Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus

Manuela Hische\textsuperscript{1,2}, Olga Luis-Dominguez\textsuperscript{3}, Andreas F H Pfeiffer\textsuperscript{1,4}, Peter E Schwarz\textsuperscript{3}, Joachim Selbig\textsuperscript{2} and Joachim Spranger\textsuperscript{1}

\textsuperscript{1}Department of Endocrinology, Diabetes and Nutrition, Charité-Universitätsmedizin Berlin, Campus Benjamin Franklin, Hindenburgdamm 30, 12200 Berlin, Germany, \textsuperscript{2}Department of Bioinformatics, Max-Planck-Institute for Molecular Plant Physiology and Department of Bioinformatics, University of Potsdam, 14476 Potsdam, Germany, \textsuperscript{3}Division of Prevention and Care of Diabetes, Department of Medicine, University of Dresden, 01307 Dresden, Germany and \textsuperscript{4}Department of Clinical Nutrition, German Institute of Human Nutrition, 14558 Nuthetal, Germany

(c) 2010 European Society of Endocrinology

Abstract

Objective: The prevalence of unknown impaired fasting glucose (IFG), impaired glucose tolerance (IGT), or type 2 diabetes mellitus (T2DM) is high. Numerous studies demonstrated that IFG, IGT, or T2DM are associated with increased cardiovascular risk, therefore an improved identification strategy would be desirable. The objective of this study was to create a simple and reliable tool to identify individuals with impaired glucose metabolism (IGM).

Design and methods: A cohort of 1737 individuals (1055 controls, 682 with previously unknown IGM) was screened by 75 g oral glucose tolerance test (OGTT). Supervised machine learning was used to automatically generate decision trees to identify individuals with IGM. To evaluate the accuracy of identification, a tenfold cross-validation was performed. Resulting trees were subsequently re-evaluated in a second, independent cohort of 1998 individuals (1253 controls, 745 unknown IGM).

Results: A clinical decision tree included age and systolic blood pressure (sensitivity 89.3%, specificity 37.4%, and positive predictive value (PPV) 48.0%), while a tree based on clinical and laboratory data included fasting glucose and systolic blood pressure (sensitivity 89.7%, specificity 54.6%, and PPV 56.2%). The inclusion of additional parameters did not improve test quality. The external validation approach confirmed the presented decision trees.

Conclusion: We proposed a simple tool to identify individuals with existing IGM. From a practical perspective, fasting blood glucose and blood pressure measurements should be regularly measured in all individuals presenting in outpatient clinics. An OGTT appears to be useful only if the subjects are older than 48 years or show abnormalities in fasting glucose or blood pressure.

European Journal of Endocrinology 163 565–571

Introduction

Despite the enormous increase in the incidence of diabetes and the well-known impact on morbidity and mortality, about 50% of individuals having type 2 diabetes mellitus (T2DM) are not aware of the disease (1, 2). The mean duration of T2DM is about 5 years at the time of diagnosis with the consequence of prevalent long-term complications at diagnosis. Current diagnostic criteria are based on fasting glucose and the oral glucose tolerance test (OGTT) to characterize postprandial glucose metabolism. However, these procedures are invasive, rather cost intensive and time consuming for both patient and medical doctor. Thus, identifying individuals at risk for having or developing T2DM by a simple and reliable tool might improve clinical care of diabetes patients by offering the opportunity of earlier therapy. Specifically, in individuals with impaired fasting glucose (IFG) and/or impaired glucose tolerance (IGT), various intervention trials demonstrated a successful prevention of diabetes by lifestyle modification. Various risk scores for identification of prevalent or incident cases of impaired glucose metabolism (IGM), based either on clinical (3–6), laboratory (7), or even genetic (8) analysis, have been described. Most of them are based on regression analyses usually including numerous confounders. Although regression analysis is a good statistical option to identify all independently associated confounders, this does not necessarily imply that all those confounders are informative in terms of disease prediction or identification. Indeed, the informative value of most of those confounders identified by regression analysis is rather limited. Another option to establish tools for clinical practice is decision tree.
modeling, which is based on simple and robust heuristics and implements feature selection and threshold generation per se. A major advantage of using decision trees is the clear presentation of complex data, which allows easy interpretation and application in clinical practice. Indeed, decision trees have been shown to be a feasible method for diagnostic purposes (9–11).

We therefore applied this supervised machine learning approach to automatically generate decision trees on a cohort of 1737 individuals without previously known diabetes (1055 individuals with normal glucose tolerance (NGT), 493 individuals with IGM (IFG or IGT), and 189 with yet unknown T2DM). To evaluate the prediction accuracy tenfold cross-validation was used. The resulting decision trees were optimized to show high sensitivity, as required in the screening outpatient situation. Sensitivity and specificity of the optimized decision trees were confirmed in a second, independent cohort including 1998 individuals without known diabetes mellitus, and the quality of the resulting decision trees was compared to established questionnaires.

**Subjects and methods**

**Metabolic syndrome Berlin Potsdam Study cohort**

In total, 1737 individuals without known diabetes mellitus (1175 females and 562 males) of the cross-sectional Metabolic Syndrome Berlin Potsdam Study (Mesy-Bepo) were analyzed. This study included individuals aged >18 years from the cities of Berlin and Potsdam as well as the surrounding area. Individuals were recruited by paper advertisements and by collaborating outpatient clinics. Exclusion criteria for the sample used in this study were existing diseases such as diabetes, liver disease, renal failure, heart disease, or cancer. Thus, the study population used in this study is not representative for the general population of the region. Overt health problems were excluded by medical history, physical examination, and standard laboratory. Details of the phenotyping procedures were described previously (5). Mean age of the cohort was $52.1 \pm 13.3$ years, body mass index (BMI) $29.1 \pm 6.0$ kg/m$^2$, waist $95.6 \pm 15.2$ cm, systolic and diastolic blood pressure $125 \pm 17$ and $78 \pm 11$ mmHg respectively. The cohort was based on complete information regarding diabetes status and 13 anthropometric and lifestyle attributes, which were age, gender, waist circumference, hip circumference, BMI, waist to hip ratio (WHR), systolic and diastolic blood pressure, smoking status, parental diabetes, and a history of dyslipidemia, gout, and hypertension. A subset of individuals ($n = 1527$) was characterized by the additional complete availability of laboratory data, which were HbA1c, uric acid, creatinine, leukocytes, cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides (TG), fasting plasma glucose, fasting plasma insulin, and homeostasis model assessment-insulin resistance (HOMA-IR) (calculated according to (12)). Both cohorts did not substantially differ regarding baseline characteristics, specifically the case–control-ratio (39–61%) was comparable. Details of phenotypic characteristics of this sub-cohort are presented in Table 1.

Glucose metabolism was analyzed by one 75 g OGTT and groups were formed according to 2003 American Diabetes Association (ADA) criteria (13), which were relevant when this study was initiated. In all, 1055 participants had NGT, 137 had IFG, 268 had IGT, 88 had IFG and IGT, and 189 had unknown T2DM. Individuals with T2DM, IFG, and/or IGT were summarized as having ‘IGM.’

After sampling in EDTA tubes, blood was immediately chilled on ice, centrifuged, and aliquots were immediately frozen at $-80^\circ C$. Blood samples were analyzed by standard methods.

The experimental protocol of the study was approved by the institutional review board, and all subjects gave written informed consent.

**Dresden cohort**

In all, 1998 (1106 females and 892 males) subjects from German families with a family history of T2DM, obesity, or dyslipidemia participated in this study. All subjects were from the city of Dresden and adjacent area and

<table>
<thead>
<tr>
<th></th>
<th>Mesy-Bepo ($n = 1527$)</th>
<th>Dresden ($n = 1998$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>52.2 ± 13.4</td>
<td>51.6 ± 17.2</td>
</tr>
<tr>
<td>Sex (% female)</td>
<td>0.68</td>
<td>0.55</td>
</tr>
<tr>
<td>NGT (n)</td>
<td>926</td>
<td>1253</td>
</tr>
<tr>
<td>IFG and/or IGT (n)</td>
<td>433</td>
<td>551</td>
</tr>
<tr>
<td>T2DM (n)</td>
<td>168</td>
<td>194</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>95 ± 15</td>
<td>92 ± 14</td>
</tr>
<tr>
<td>Hip circumference (cm)</td>
<td>107 ± 12</td>
<td>104 ± 11</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>29.0 ± 5.8</td>
<td>26.7 ± 4.6</td>
</tr>
<tr>
<td>WHR</td>
<td>0.89 ± 0.10</td>
<td>0.88 ± 0.09</td>
</tr>
<tr>
<td>Systolic RR (mmHg)</td>
<td>125 ± 17</td>
<td>132 ± 20</td>
</tr>
<tr>
<td>Diastolic RR (mmHg)</td>
<td>78 ± 11</td>
<td>82 ± 12</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>5.5 ± 2.8</td>
<td>5.4 ± 0.6</td>
</tr>
<tr>
<td>Uric acid (µmol/l)</td>
<td>275 ± 79</td>
<td>316 ± 88</td>
</tr>
<tr>
<td>Creatinine (µmol/l)</td>
<td>81 ± 13</td>
<td>86 ± 18</td>
</tr>
<tr>
<td>Leukocytes (G/l)</td>
<td>5.8 ± 1.6</td>
<td>6.2 ± 1.8</td>
</tr>
<tr>
<td>Cholesterol (mmol/l)</td>
<td>5.4 ± 1.1</td>
<td>5.5 ± 1.2</td>
</tr>
<tr>
<td>HDL (mmol/l)</td>
<td>1.4 ± 0.4</td>
<td>1.5 ± 0.5</td>
</tr>
<tr>
<td>LDL (mmol/l)</td>
<td>3.38 ± 0.93</td>
<td>3.3 ± 1.04</td>
</tr>
<tr>
<td>TG (mmol/l)</td>
<td>1.4 ± 0.8</td>
<td>1.6 ± 1.8</td>
</tr>
<tr>
<td>Fasting glucose (mmol/l)</td>
<td>5.16 ± 0.78</td>
<td>5.58 ± 0.85</td>
</tr>
<tr>
<td>Fasting insulin (µmol/l)</td>
<td>53 ± 38</td>
<td>80 ± 54</td>
</tr>
<tr>
<td>HOMA-IR</td>
<td>2.1 ± 1.7</td>
<td>3.4 ± 2.7</td>
</tr>
</tbody>
</table>

RR, Riva-Rocci.
were at least 18 years old. Exclusion criteria were known diabetes, impaired renal function (creatinine > 1.5), presence of any disease with a strong impact on life expectancy (e.g. malignancy), and therapy with drugs known to influence glucose tolerance (e.g. steroids, β blockers, and thiazide diuretics).

The mean age of the cohort was 51.6 ± 17.2 years, BMI 26.71 ± 4.6 kg/m², waist 92.1 ± 14 cm, systolic and diastolic blood pressure 132 ± 20 and 82 ± 12 mmHg respectively. The cohort was based on complete information regarding diabetes status and 12 anthropometric and lifestyle attributes, which were age, gender, waist circumference, hip circumference, BMI, WHR, systolic and diastolic blood pressure, smoking status, parental diabetes, and a history of dyslipidemia and hypertension. All individuals underwent a 75 g OGTT after a 10 h overnight fast and plasma glucose and insulin concentrations were measured at baseline (−30, −15, and 0 min) and at 30, 60, 90, and 120 min after glucose ingestion. Plasma was immediately separated from whole blood and kept on ice for further analysis. Lipid profile and HbAlc were measured in the blood sample obtained at baseline. The detailed procedure of subject recruitment and the methods used have been described previously (14). The diagnosis of diabetes was based on American Diabetes Association criteria: 2 h plasma glucose ≥ 200 mg/dL or fasting plasma glucose ≥ 126 mg/dL. Subjects on insulin or oral antihyperglycemic medications were also considered to have diabetes.

**Laboratory measurements** Plasma glucose was measured using the hexokinase method (interassay coefficient of variation (CV) = 1.5%). Plasma insulin concentration was measured with RIA (Biosource EUROPE, Nivelles, Belgium; interassay CV = 7.5% with no cross reactivity with proinsulin). HbAlc was measured with HPLC.

**Statistical analysis**

A standard decision tree algorithm (15) was applied to Mesy-Bepo, either based on clinical data alone or including laboratory markers from fasting blood samples. All automatically created decision tree models were validated via tenfold cross-validation (16). Several parameters, such as the penalty term, the minimal number of subjects per leaf, and the confidence factor, were varied to maximize the resulting sensitivity and specificity (15). From all models with a mean sensitivity above 90% after cross-validation, we further analyzed the four models with maximal mean specificity. From this group we formed a robust hybrid and manually pruned this to allow up to two node levels. For all data sets, the positive predictive value (PPV) and the negative predictive value (NPV) were additionally calculated to describe the models’ performance. Both resulting models (with and without laboratory parameters) were externally validated within an independent cohort from the region of Dresden.

Established risk scores and tools were calculated within the Mesy-Bepo cohort according to the information provided in the respective publications (7, 17, 18).

**Results**

Depending on the selected parameters of the decision tree modeling, the threshold of discriminating nodes and consequently the sensitivity and specificity of the respective decision trees varied considerably (Fig. 1). As expected, the use of laboratory in addition to clinical parameters improved the quality of identification.
In the model based on clinical data, age was the strongest discriminating factor with a threshold of 48.3 years. Individuals younger than 48.3 years were further sub-classified by a second node based on systolic blood pressure with a threshold at 127 mmHg. The sensitivity of the complete model was 89.3% (notably in contrast to a mean sensitivity slightly above 90% after cross-validation) and the specificity was 37.4%. The PPV was 48.0% and the NPV was 84.4%. In total, 9 of 189 (5%) individuals with existing T2DM were misclassified as NGTs (Fig. 2A). Aiming for an external validation, this tree was applied to the Dresden cohort. Remarkably, this clinical-data-based decision tree had a quality, in that independent cohort, which was comparable to the quality in Mesy-Bepo (Dresden cohort: sensitivity 90.3%, specificity 32.1%, PPV 44.2%, and NPV 84.8%) (Fig. 3A). Finally, 12 of 194 (6%) individuals with yet unknown T2DM were misclassified as NGTs.

Subsequently, an additional model including clinical and laboratory data was calculated. Within that model, fasting glucose was the most important discriminating factor and individuals were classified according to a threshold of 5.07 mmol/l at a first node. Subjects with a fasting glucose below 5.07 mmol/l were sub-classified by systolic blood pressure with a threshold at 130 mmHg (Fig. 2B). The sensitivity of this model was 89.7% (also in comparison with a sensitivity above 90% after cross-validation) and the specificity was 54.6%. Two of the 168 (1%) cases of T2DM were misclassified as NGM. The PPV was at 56.2%, and the NPV was at 89.1%.

Applying the decision tree including the laboratory data to the Dresden cohort revealed comparable quality to identify individuals with IGM. The sensitivity was 95.0% and the specificity was 27.8%. One case of T2DM was misclassified as NGM (0.5%). The PPV was at 43.9%, and the NPV was at 90.4% (Fig. 3B).

Those data were finally compared to established tools. Those tools were established to identify individuals with existing T2DM, but not IGM. Notably, the quality of those questionnaires and tools was, in summary, comparable with that of the proposed models presented in this study, although these decision trees are based on two nodes only (Table 2).

**Discussion**

Lifestyle or pharmacological interventions can effectively decrease diabetes risk in high-risk individuals, i.e. subjects with IFG or IGT (19–21). Thus, the identification of those risk groups is highly desirable to imply individualized preventive approaches. Ideally, a practically useful tool would be able to identify both, individuals at high risk for diabetes and those with existing, but yet unknown, diabetes. This can effectively help to direct preventive or therapeutic interventions with efficient use of resources. The decision trees proposed in this study offer a simple and robust tool to identify individuals with IFG, IGT, or existing, but yet unknown, T2DM within a cohort of apparently healthy subjects. Notably, the trees were successfully validated in a second, independent cohort.
Various tools were developed to identify individuals with existing T2DM (17, 18) or those at high risk to develop diabetes (4, 5). Most of the existing questionnaires are based on regression analyses, which consider all independently associated confounders, even if the respective effect of that confounder is rather limited. In result, the questionnaires often include numerous attributes and may be quite complex.

Decision tree modeling is a supervised machine learning approach and as such it is a feasible method to calculate robust, reliable, and simple-to-use tools. In practice, such a method would be desirable to identify individuals, who would benefit from a lifestyle intervention, but to avoid oral glucose tolerance testing in all subjects. The decision trees proposed in this study include well-established risk factors of diabetes, such as age, blood pressure, and fasting glucose. The proposed decision trees would avoid OGTTs in about 25% of individuals with a reasonable rate of false-negative results. This is comparable to the existing, regression-based questionnaires (22), although the decision proposed in this study is much simpler to use. However, it needs to be considered that those questionnaires were not developed to identify subjects at risk and with existing T2DM. Thus, a regression-analysis-based questionnaire may have a better screening quality, if specifically individuals with IGM are targeted. A recent publication also proposed a decision tree to identify subjects with existing T2DM (7). We also applied this tree to our cohort. Comparable to the results of the initial publication, the sensitivity (89%) and specificity (95%) to identify individuals with existing, but unknown, diabetes was high. In contrast, the sensitivity to identify individuals with IFG, IGT, or T2DM was only slightly above 30%, thus not helpful to identify individuals with a prediabetic situation, although it is well known that those patients carry considerable cardiovascular risk. Since this tree was not developed to identify individuals with a prediabetic situation, the low sensitivity with respect to this end point is therefore not surprising.

As expected, a decision tree including laboratory parameters improved the quality characteristics compared with a tree based exclusively on clinical data. Our results suggest that fasting glucose is highly informative to identify individuals at risk for future diabetes and those with existing diabetes. The suggested thresholds at the tree nodes appear to be valid, since application of those thresholds to an entirely independent cohort from the region of Dresden revealed comparable quality measures.

Some limitations of our study should be mentioned. The proposed trees are of limited value to identify individuals with IGT. Actually up to 25% of subjects with IGT are misclassified, although those individuals need to be considered as being at risk for future diabetes and should be offered a preventive intervention. On the other hand, all the proposed tools are far from perfectly identifying individuals at risk or with existing diabetes. According to the quality measures analyzed in Mesy-Bepo, the proposed decision trees are at least comparable to other established tools, despite the much higher complexity of those questionnaires. A drawback of our model is the limited specificity. The model is intended to support the decision in clinical practice, whether or not an OGTT in apparently healthy subjects is required. Clearly, we aimed to identify all individuals having diabetes, rather than improving specificity. Under these assumptions, our trees will help to avoid OGTTs in up to 20% of all individuals. Nevertheless, the low specificity apparently results in the recommendation of OGTTs on a still considerable number of finally healthy subjects. Apparently, clinical data (and fasting plasma glucose (FPG)) are not sufficiently informative to solve this problem. Thus, although our trees help to avoid unnecessary OGTT, they do not entirely exclude them. This is comparable to the quality data of all other available tools, despite the lower complexity of our decision trees. The major advantage of our models is therefore the scientific basis of thresholds and the low complexity of the models, while the trees do not solve the problem of low specificity in the prediction of existing diabetes.

The trees were additionally validated in an independent cohort from the region of eastern Germany. Apparently, the specificity was even worse after validation in that second, independent cohort. The decrease in specificity suggests an overfiling of the original data by our model, which was not completely

---

Table 2: External validation of the novel decision trees in an independent cohort from the region of Dresden, Germany and comparison of performance characteristics to three established risk scores applied to the Mesy-Bepo cohort. Performance characteristics were calculated to identify individuals with impaired fasting glucose (IFG), impaired glucose tolerance (IGT), or type 2 diabetes mellitus (T2DM). Please note that the compared risk scores in this study (7, 17, 18) were initially established to identify individuals with prevalent T2DM, but not IFG or IGT.

<table>
<thead>
<tr>
<th>Data characteristics</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Positive predictive value (%)</th>
<th>Negative predictive value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>Clinical data</td>
<td>89.3</td>
<td>37.4</td>
<td>48.0</td>
</tr>
<tr>
<td>Mesy-Bepo</td>
<td>Lab data</td>
<td>89.7</td>
<td>54.6</td>
<td>56.2</td>
</tr>
<tr>
<td>Decision tree</td>
<td>Clinical data</td>
<td>90.3</td>
<td>32.1</td>
<td>44.2</td>
</tr>
<tr>
<td>Dresden</td>
<td>Lab data</td>
<td>95.0</td>
<td>27.8</td>
<td>43.9</td>
</tr>
<tr>
<td>Glümer et al.</td>
<td>Clinical data</td>
<td>91.7</td>
<td>20.1</td>
<td>42.7</td>
</tr>
<tr>
<td>Al-Lawati et al.</td>
<td>Clinical data</td>
<td>90.2</td>
<td>28.6</td>
<td>45.1</td>
</tr>
<tr>
<td>Manley et al.</td>
<td>Lab data</td>
<td>30.5</td>
<td>96.1</td>
<td>83.6</td>
</tr>
</tbody>
</table>

---

www.eje-online.org
avoids despite tenfold internal cross-validation within the initial set of data. Overfitting is a well-known effect in statistic modeling and leads to overoptimistic test characteristics. We feel that the very thorough cross-validation and the additional external validation are a strength of our study; because the results are likely to give a realistic picture of prediction quality with the given data. Numerous previous publications did not perform such a rigorous validation, thereby questioning the applicability of those results for clinical practice. Thus, despite the resulting low specificity (which demonstrates the limited value of currently used clinical data for prediction), the rigorous validation process of our study is a considerable strength of our study.

Both analyzed cohorts included predominantly Caucasian. It is unclear, whether these results, specifically the suggested thresholds, can be transferred to other populations.

The classification of individuals was in both cohorts based on one OGTT. It is well known that the results of OGTT can vary considerably over time. We therefore cannot exclude that some individuals may have been misclassified according to one OGTT, although this is unlikely to challenge the overall results.

Our work is at risk for incorporation bias, since fasting plasma glucose was used both for defining the glucose metabolism status and generating the decision tree. Therefore, the test characteristics might be overestimated. However, the overestimation is rather small, since an extreme bias would have resulted in a trivial tree containing only one node, FPG. The risk for incorporation bias is a problem of diabetes diagnosis itself, which is based on biochemical values. It is basically impossible to avoid the risk of incorporation within the given setting and without accepting a substantial loss of information. This is apparent by comparing our trees with and without laboratory data (which has no risk of incorporation bias). The tree with FPG is substantially more informative. Therefore, not considering fasting glucose as a potential marker would be a loss of valuable information, which is why we felt that the use of that parameter was justified despite the potential bias.

In summary, the data presented in this study suggest that individuals with IFG, IGT, or existing, but yet unknown, diabetes can be identified with reasonable accuracy by simple measures such as age, systolic blood pressure, and fasting glucose. All results were confirmed in a second, independent cohort. Although the relatively low specificity of the trees limits a broad use as screening tools, the data support that fasting blood glucose and blood pressure measurements are most informative to estimate existing diabetes and should be regularly performed in all individuals presenting to a clinical practice. If the subjects are older than 48 years or show abnormalities in fasting glucose or blood pressure, an OGTT should be performed.

Declaration of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding
This work was supported by a research group (Molecular Nutrition) of the Bundesministerium für Bildung und Forschung (BMBF), a Heisenberg-Professorship (SP716/2-1), and a clinical research group (KFO218/1) of the Deutsche Forschungsgemeinschaft.

Author contribution statement
M Hische, P E Schwarz, J Selbig, and J Spranger contributed to design of the study, statistical analyses, and writing of the manuscript. O Luis-Dominguez performed statistical analysis and contributed to the establishment of the Dresden cohort. A F H Pfeiffer contributed to the establishment of the Mesy-Bepo cohort.

Acknowledgements
The authors thank Katrin Sprengel for excellent technical assistance.

References
Identification of impaired glucose metabolism


Received 29 July 2010
Accepted 6 August 2010