and insulin resistance–related regions in our heatmaps. Unexpectedly, the deep lower abdomen differentiated diabetes from solitary insulin resistance. As the pancreas is not located in this area, our results suggest that additional biological signals that originate from the lower abdomen and target pancreatic islets could impair insulin release. The pancreas probably did not emerge directly in our machine-learning approach, as diabetes-related changes only occur in the islets that represent a minute proportion of the entire organ and can therefore hardly be detected by imaging. Another organ with known important contribution to diabetes, the liver, could correspond

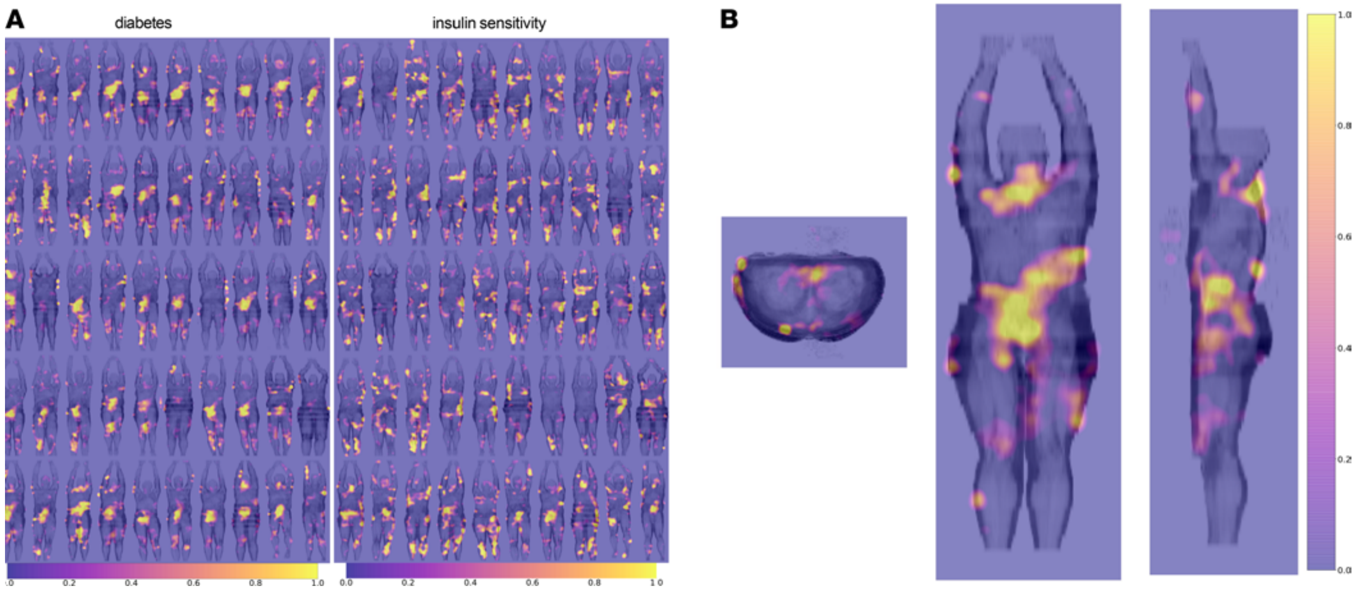

Figure 2. Gradient maps visualizing voxels with large influence on the classification/regression outcome. (**A**) Gradient maps for diabetes and insulin sensitivity, computed for 50, randomly selected, persons with prediabetes. The body scans, as well as the gradient maps, were averaged along the coronal projection to generate 2-dimensional representations. (**B**) An example gradient heatmap for the diabetes label in 3 projections. For assignment of gradient maps to body regions by raters, similar 3-dimensional gradient map representations, were used.

to highlighted areas in the deep right upper abdomen, appearing in 64% of diabetes with IFG+IGT classifiers. Accordingly, there was considerably less highlighting in the left upper abdomen, i.e., outside of the liver (13%). We have previously shown that a disruptive organ crosstalk among fat, liver, and pancreatic β cells could contribute to a deterioration of insulin secretion (13). Our findings about diabetes-related features of whole-body MRI stress the multiorgan nature of diabetes pathology.

The cluster analysis of the embeddings generated by machine learning from MRI scans shows a clear discrimination of 4 groups. This is not just a sole clustering of random image information but has biological meaning, as the clusters delineate different demographic and metabolic entities. As one of the identified subphenotypes was also associated with future diabetes and microalbuminuria, the most important early marker of diabetic kidney disease, the information content of the MRI images is also highly relevant for prediction of glycemic deterioration and a diabetes complication.

The results of sensitivity analyses using images restricted to the abdominal region suggest that future investigations could mainly focus on abdominal MR imaging. Using state-of-the art MR imaging techniques, higher resolution and faster acquisition times could be yielded, which might contribute to a better understanding of abdominal anatomy to diabetes pathology.

Our work has some limitations. Although different scanners were used to produce our data, this was a single-center study without external replication. Despite splitting our data into training, test, and replication sets, how our classifier will perform on data from different centers still needs to be tested. Furthermore, some of the data were repeated measurements in the same person, which we were unable to explicitly address in the machine-learning procedure. However, labels were updated concurrently (from oral glucose tolerance tests [OGTTs] performed at the time of each MRI scan), linking the respective metabolic status to anatomic patterns, and sensitivity analyses using a subset of the data without repeated measurements show comparable results for some labels. Capturing the intuition behind machine learning is still challenging, and there is no generally accepted method for this. To our knowledge, this is the first work to utilize 3-dimensional MRI whole-body scans for the analysis of diabetes and related features as well as to investigate a combination of heatmaps and their assignment to anatomic hot spots by human experts.

In summary, our work provides evidence that machine learning can classify diabetes from whole-body MRI. Diabetes, but not insulin sensitivity, was particularly associated with the features of the deep lower abdomen. This points toward considerable heterogeneity in the metabolic role of fat located in different parts of the visceral adipose tissue that has not been described so far. Further research is warranted on underlying molecular pathways that could represent important novel pathomechanisms in diabetes development.

Table 2. Human expert classification of gradient heatmaps generated from the output nodes of the machine-learning classifier network

		Arms	Head	Lower legs	Upper legs	Lower abdominal visceral region	Mediastinum	Neck	Thighs, lower abdominal s.c. regions	Abdomen, upper left visceral region	Abdomen, upper right visceral region	Breasts, upper thorax
1	Sex	67%	28%	51%	61%	42%	10%	12%	45%	5%	17%	73%
2	Age	46%	5%	34%	63%	62%	1%	46%	50%	11%	31%	13%
3	BMI	62%	2%	22%	64%	19%	9%	29%	60%	44%	36%	39%
4	Diabetes	51%	14%	28%	66%	89%	4%	50%	27%	14%	42%	12%
5	Diabetes and IFG+IGT	44%	8%	16%	49%	84%	2%	51%	22%	13%	64%	7%
6	Prediabetes	49%	4%	31%	53%	69%	1%	50%	20%	12%	59%	12%
7	HbA1c	64%	24%	36%	40%	62%	4%	42%	30%	5%	38%	23%
8	Insulin sensitivity	45%	38%	61%	70%	22%	9%	6%	51%	4%	14%	51%

The values show the mean percentage of an anatomical region appearing in the heatmaps of a given trait.

Methods

Study population and MRI

MRI was performed on individuals who participated in metabolic screenings within the framework of multiple studies performed at the Department of Medicine IV, University Hospital Tübingen. In most of these studies, participants were generally healthy, without known type 2 diabetes but with an increased risk for the disease. This was defined as either family history of type 2 diabetes, BMI of greater than 27 kg/m^2, or known prediabetes. The participants came fasted to the study facility and underwent whole-body MRI in the early morning, which was followed by a health examination, assessment of medical history, and an OGTT. OGTT does not only allow assessment of insulin sensitivity and glucose tolerance, but it is also the gold-standard detection of diabetes. Follow-up data for diabetes incidence and assessment of complications such as microalbuminuria was available for a subset of the subjects.

During the MRI, subjects were lying in a prone position with extended arms, and images were recorded from fingers to toes. A T1-weighted fast spin–echo technique with a slice thickness and an interslice gap of 10 mm was applied, allowing discrimination of adipose and lean tissue due to inherent different longitudinal relaxation times T1. The patient table was shifted by 10 cm after each measurement (12 seconds each). Total acquisition time, including 1 rearrangement, was 20–25 minutes (15).

Data acquisition

Within the whole-body MRI scanning procedure, 90–120 parallel transverse slices were generated per participant, depending on body height. We quantified total adipose tissue volume, visceral adipose tissue volume, and upper extremities adipose tissue volume from these images for the benchmark models, using methods described previously (22). MR voxel arrays were provided in the Digital Imaging and Communications in Medicine file format. The original data set consisted of 2555 whole-body scans of 1080 participants, as some had been scanned multiple times. The number of MRI scans involved at different steps of the analysis is shown in Supplemental Figure 1. We used 8 ground truth labels. Four of these were binary labels; 1 was for sex, and the remaining 3 were for different diabetes definitions, including diabetes (Dα), prediabetes (Dβ), and a third definition, Dγ, that denoted diabetes cases extended with participants having concomitant IFG and IGT. In addition to the 4 binary labels, we used age (years), BMI (kg/m^2), insulin sensitivity (determined with the Matsuda index; ref. 23), and glycated hemoglobin (HbA1c) (%) as target labels for the network. An overview of the characteristics of participants and the labels is provided in Supplemental Table 1. Laboratory measurements were performed as previously described (24). The diagnosis of diabetes was established by one of the following: fasting glucose >7.0 mmol/l, postchallenge glucose ≥11.1 mmol/l or a glycated hemoglobin ≥48 mmol/mol. Microalbuminuria was established by a uACR ≥30 mg/g creatinine.

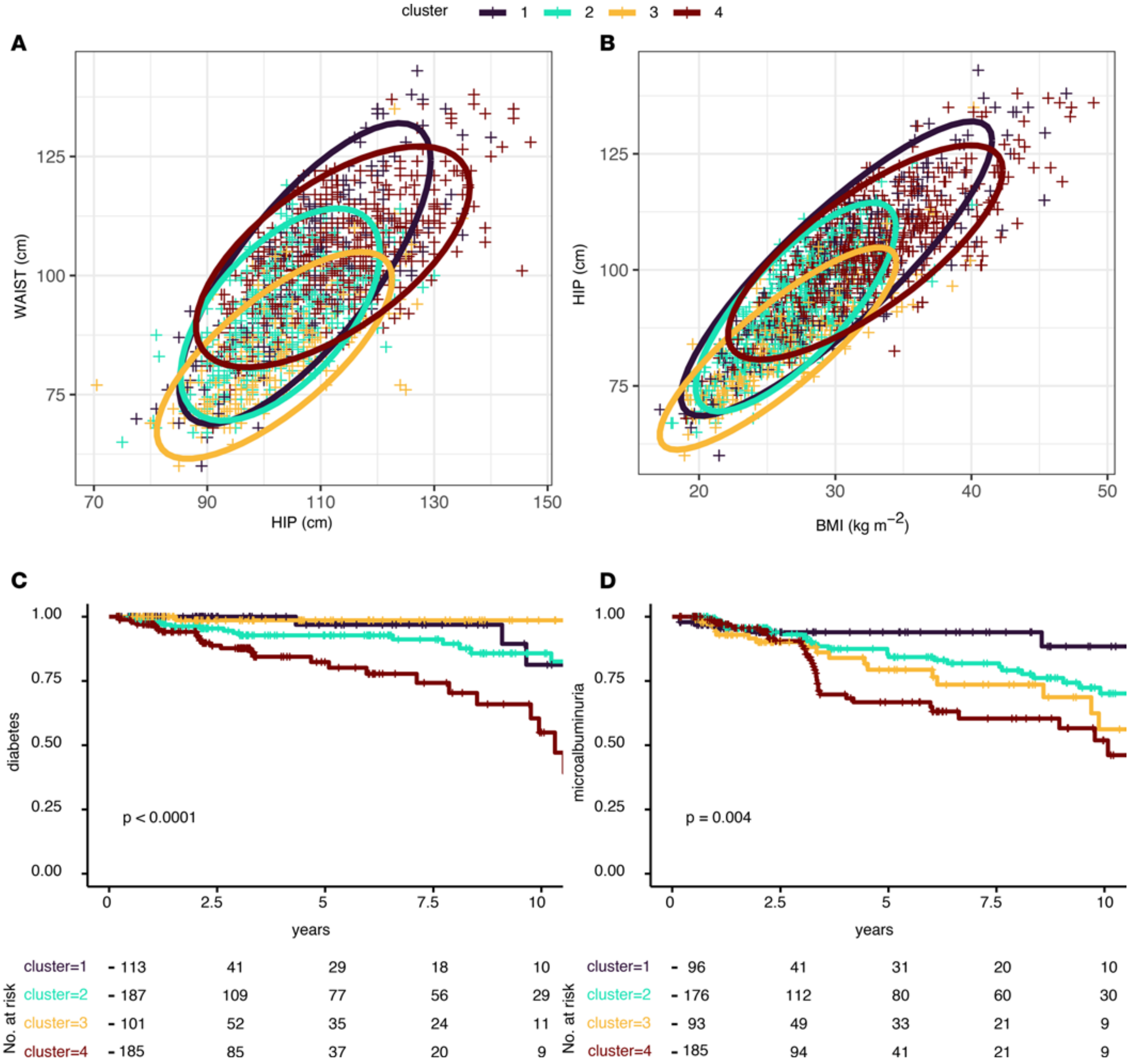

Figure 3. Partitioning of MRI images. Data-driven clustering was performed from embedding layers, which are numeric representations of MRI scans generated during inference (n = 2048). The MRI-based clusters have different distributions of waist and hip circumference (**A**) and BMI (**B**). For the participants with follow-up data, these MRI-data based clusters also define different risk profiles not only for new-onset diabetes (n = 586) (**C**), but also for the diabetes complication microalbuminuria (n = 550) (**D**). Diagrams showing incidence-free survival were compared with log-rank tests.

Data preprocessing

Shape normalization. As mentioned, MRI scans were acquired by generating image slices along the body's horizontal plane. Unlike the slice dimension, the number of slices varied according to body height. The most frequent number of slices was 95. All scans with different heights were linearly interpolated along the vertical body axis to produce volumes with normalized dimensions of 95 × 150 × 250 voxels. The voxel grid resolution of the 2 horizontal axes (considering a standing person) was considerably higher than the resolution along the body height axis, with negligible differences in coronal (z axis) scaling due to the aforementioned interpolation. We did not correct the lower resolution on the axis corresponding to the body height with further interpolations. However, we downsampled the standardized voxel grids for computational efficiency to their final dimension of 85 × 110 × 135 voxels.

Table 3. Characteristics of clusters generated from embedding layer representation of the MRI scans

	Clusters				
	1	2	3	4	*P* value
n	366	663	341	678	
Male sex (%)	251 (68.6)	280 (42.2)	40 (11.7)	217 (32.0)	<0.001
Female sex (%)	115 (31.4)	383 (57.8)	301 (88.3)	461 (68.0)	
Age in yrs (mean [SD])	42.94 (11.90)	55.14 (10.62)	37.72 (10.86)	55.66 (9.95)	<0.001
BMI (kg/m^2) (mean [SD])	30.44 (5.56)	27.13 (3.43)	26.23 (4.18)	32.47 (4.82)	<0.001
Waist circumference (cm) (mean [SD])	100.33 (14.79)	91.83 (10.14)	84.36 (10.63)	104.23 (11.08)	<0.001
Hip circumference (cm) (mean [SD])	107.73 (10.98)	103.11 (8.28)	102.33 (10.01)	112.51 (11.13)	<0.001
Total adipose tissue MRI (L) (mean [SD])	35.61 (14.44)	28.62 (8.71)	28.57 (9.50)	41.35 (12.57)	<0.001
s.c. adipose tissue MRI (L) (mean [SD])	11.91 (6.69)	8.96 (3.98)	8.56 (3.93)	14.96 (6.36)	<0.001
Visceral adipose tissue MRI (L) (mean [SD])	4.34 (2.57)	3.43 (1.84)	1.53 (1.25)	5.01 (2.03)	<0.001
s.c.-to-visceral adipose ratio (mean [SD])	3.44 (2.25)	3.25 (1.94)	7.23 (3.62)	3.45 (1.96)	<0.001
Visceral adipose % of total (mean [SD])	0.12 (0.06)	0.12 (0.06)	0.05 (0.04)	0.13 (0.06)	<0.001
% Liver fat content (mean [SD])	6.02 (5.68)	4.40 (4.32)	2.25 (2.80)	9.55 (7.67)	<0.001
% Fatty liver disease (mean [SD])	138 (37.9)	165 (25.4)	26 (7.8)	381 (57.0)	<0.001
Systolic blood pressure (mmHg) (mean [SD])	131.19 (16.20)	129.95 (16.32)	120.52 (13.73)	137.71 (16.60)	<0.001
Diastolic blood pressure (mmHg) (mean [SD])	83.25 (11.87)	81.38 (10.95)	76.73 (10.09)	87.62 (11.28)	<0.001
Heart rate (bpm) (mean [SD])	68.57 (12.57)	67.98 (10.49)	69.80 (10.61)	71.64 (10.24)	<0.001
Fasting glucose (mmol/l) (mean [SD])	5.30 (0.51)	5.43 (0.52)	5.02 (0.40)	5.65 (0.61)	<0.001
Postchallenge glucose (mmol/l) (mean [SD])	6.47 (1.53)	6.95 (1.77)	6.13 (1.46)	7.61 (2.10)	<0.001
Glycemic category					<0.001
NGT (%)	237 (64.8)	348 (52.5)	272 (79.8)	246 (36.3)	
IFG (%)	63 (17.2)	136 (20.5)	29 (8.5)	147 (21.7)	
IGT (%)	38 (10.4)	90 (13.6)	37 (10.9)	98 (14.5)	
IFG+IGT (%)	25 (6.8)	57 (8.6)	2 (0.6)	108 (15.9)	
DIA (%)	3 (0.8)	32 (4.8)	1 (0.3)	79 (11.7)	
Glycated hemoglobin (mmol/mol) (mean [SD])	36.36 (3.59)	38.52 (4.00)	35.29 (3.49)	39.86 (4.27)	<0.001
Triglycerides (mmol/l) (mean [SD])	1.36 (0.80)	1.30 (0.73)	1.03 (0.56)	1.70 (1.51)	<0.001
Insulin sensitivity (Matsuda, arbitrary units) (mean [SD])	14.80 (10.17)	15.63 (8.75)	19.50 (9.29)	8.91 (5.07)	<0.001
Fasting insulin (pmol/l) (mean [SD])	64.55 (41.69)	51.20 (32.15)	46.90 (26.86)	86.24 (47.78)	<0.001
Insulinogenic index (arbitrary units) (mean [SD])	145.93 (137.84)	96.31 (83.69)	139.64 (184.45)	130.14 (99.14)	<0.001
Disposition index (arbitrary units) (mean [SD])	1781.41 (3409.53)	1287.56 (1252.09)	2834.08 (6596.89)	990.99 (852.42)	<0.001
C-reactive protein (mg/dl) (mean [SD])	0.20 (0.30)	0.18 (0.27)	0.23 (0.35)	0.41 (0.50)	<0.001
Cholesterol (mmol/l) (mean [SD])	4.96 (0.94)	5.22 (0.93)	4.58 (0.86)	5.28 (1.01)	<0.001
LDL (mmol/l) (mean [SD])	3.07 (0.85)	3.18 (0.78)	2.66 (0.79)	3.21 (0.85)	<0.001
HDL (mmol/l) (mean [SD])	1.27 (0.33)	1.43 (0.34)	1.48 (0.32)	1.34 (0.32)	<0.001
Aspartate aminotransferase (U/l) (mean [SD])	26.85 (20.34)	23.68 (8.16)	20.76 (8.40)	25.17 (9.88)	<0.001
Alanine aminotransferase (U/l) (mean [SD])	32.69 (18.81)	24.31 (11.90)	19.91 (12.18)	30.38 (17.05)	<0.001
γ-Glutamyl transferase (U/l) (mean [SD])	30.63 (21.11)	25.55 (24.48)	15.94 (15.55)	34.30 (29.85)	<0.001
Serum creatinine (mg/dl) (mean [SD])	0.85 (0.16)	0.83 (0.16)	0.78 (0.15)	0.78 (0.16)	<0.001
Urine albumin-creatinine ratio (mean [SD])	24.48 (107.36)	18.50 (46.32)	21.20 (37.48)	20.69 (37.70)	0.558

Anthropometric, clinical, and laboratory characteristics of clusters generated from embedding layer representation of the MRI scans. DIA, diabetes; NGT, normal glucose tolerance; IFG, impaired fasting glucose; IGT, impaired glucose tolerance.

Voxel value normalization. Voxels that did not belong to the body (e.g., caused by motion artifacts inherent to MRI) were identified using value distributions and set to 0. We standardized body voxel values to have a mean of 0 and a SD of 1 and truncated and subsequently shifted the distribution to strictly positive values to keep the distinction from the surrounding air. We transformed all scan samples equally.

Labels

A total of 8 outcome label variables were used as described. Samples with missing labels were excluded, continuous variables were normalized to a range between 0 and 1. Outliers were removed using the isolation forest algorithm and fitted on a subset of the medical features, namely insulin sensitivity, BMI, HbA1c, as well as total adipose tissue estimate (25).

Data partitioning

In order to assess the generalization capabilities of our models, we applied a stratified random split to separate the entire data set into training (70%), validation (15%), and test (15% of all data) folds. The folds were stratified by BMI, insulin sensitivity, and diabetes. Our stratification algorithm required that each multivariate stratum contains more than 1 sample (Supplemental Figure 1).

Augmentation

To increase model robustness, the input volumes were augmented in several ways. Additional 0 padding was added to each dimension, increasing the size of the 3-dimensional image array. To augment the training samples, the degree of padding was dynamically adjusted at random during training. Furthermore, a series of rotations were randomly performed on each input array as well as addition of Gaussian noise to all body voxels (Supplemental Figure 2). For testing and validation, the body volumes were centered, and no rotation or noise was applied. We also performed sensitivity analyses using additional random zooming of the images during the training. Furthermore, we tested the training on restricted images cropped to the torso and abdominal area (Supplemental Figure 3, A and B).

Model architecture

Densely connected convolutional layers. We built the network in accordance to the DenseNet architecture (26). The 3-dimensional input volumes were fed to the input layer in batches consisting of 8 samples each. The initial layer consisted of a fully connected convolutional layer with a kernel size of 5 and 8 convolutional filters. The initial convolutional layer was followed by a batch normalization layer (27, 28). The dimensions of the intermediary feature maps were subsequently downsampled using a pooling layer to improve computation efficiency. The output was fed to the first dense block. Alternating dense blocks and transition layers were sequentially added to process the input.

Following the final transition layer, the activation maps were flattened to a 1-dimensional array and passed to 3 sequential densely connected layers with dropout. The output of the dense layers had the dimension 1 × 128 units and was referred to as the embedding layer, embodying low-dimensional representations of MRI voxels as "learned" by the neural network. The embedding layer was used for the prediction of the desired target labels and for unsupervised clustering analysis. For the prediction of the output nodes, subsequent dense layers were added to the embedding layer. A schematic of the entire model is provided in Supplemental Figure 2.

Gradient heatmaps. Predictions for the various target variables were directly represented by the output nodes. Differentiating these with respect to previous convolutional layers yields pixel-wise gradients. The chosen approach for heatmap generation was gradients × input (27); hence, we computed the gradients of the individual outputs with respect to the image input. Differentiating resulted in target-specific gradients of the same dimension as the input scans. Gradient heat maps were assigned to anatomical regions by 3 human medical experts (experienced physicians working in hospitals) who had experience in the evaluation and interpretation of medical imaging. All expert raters were blinded to the subject characteristics as well as to the trait they rated. Results were averaged for the 3 raters.

Hyperparameters. We used a growth factor kGrowth = 18. The initial convolution layer, prior to the first dense block had a kernel size of 5 × 5 × 5 voxels, and it generated 4 activation maps. We chose to evaluate the model with 3 dense blocks and 3 subsequent fully connected layers, downsampling the flattened representation to 512, 256, and 128, respectively. With the sole exception of the final regression output layers, all activation functions throughout the network were exponential linear units (28). We chose the initialization proposed by He et al. (29) for all kernel weights. Adam (30) was used as optimizer with an initial learning rate of 10-4. All other hyperparameters of the optimizer were kept at their Tensorflow (31) implementation defaults. The learning rate is adapted during training through a Tensorflow variable and has a cyclic, exponentially decay.

Training. Network training was performed on a Nvidia Tesla V100-PCIE 32GB GPU, using the CUDA framework (32). The network was trained for a maximum of 250 epochs with a batch size of 8, due to the considerable memory requirements of our 3-dimensional voxel grids. The network converged after approximately 2–3 days, depending on network depth, i.e., number of trainable parameters as well as batch size and other hyperparameters.

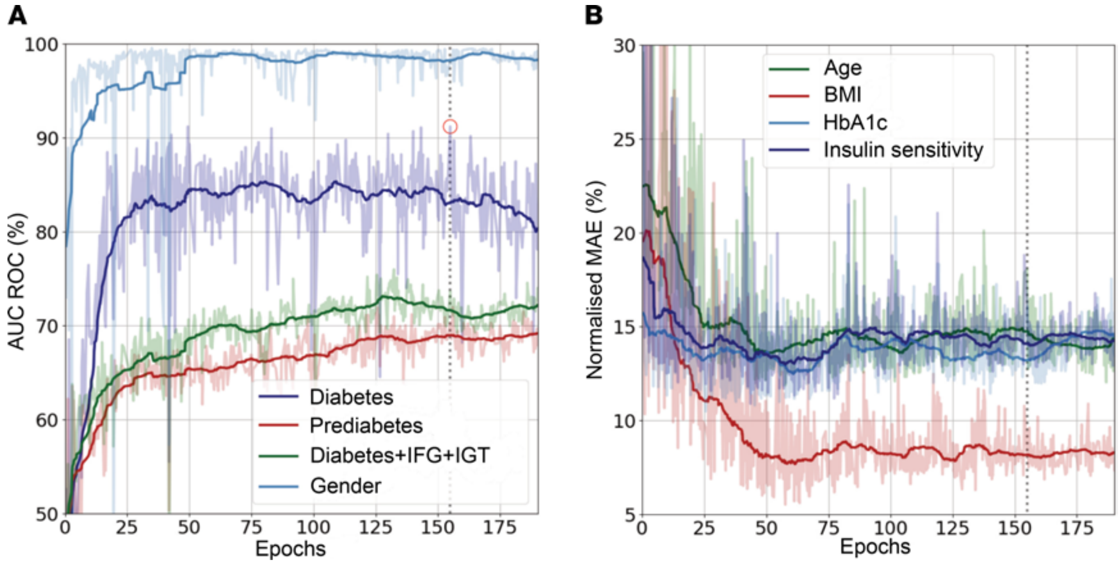

Figure 4. Training metrics and model selection. Performance of models in subsequent computation runs for classifications (area under the receiver operating characteristic (ROC) curves (**A**) and regressions (normalized mean absolute error) (**B**). The circled point in **A** indicates the highest achieved ROC for diabetes in the validation set.

Model selection. We frequently evaluated the network's performance and selected the model according to the highest diabetes (Dα) AUROC score on the validation set.

We first summarized the training progress, using the AUROC and MAE metric on the validation set over all trained epochs (Figure 4). Metrics were generally computed on the test set with the exception of diabetes (Dα) and prediabetes (Dβ). Due to the critically low number of positives for these labels, we chose to concatenate the data sets for testing and validation to compute the classification performance metrics. To assess the performance of our approach, benchmark models were computed using linear regression and k-nearest neighbor classifier, random forest models, and support vector machines for comparison. Body fat compartment volumes that have been segmented from MR images (total, visceral, and upper limb adipose tissue) were used as model inputs.

Postprocessing. The output of the model consisted of a set of predictions for each sample in addition to the respective gradient maps as well as its embedding space representation.

We used gradient maps to compute target specific heat maps, using the gradients×input method, proposed by Shrikumar et al. (27). The individual output nodes were differentiated with respect to the input to produce 3-dimensional feature-specific gradient maps. For visualization, the gradient maps were postprocessed using Gaussian filters in addition to contrast enhancements and averaging to 2 dimensions. Furthermore, for the classification nodes, only the output node corresponding to the correct label was considered. In other words, for a patient with label female, only the gradient that increased the female probability prediction was used for visualization. All positive gradients were considered for the regression tasks.

Data availability

All requests for data and materials are promptly reviewed by the Data Access Steering Committee of the Institute of Diabetes and Metabolic Research, Tübingen, Germany, to verify if the request is subject to any intellectual property or confidentiality obligations. Individual level data may be subject to confidentiality. Any data and materials that can be shared will be released via a material transfer agreement.

Statistics

Cluster analysis was performed on the embeddings using partitioning around medoids with Gower's distances. The optimal number of clusters was selected using average Silhouette widths. To investigate the robustness of the clusters, we performed a bootstrap validation showing a Jaccard similarity index of 0.73 over all clusters. A subset of the participants was tested during follow-up visits for incident diabetes (n = 586) and microalbuminuria (n = 550). Comparison of risks was performed using Kaplan-Meier diagrams and log-rank tests. Further analyses with adjustments for potential confounders were performed with proportional hazards models. Proportional hazards assumptions were tested by visualizing Schoenfeld residuals.

Study approval

The studies were carried out with written informed consent from all subjects in accordance with the Declaration of Helsinki. All protocols were approved by the ethics committee of the University of Tübingen.

Author contributions

JM and RW contributed to data acquisition, analyzed the data, and wrote the manuscript. BD, VA, and SB analyzed the data and wrote the manuscript. JD and SR contributed to data acquisition and the interpretation of data and edited the manuscript. MH, HUH, and AF contributed to the concept of the work and edited the manuscript. PS, ALB, FS, NS, HP, and BS contributed to the interpretation of data and edited the manuscript. All authors have reviewed the manuscript. BD constructed the first machine-learning models and is therefore listed first among the co–first authors.

Acknowledgments

This study was supported in parts by a grant (01GI0925) from the Federal Ministry of Education and Research to the German Center for Diabetes Research and from the state of Baden-Württemberg (32-5400/58/2, Forum Gesundheitsstandort Baden-Württemberg). We thank all the research volunteers for their participation. We gratefully acknowledge the excellent technical assistance of the Diabetes Research Unit Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tübingen. We thank Benjamin A. Jaghutriz, University Hospital Tübingen, for his help in heatmap assignment.

Address correspondence to: Robert Wagner, Otfried-Müller-Str 10, 72076 Tübingen, Germany. Phone: 49.7071.2968967; Email: robert.wagner@med.uni-tuebingen.de.

1. Chatterjee S, et al. Type 2 diabetes. *Lancet*. 2017;389(10085):2239–2251.
2. Kahn BB, Flier JS. Obesity and insulin resistance. *J Clin Invest*. 2000;106(4):473–481.
3. Stefan N, et al. Identification and characterization of metabolically benign obesity in humans. *Arch Intern Med*. 2008;168(15):1609–1616.
4. Stefan N, et al. Causes, characteristics, and consequences of metabolically unhealthy normal weight in humans. *Cell Metab*. 2017;26(2):292–300.
5. Goodpaster BH, et al. Subcutaneous abdominal fat and thigh muscle composition predict insulin sensitivity independently of visceral fat. *Diabetes*. 1997;46(10):1579–1585.
6. Stefan N, et al. Metabolically healthy obesity: the low-hanging fruit in obesity treatment? *Lancet Diabetes Endocrinol*. 2018;6(3):249–258.
7. McLaughlin T, et al. Preferential fat deposition in subcutaneous versus visceral depots is associated with insulin sensitivity. *J Clin Endocrinol Metab*. 2011;96(11):E1756–E1760.
8. Siegel-Axel DI, Häring H-U. Perivascular adipose tissue: an unique fat compartment relevant for the cardiometabolic syndrome. *Rev Endocr Metab Disord*. 2016;17(1):51–60.
9. Rittig K, et al. Perivascular fatty tissue at the brachial artery is linked to insulin resistance but not to local endothelial dysfunction. *Diabetologia*. 2008;51(11):2093–2099.
10. Wagner R, et al. Exercise-induced albuminuria is associated with perivascular renal sinus fat in individuals at increased risk of type 2 diabetes. *Diabetologia*. 2012;55(7):2054–2058.
11. Wagner R, et al. The protective effect of human renal sinus fat on glomerular cells is reversed by the hepatokine fetuin-A. *Sci Rep*. 2017;7(1):2261.
12. Heni M, et al. Pancreatic fat is negatively associated with insulin secretion in individuals with impaired fasting glucose and/or impaired glucose tolerance: a nuclear magnetic resonance study. *Diabetes Metab Res Rev*. 2010;26(3):200–205.
13. Gerst F, et al. Metabolic crosstalk between fatty pancreas and fatty liver: effects on local inflammation and insulin secretion. *Diabetologia*. 2017;60(11):2240–2251.
14. Gerst F, et al. What role do fat cells play in pancreatic tissue? *Mol Metab*. 2019;25:1–10.
15. Machann J, et al. Follow-up whole-body assessment of adipose tissue compartments during a lifestyle intervention in a large cohort at increased risk for type 2 diabetes. *Radiology*. 2010;257(2):353–363.
16. Rajkomar A, et al. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–1358.
17. Holm EA. In defense of the black box. *Science*. 2019;364(6435):26–27.
18. Hayashi T, et al. Visceral adiposity, not abdominal subcutaneous fat area, is associated with an increase in future insulin resistance in Japanese Americans. *Diabetes*. 2008;57(5):1269–1275.
19. Thamer C, et al. High visceral fat mass and high liver fat are associated with resistance to lifestyle intervention. *Obesity (Silver Spring)*. 2007;15(2):531–538.
20. Thamer C, et al. Interscapular fat is strongly associated with insulin resistance. *J Clin Endocrinol Metab*. 2010;95(10):4736–4742.
21. Matthaei S, et al. Pathophysiology and pharmacological treatment of insulin resistance. *Endocr Rev*. 2000;21(6):585–618.
22. Würslin C, et al. Topography mapping of whole body adipose tissue using A fully automated and standardized procedure. *J Magn Reson Imaging*. 2010;31(2):430–439.
23. Matsuda M, DeFronzo RA. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care*. 1999;22(9):1462–1470.

24. Babbar R, et al. Prediction of glucose tolerance without an oral glucose tolerance test. *Front Endocrinol (Lausanne)*. 2018;9:82.
25. Liu FT, et al. Isolation Forest. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q = isolation-forest. Accessed September 29, 2021.
26. Cambria E, White B. Jumping NLP curves: a review of natural language processing research. *IEEE Computat Intell Mag*. 2014;9:48–57
27. Shrikumar A, et al. Learning Important Features Through Propagating Activation Differences. http://arxiv.org/abs/1704.02685. Accessed September 29, 2021.
28. Clevert D-A, et al. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). https://arxiv.org/abs/1511.07289. Accessed September 29, 2021.
29. He K, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. https://arxiv.org/abs/1502.01852. Accessed September 29, 2021.
30. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980. Accessed September 29, 2021.
31. Abadi M, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://arxiv.org/abs/1603.04467. Accessed September 29, 2021.
32. Nickolls J, et al. Scalable parallel programming with CUDA: is CUDA the parallel programming model that application developers have been waiting for?. *ACM Queue*. 2008;6(2):40–53.