How to distinguish between type 1 and type 2 diabetes at diagnosis in ≥ 16-year-old patients

Peik Romantschuk

Bachelor of Medicine

University of Helsinki

Helsinki 29.11.2021

Thesis

peik.pietila@helsinki.fi

Supervisor: Tiinamaija Tuomi

UNIVERSITY OF HELSINKI

Faculty of Medicine

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department
Medicinska fakulteten	

Tekijä – Författare – Author

Peik Romantschuk

Työn nimi – Arbetets titel – Title

Hur man skiljer på typ 1 och typ 2 diabetes vid diagnos hos patienter över 16 år

Oppiaine – Läroämne – Subject

Medicin

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä -Sidoantal - Number of pages
Studie	11/2021	17+1

Tiivistelmä – Referat – Abstract

Mål: Diabetes är en sjukdom som uppstår då bukspottkörteln inte kan producera tillräckliga mängder av insulin för att upprätthålla en fysiologisk nivå på blodsockret. Olika patofyisologiska mekanismer ligger bakom detta tillstånd och är beroende på typen av diabetes. För optimal vård av diabetes, är det viktigt att vid ett tidigt skede kunna utgöra vilken typ av diabetes en patient har insjuknat i.

Metod: I denna studie undersökte vi vilka kliniska variabler är mest betydande för att skilja mellan typ 1 och typ 2 diabetes, vid tidpunkten av diagnos. Vi använder dessa variabler för att träna och validerar en CART maskininlärnigsmodell för att kunna skilja på typ 1 och typ 2 diabetes, speciellt i fall där det är oklart vilken subtyp patenten hör till.

Resultat: Blodsockernivå, C-peptidnivå samt deras förhållande och BMI samt ålder vid insjunkning i diabetes visade sig vara de mest signifikanta kliniska variablerna. Vår modell klarade av att skilja på typ 1 och typ 2 diabetes med 91,8 % noggrannhet av testdatat som bestod av 1175 patienter. CART modellen är således en användbar modell för att differentiera diabetes typer vid tidpunkten av diagnos hos patienter över 16 år. (190 ord)

Avainsanat – Nyckelord – Keywords

Diabetes, Machine learning, Classification tree, CART

Säilytyspaikka – Förvaringställe – Where deposited

Terkko, Helda

Muita tietoja – Övriga uppgifter – Additional information

Table of Contents

1 Introduction

Diabetes occurs when the beta-cells of the pancreas cannot secrete efficient amounts of insulin to keep the blood glucose on a physiological level. This happens due to various causes leading to the destruction of beta cells, impaired response to increases in blood glucose, decrease in insulin sensitivity of target tissues or a combination of all these outcomes[1] [2]. The result is a supra-physiological blood glucose level, which apart from the acute complications, over time can cause multiple chronic complications in patients including retinopathy, nephropathy, neuropathy and cardiovascular disease. The underlying cause of diabetes affects the treatment, which is why the correct classification is of the essence.

Current general classification of diabetes by the American Diabetes Association**:**[3]

1. *Type 1 diabetes (due to β-cell destruction, usually leading to absolute insulin deficiency)*
2. *Type 2 diabetes (due to a progressive insulin secretory defect on the background of insulin resistance)*
3. *Gestational diabetes mellitus (GDM) (diabetes diagnosed in the second or third trimester of pregnancy that is not clearly overt diabetes)*
4. *Specific types of diabetes due to other causes, e.g., monogenic diabetes syndromes (such as neonatal diabetes and maturity-onset diabetes of the young [MODY]), diseases of the exocrine pancreas (such as cystic fibrosis), and drug- or chemical-induced diabetes (such as in the treatment of HIV/AIDS or after organ transplantation)*

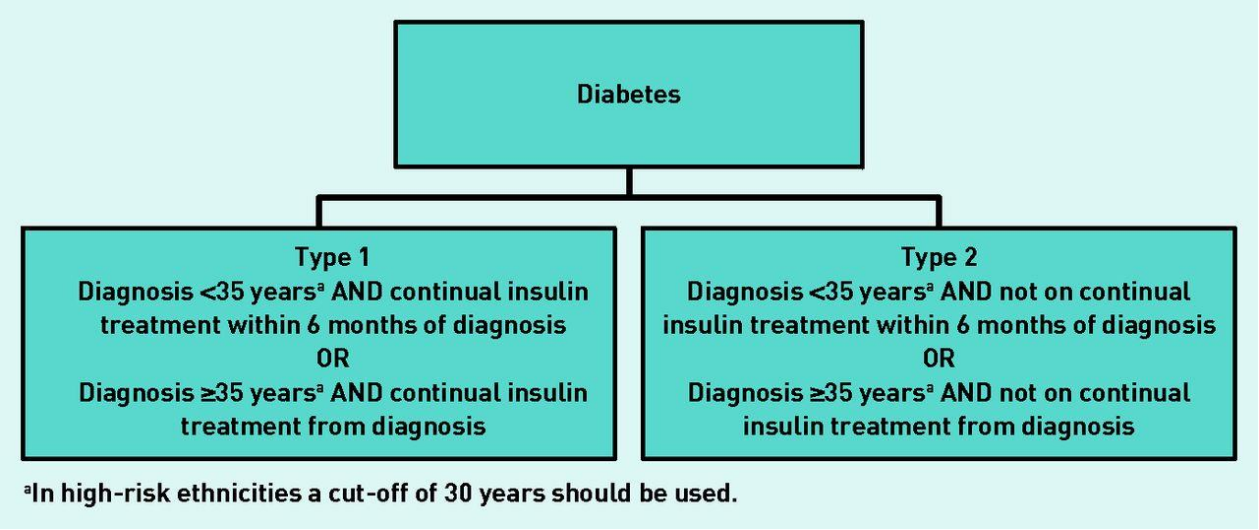
In Finland, the total number of patients entitled to reimbursement for medication purchases due to diabetes was 6.3% (n = 346 887) of the population in 2018[4]. The number of people with diabetes is likely higher since patients with only one oral diabetes medication may not get reported to the national medical register and many patients with T2DM are undiagnosed. It is estimated that 75-80% of the patients have T2DM, and the rest have other forms of diabetes with T1DM as the majority. T1DM incidence in Finland is among the highest in the world with 65 incidences per 100 000 person-years in 2005[5].

Correct classification of a patient's diabetes is essential for optimal management of the disease and to minimize the risks for chronic complications. Current clinical practice does not

accurately classify patients. 7-15% of patients with diabetes are misclassified in all age groups and the proportion is even higher when the age at diagnosis is >30 years[6].

The current clinical guidelines for diabetes classification do not provide criteria that would unambiguously differentiate between the different types of diabetes. Pragmatic guidelines on diabetes classification have been developed but are based on consensus expert clinical opinion and are only partly evidence-based (e.g. Figure 1). In practice, differential diagnostics is based on the general traits of the different types of diabetes, rather than a clear clinical criterion. Ketoacidosis, low C-peptide, autoantibodies (to GAD, IA2, etc.), young age and involuntary weight loss are characteristic for type 1 diabetes[7]. Another factor taken into account is age, with over 70% of diabetes occurring before the age of 20 being type 1 diabetes[8]. On the other hand, a recent population-based genetic stratification analysis has suggested that 42% of type 1 diabetes occurs after the age of 30[9]. Higher BMI is strongly correlated with insulin resistance and is therefore a clinical parameter associated with type 2 diabetes[10]. However, with the population becoming increasingly overweight this parameter is becoming less predicting. Although clinical traits can point towards a certain type of diabetes, none of these traits alone can ensure type 1 and exclude type 2 or vice versa.

We aimed to evaluate which variables could accurately discriminate T1DM from T2DM by combining follow-up data with data for clinical variables measured at the time of diagnosis.



UK Practical Classification Guidelines for Diabetes (extract showing an algorithm of classification guidelines for type 1 and type 2 diabetes). 2010 (Figure 1)

2 Objective

The first objective of this study is to compare clinical features of T1DM and T2DM. The aim is to find differences in clinical and serological parameters, which could help differentiate between adult-onset diabetes types at the stage of diagnosis (DG), particularly in cases where the clinical representation of diabetes is not overwhelmingly pointing towards a certain type. The second objective is to create and validate a diabetes type classification model using machine learning.

3 Material and methods

3.1 Data

Data for this study was acquired from the Diabetes Registry Vaasa (DIREVA)[11], a register in the Vasa hospital region upheld by the University of Helsinki and Lund University (Sweden). The DIREVA register consists of patients with diabetes within western Finland (~200 000 inhabitants) and included 6674 individuals with diabetes recruited between 2009 – 2018; recruitment is still ongoing. The project was approved by the ethics committee of Vasa Central Hospital (6/2007). It aims to identify factors affecting response to therapy and complications in patients with diabetes.

Additionally, we used data on patients diagnosed with T1DM within the Helsinki hospital region in southern Finland (~1.68 million inhabitants). Patients, treated at the Helsinki University Hospital (HUCH) between 2010-2015, who at some point were diagnosed with T1DM were called for a follow-up visit 3-5 years after the DG. Clinical variables were measured, and a blood sample was taken. Additional data of clinical variables at the stage of DG was gathered manually from electronical medical records Oberon and WebLab. In the end, the HUCH T1DM dataset consisted of 902 individuals (includes patients that part took in the study without coming to a follow-up visit).

From the initial DIREVA and HUCH T1DM datasets, we selected individuals ≥16 years of age at DG with clinical variables available “at the time of DG”. To define “at the stage of DG” we accepted data measured within one year from DG for T2DM, whereas for T1DM the data had to be measured truly at the stage of DG (< 1 week after DG). Due to the nature of

T2DM, significant changes in clinical variables are not expected in one year and should resemble values at the stage of DG. Finally, cases with more than 3 missing variables were excluded.

The number of people who met the study inclusion criteria in DIREVA and HUCH T1DM database was 2224 (Table 1).

Variables that were available for input in the acquired data:

- Age
- BMI
- Fasting C-peptide
- Fasting glucose
- Blood pressure
- HDL
- LDL
- Triglycerides
- Glutamic acid decarboxylase antibodies
- Follow-up C-peptide

3.2 Defining the dependent variable

We defined the DG, T1DM or T2DM, based on GAD antibody positivity and the level of insulin secretion two years after the DG of diabetes according to the following criteria:

True T1DM was defined according to C-peptide value ≤ 0.2 nmol/l two years from DG and positivity for pancreatic autoantibodies (GADAb > 10 IU/ml) at DG. True T2DM was defined as C-peptide > 0.2 nmol/l after two years and no autoantibodies at DG. We did this to circumvent possible misclassifications of diabetes type in the data and to have a clear definition of T1DM and T2DM.

3.3 Statistical analysis and methods

Initial analysis and data visualization was done on SPSS 25, to select which variables to use as input for the machine learning algorithm. Next, a decision tree model was built and tested using a CART[12] algorithm. Finally, a variable importance computation was performed using a random forest algorithm. All machine learning modeling was done using R software, version 3.6.1.

A decision tree is a supervised machine learning modeling technique for regression and classification problems. It works by predicting an outcome variable based on input variables in a flow chart manner. The algorithm computes optimal splits in the flowchart, based on

training data, so that it maximizes homogeneity in the subdivision of data after each split. Different algorithms use different metrics for computing the optimal splitting point e.g. information gain, variance reduction or Gini index. We chose a CART algorithm because we wanted to create a model that favors generalizability more than accuracy and a model that can be visualized instead of a "black box". Hence, making it easier to interpret and perhaps more useful in clinical practice.

A random forest is similar to a classification tree algorithm but goes one step further. It works by training a collection of decision trees and letting each tree vote on the outcome. In a classification setting, a single output is then given by the majority of votes. In the process of creating the random forest, the algorithm randomly selects data points in a process called "bagging". The accuracy of the random forest can then be estimated by predicting classes of the samples that were not selected in the training process (so-called "out-of-bag samples"). To measure the importance of an input variable, said variable is permuted while other variables remain unchanged. By assessing the random forests decrease in the correct classification of "out of the bag samples", when a specific variable is randomized a comparable measurement of "variable importance" can be calculated[14].

4 Results

Higher age at DG and features of metabolic syndrome (high BMI, high LDL, low HDL, high blood pressure) are more common in patients with T2DM. Therefore, it is expected that there would be a difference among these features between T1DM and T2DM at the stage of DG, and thus be predictive of the diabetes type.

As expected, we found differences in BMI and age at DG between T1DM and T2DM (Figure 2 and Table 1). We also found clustering of DM types when plotting fasting glucose and C-peptide in a scatter plot (Figure 3), but surprisingly no major differences in lipids or blood pressure values (Figures 4, 5, 6 and Table 1). This is likely due to vigorous monitoring and treatment of dyslipidemia and hypertension that is common in primary care. Since medical records of statins and antihypertensive drugs were not collected, we were not able to correct for the use of these medications. It is plausible that a bigger difference had been found if the use of medication would have been taken into account.

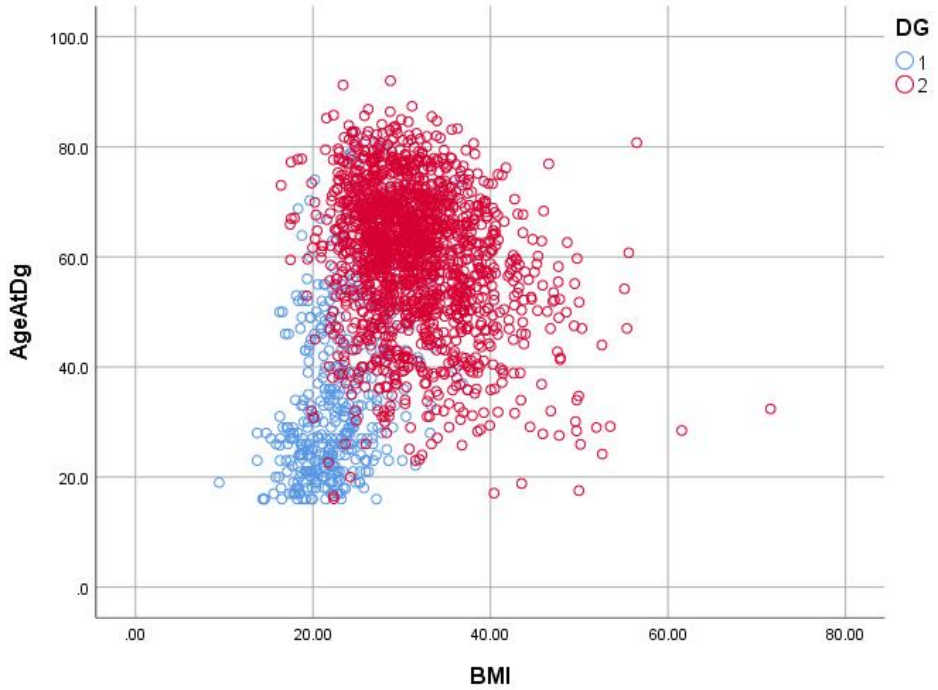


Figure 2. Relationship age and BMI in T1DM and T2DM at the stage of diagnosis

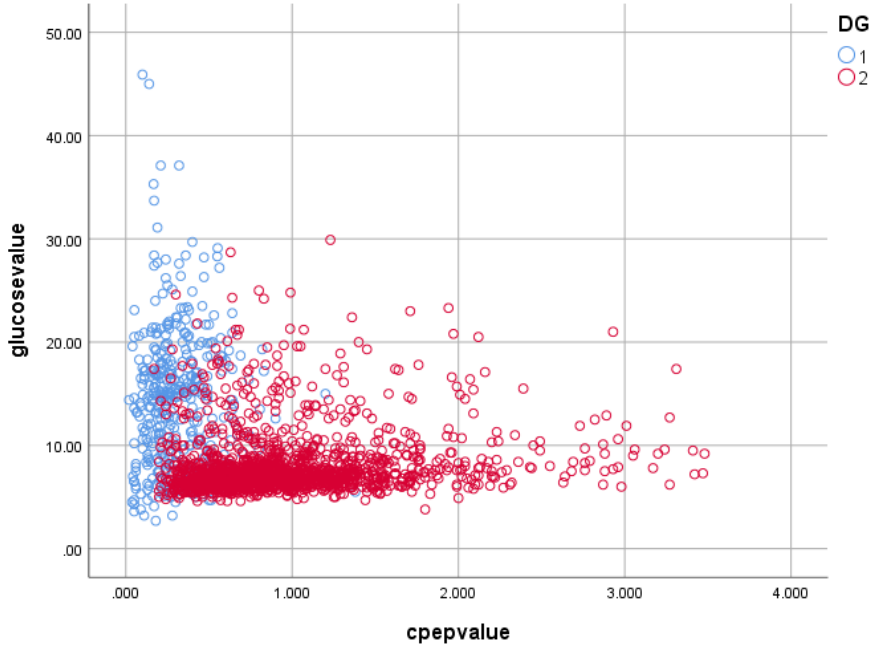


Figure 3. Relationship of plasma glucose and C-peptide in T1DM and T2DM at the stage of diagnosis

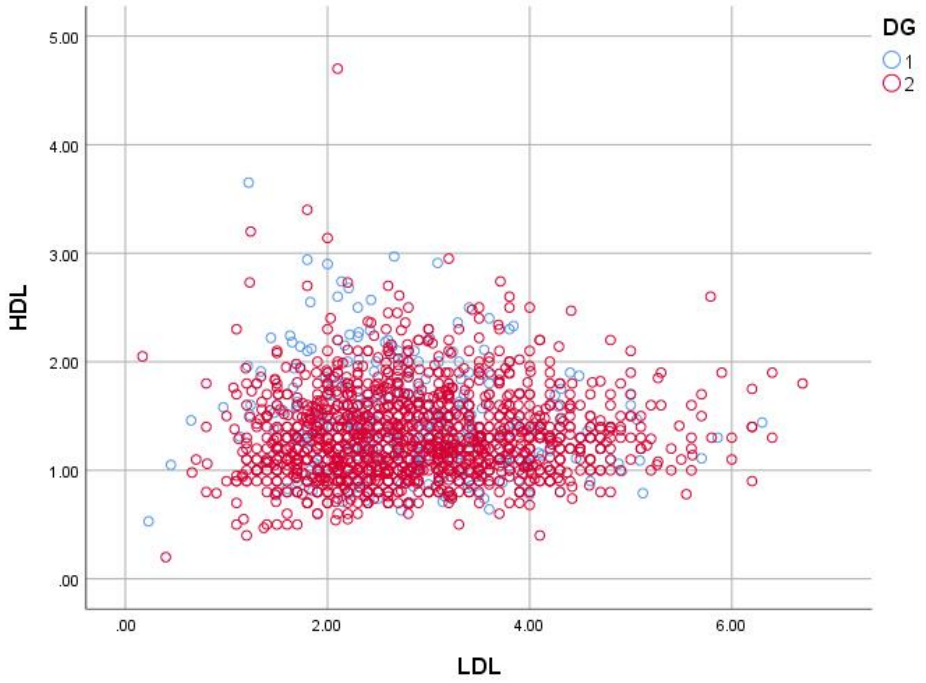


Figure 4. Relationship of HDL and LDL in T1DM and T2DM at the stage of diagnosis

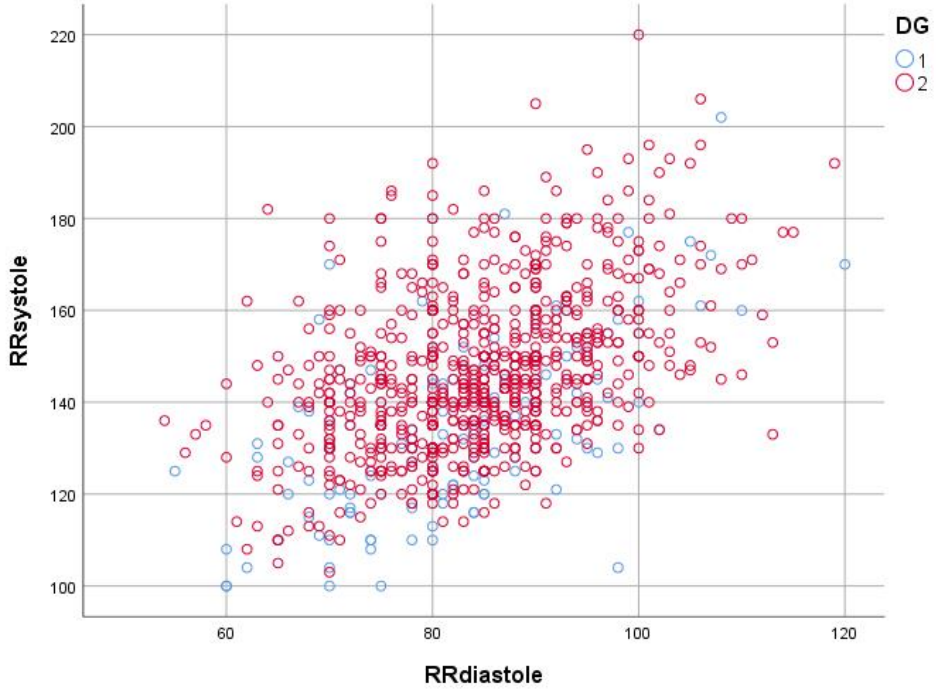


Figure 5. Relationship of systolic and diastolic blood pressure in T1DM and T2DM at the stage of diagnosis

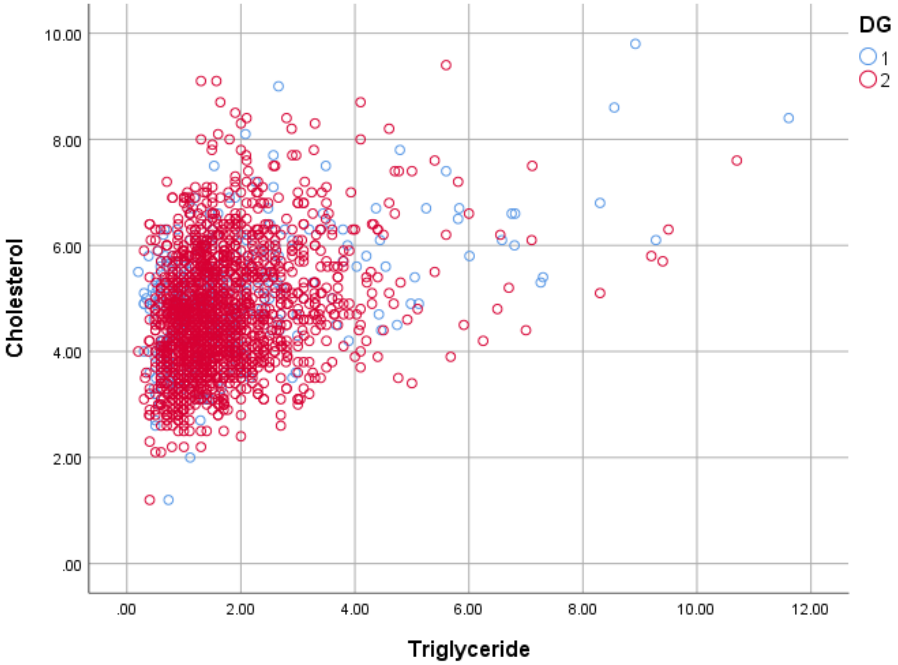


Figure 6. Relationship of total cholesterol and triglycerides in T1DM and T2DM at the stage of diagnosis

Group Statistics

	DG	N	Mean	Std. Deviation	Std. Error Mean
cpepvalue	1	465	.3138	.18384	.00853
	2	1717	.9549	.51949	.01254
glucosevalue	1	433	14.0596	6.81265	.32740
	2	1705	8.0339	3.21021	.07774
cpg_ratio	1	432	2.86722	2.499137	.120240
	2	1705	12.69197	6.896455	.167018
AgeAtDg	1	504	34.272	15.0333	.6696
	2	1718	59.850	12.8167	.3092
BMI	1	469	22.4640	4.09144	.18892
	2	1705	31.3573	5.89175	.14269
RRsyst	1	173	133.71	18.734	1.424
	2	837	145.85	17.111	.591
RRdiast	1	173	81.33	11.785	.896
	2	838	84.51	9.940	.343
LDL	1	397	2.8594	1.03474	.05193
	2	1648	2.8684	.99067	.02440
HDL	1	432	1.4366	.53507	.02574
	2	1594	1.3012	.40123	.01005
Triglyceride	1	435	2.4588	6.00379	.28786
	2	1624	1.7399	1.12602	.02794
Cholesterol	1	432	5.1253	1.71099	.08232
	2	1636	4.7305	1.12727	.02787

Table 1. Descriptive statistics by diabetes type.

4.1 Training the tree

We decided to use C-peptide to glucose ratio, BMI and age as input variables for the classification tree. As training data, we used all samples that had no missing values regarding GADAb, C-peptide, BMI and age at the stage of DG. Follow-up C-peptide also had to be available so that we could give our own true diagnosis to each sample. "True DG", defined earlier, was used as the dependent variable when training the model. Training data matched the total data in variable means and variance reasonably well when taking the distribution of diabetes types into account. To avoid overfitting, we used K-folds cross-validation for pruning the tree. The complexity parameter was set at 0.03.

TrueDG

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	153	49.0	49.0	49.0
	2	159	51.0	51.0	100.0
	Total	312	100.0	100.0	

Table 2. Structure of training data

A total of 312 patients met the training inclusion criteria. The training population consisted of 153 (49%) true T1DM and 159 (51%) true T2DM. The mean age at DG, BMI and C-peptide to glucose ratio was 46.8 ± 18.1 years, 26.2 ± 6.9 kg/m² and 6.6 ± 6.3 nmol/mmol respectively. The most predictive variable was C-peptide to glucose ratio with a cutoff value of 3.83 nmol/mmol (CPG ratio = 100 * C-peptide/Glucose). A CPG ratio less than 3.83 nmol/mmol was classified as T1DM. If the CPG ratio was ≥ 3.83 nmol/mmol the subset was split using age at DG with a cutoff value of 36.5 years. A CPG ratio ≥ 3.83 nmol/mmol and age at DG ≥ 36.5 years was classified as T2DM. If the CPG ratio was ≥ 3.86 nmol/mmol but the age at DG was under 36.5 years, a final split of the subset was made using BMI with a cutoff value of 32.6 kg/m², less than this classified as T1DM, ≥ 32.6 kg/m² classified as T2DM. (Figure 7)

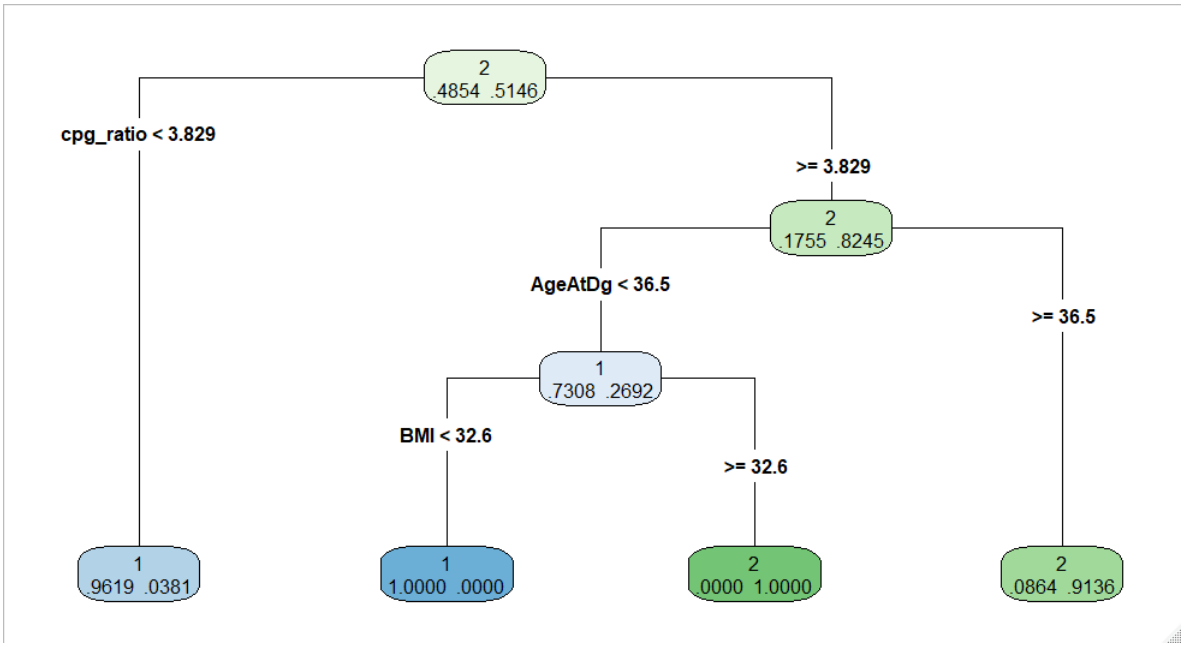


Figure 7. Visual representation of CART classification tree. The majority category and the relative ratios of each category are represented in the nodes

4.2 Validating the tree

A total of 1175 patients met the testing inclusion criteria (no missing values in C-peptide, glucose, age at DG and BMI, but missing values for GADAb and follow-up C-peptide allowed). Testing population consisted of 289 (24.6%) T1DM and 886 (75.4%) T2DM. The mean age at DG, BMI and C-peptide to glucose ratio was 48.1 ± 13.7 years, 30.0 ± 7.3 kg/m² and 10.1 ± 7.2 nmol/mmol, respectively.

DG

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	289	24.6	24.6	24.6
	2	886	75.4	75.4	100.0
	Total	1175	100.0	100.0	

Table 3. Structure of testing data

```
Confusion Matrix and Statistics

predictions_MODEL   1   2
                1 260  67
                2  29 819

               Accuracy : 0.9183
                 95% CI : (0.9011, 0.9333)
    No Information Rate : 0.754
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7891

 Mcnemar's Test P-Value : 0.0001592

            Sensitivity : 0.8997
            Specificity : 0.9244
         Pos Pred Value : 0.7951
         Neg Pred Value : 0.9658
             Prevalence : 0.2460
         Detection Rate : 0.2213
   Detection Prevalence : 0.2783
      Balanced Accuracy : 0.9120

       'Positive' Class : 1

 Area under curve = 0.9053442
```

Table 4. Descriptive statistics of model predictions

The total accuracy of the model was 91.83% with a sensitivity of 0.8997 and specificity of 0.9244 in regard to T1DM with an area under the curve of 0.905 (Table 4).

4.3 Variable importance

Variable importance in the random forest when taking all available variables into account in the testing data, showed that imputing the CPG ratio resulted in the biggest reduction in model accuracy while randomizing gender did not affect accuracy.

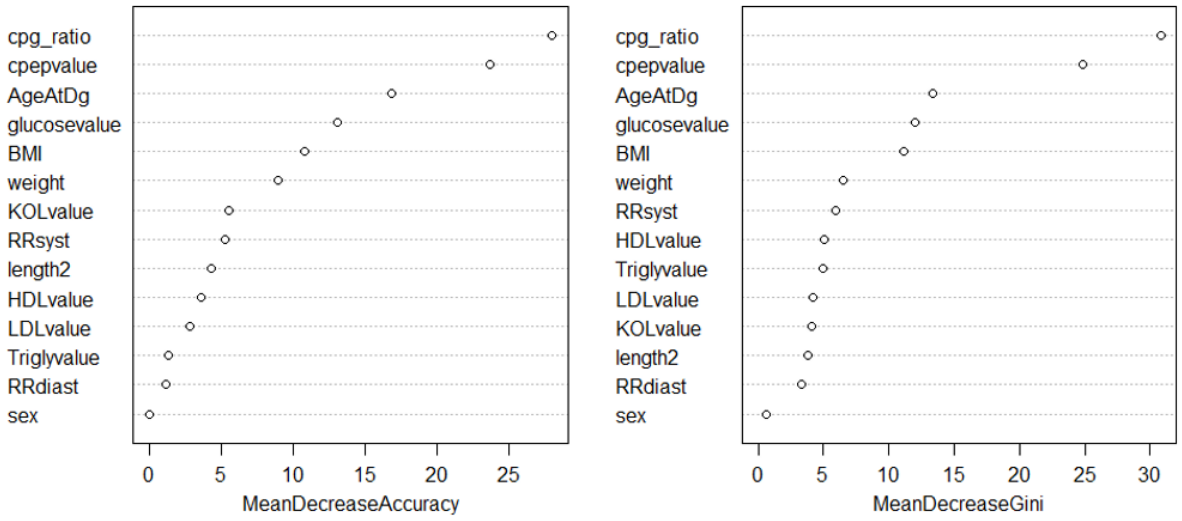


Figure 8. Variable importance of testing data

5 Discussion

The most predictive variable for distinguishing T1DM from T2DM turned out to be the C-peptide to glucose ratio, and the least predicting was BMI. This is understandable, as the population is becoming more overweight and since high BMI is not protective against T1DM, but still increases the risk for T2DM. Data from the Finnish Institute for Health and Welfare shows that the number of 15-64-year old's with BMI > 25 kg/m² went up by 7% between 2000-2014. The model reflects the findings of other studies regarding clinical parameters that suggest that age under 30-40 years and BMI around 27-28 kg/m² is the optimal cutoff for differentiating between T1DM and T2DM[15] (review article). It should be noted that these studies looked at the criteria separately unlike our model that sets the criteria subsequently.

variable.importance	double [3]	85.9 18.8 10.1
cpg_ratio	double [1]	85.92375
AgeAtDg	double [1]	18.78693
BMI	double [1]	10.08

*Table 3. An overall measure of variable importance given by the rpart() function in the caret package is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness * adjusted agreement for all splits in which it was a surrogate. The values are scaled to sum to 100.*

5.1 The role of C-peptide and CPG ratio

In practice, a C-peptide under 0.6-0.7 nmol/l is used as a cutoff pointing towards T1DM if the fasting glucose is above reference. However, a healthy individual has a fasted C-peptide concentration of 0.3-0.6 nmol/l under physiological glucose levels[16]. This creates an issue of incomparability. The only time C-peptide alone can be useful for distinguishing between diabetes types is when it is under 0.2 nmol/l. Incomparability is not as big of an issue when comparing the CPG ratio. A healthy individual will rarely have a ratio under 6 nmol/mmol (assuming physiological blood glucose to be between 4-6 nmol/l in ≥ 16-year-olds and C-peptide between 0.3-0.6 nmol/l), thus differentiating between healthy individuals and an individual with T1DM. As presented in this study, the CPG ratio can also differentiate between an individual with T1DM and T2DM with reasonably good accuracy, especially when combined with age at DG and BMI. Other studies of C-peptide to glucose ratio are limited. One study by Hiroshi et al. looked at the correlation between the CPG ratio and other

clinical variables in patients with diabetes. They found that the average fasted CPG ratio for groups of patients with insulin dependency was 3.35-3.94 nmol/mmol[17].

To acquire the CPG ratio, C-peptide naturally needs to be measured. As of today, national guidelines do not advocate routine measurement of C-peptide levels in primary care. The role of C-peptide is currently limited to unusual or ambiguous cases of diabetes. For example, C-peptide has a role in differentiating T1DM from MODY, where C-peptide levels are persistent over time[18]. Additionally, some studies suggest that C-peptide could be helpful when differentiating between LADA and T2DM, where C-peptide is lower in LADA compared with T2DM[19 20]. Finally, there is some evidence that C-peptide could be useful when selecting therapies, predicting the need for insulin in the future and even the risk of complications[21]. With this in mind, C-peptide and CPG ratio could be more readily available in the future for clinicians, and aid in guidelines relating to diagnosis and management of diabetes.

5.2 Machine learning in medicine

Machine learning is making its way into medicine and is applicable in multiple situations from diagnosis to treatment[22]. From interpreting histological samples to monitoring and predicting the survival rate of ICU patients. The accelerating creation and storing of vast amounts of health care data will probably lead to a further increase in the use of machine learning as a tool in medicine. When introducing machine learning to a new task or subtype of medicine, we suggest the use of an interpretable model rather than a black box. This makes it easier to recognize pitfalls where the algorithm might go wrong and makes it possible for the clinician to see how changes in the variables affect the outcome. In many cases, there is little to no upside in using more complicated black box models compared to simpler ones[23]. For example, a study by Qin Liu et al. found that a standard vector machine (SMV) was more accurate than a Back Propagation Neural Network (BPNN) when predicting critical illness risk in hospitalized patients with COVID-19 pneumonia[24]. A consensus in the machine learning community is that when choosing a model, the best option is the least complicated model, which produces an acceptable accuracy. This usually makes it more reliable and applicable in practice.

5.3 A practical example

An example of how our model, in particular, could be useful in practice:

Autoantibody tests on all patients is the only way to diagnose autoimmune T1DM since autoantibody positivity is required per definition. However, it would be costly to test all patients with symptoms of diabetes and so it is not standard practice. Now, consider a case where a clinician is uncertain of the type of diabetes in a patient. The clinician can then follow the algorithm and get to a prediction. If the algorithm classifies the patient as T1DM, then tests for autoantibodies should be done. This could help decrease the incidence of "missing" autoimmune T1DM and LADA without the full cost of testing all patients for autoantibodies. Correctly diagnosing patients should in theory decrease the number of new visits due to complications induced by sub-optimal treatment due to misdiagnosis, therefore compensating for the extra cost of testing for autoantibodies.

6. Conclusion

In this study, we developed and validated a CART model for differentiating between T1DM and T2DM based on clinical data at the stage of diagnosis (91.83% accuracy with a sensitivity of 0.8997 and specificity of 0.9244). We showed that C-peptide to glucose ratio in our data was more important than glucose and C-peptide alone as a predicting variable at the stage of diagnosis. We have demonstrated that age at diagnosis and BMI at diagnosis are predictive of the type of diabetes, although assessing a diabetes type solely on these variables is prone to misclassification. We have argued that simple machine learning models can be useful as a clinical tool and is better than a black-box. Whether our model will be useful in practice is still up for debate. Previous studies have shown that there are benefits in measuring C-peptide in patients with diabetes. Studies on C-peptide to glucose ratio and its importance for diabetes diagnosis and management, and research to test whether machine learning models could be useful in differentiating between T2DM and LADA or T1DM and MODY is suggested.

7 Limitations

We acknowledge that the training sample is very small and that the DG in testing data is not confirmed by follow-up data, and is merely based on the clinical assessment. This can both under- or overestimate the accuracy of our model. Furthermore, rarer types of diabetes or other autoantibodies found in autoimmune diabetes were not taken into account.

8 Acknowledgments

First and foremost, I would like to thank all the wonderful people at the Botna research group, the personnel at the Diabetes Outpatient clinic at the Meilahti Hospital and the DIREVA personnel for clinically studying the patients. Finally, I would like to express my deepest gratitude to my supervisor Tiinamaija Tuomi who with her guidance, support and understanding carried me through the process of writing this thesis.

References

[1] Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet Med*. 1998;15(7):539-553. doi:10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S

[2] Tuomi T, Santoro N, Caprio S, Cai M, Weng J, Groop L. The many faces of diabetes: a disease with increasing heterogeneity. *Lancet*. 2014;383(9922):1084-1094. doi:10.1016/S0140-6736(13)62219-9

[3] American Diabetes Association Diabetes Care 2015 Jan; 38(Supplement 1): S8-S16. https://doi.org/10.2337/dc15-S005

[4] THL (Finnish institute of health and welfare) https://sotkanet.fi/sotkanet/fi/index

[5] Marja Niemi, Klas Winell STAKES reports/2005 Diabetes in Finland

[6] Shields BM, Peters JL, Cooper C, et al. Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature. *BMJ Open*. 2015;5(11):e009088. Published 2015 Nov 2. doi:10.1136/bmjopen-2015-009088

[7] Maahs DM, West NA, Lawrence JM, Mayer-Davis EJ. Epidemiology of type 1 diabetes. *Endocrinol Metab Clin North Am*. 2010;39(3):481-497. doi:10.1016/j.ecl.2010.05.011

[8] Maahs DM, West NA, Lawrence JM, Mayer-Davis EJ. Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am. 2010;39(3):481-497. doi:10.1016/j.ecl.2010.05.011

[9] Thomas NJ, Jones SE, Weedon MN, Shields BM, Oram RA, Hattersley AT. Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol*. 2018;6(2):122-129. doi:10.1016/S2213-8587(17)30362-5

[10] Chung JO, Cho DH, Chung DJ, Chung MY. Associations among body mass index, insulin resistance, and pancreatic β-cell function in Korean patients with new-onset type 2 diabetes. *Korean J Intern Med*. 2012;27(1):66-71. doi:10.3904/kjim.2012.27.1.66

[11] Vasa Central Hospital webpage: https://www.vaasankeskussairaala.fi/en/for-professionals/recruitment-education-and-development/perusterveydenhuollon-yksikko/direva---diabetesregister/

[12] L. Breiman, J. Freidman, C. Stone (1984) Classification and Regression Trees (Wadsworth Statistics/probability)
https://scholar.google.com/scholar?q=Classification%20and%20Regression%20Trees

[14] Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[15] Shields BM, Peters JL, Cooper C, et al. Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature. *BMJ Open*. 2015;5(11):e009088. Published 2015 Nov 2. doi:10.1136/bmjopen-2015-009088

[16] Yosten GL, Maric-Bilkan C, Luppi P, Wahren J. Physiological effects and therapeutic potential of proinsulin C-peptide. *Am J Physiol Endocrinol Metab*. 2014;307(11):E955-E968. doi:10.1152/ajpendo.00130.2014

[17]Hiroshi BANDO, Koji EBE, Tetsuo MUNETA, Masahiro BANDO, Yoshikazu YONEI (2018) Investigation of Fasting Ratio of C-Peptide/Glucose and Related Markers in Diabetes Archives of Diabetes and Endocrine System Volume 1, Issue 1, 2018, PP: 17-24

[18] Gardner DS, Tai ES. Clinical features and treatment of maturity onset diabetes of the young (MODY). *Diabetes Metab Syndr Obes*. 2012;5:101-108. doi:10.2147/DMSO.S23353

[19] N, S.K., Subhakumari, K.N. Role of anti-GAD, anti-IA2 antibodies and C-peptide in differentiating latent autoimmune diabetes in adults from type 2 diabetes mellitus. *Int J Diabetes Dev Ctries* **36,** 313–319 (2016). https://doi.org/10.1007/s13410-015-0451-8

[20] Arikan E, Sabuncu T, Ozer EM, Hatemi H. The clinical characteristics of latent autoimmune diabetes in adults and its relation with chronic complications in metabolically

poor controlled Turkish patients with Type 2 diabetes mellitus. J Diabetes Complications. 2005 Sep-Oct;19(5):254-8. doi: 10.1016/j.jdiacomp.2005.02.004. PMID: 16112499.

[21] Jones AG, Hattersley AT. The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabet Med*. 2013;30(7):803-817. doi:10.1111/dme.12159

[22] Rajkomar, Dean, and Kohane. Machine Learning in Medicine. New England Journal of Medicine. 2019;380(14): 1347-1358. Doi:10.1056/NEJMra1814259

[23] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[24] Liu Q, Pang B, Li H, et al. Machine learning models for predicting critical illness risk in hospitalized patients with COVID-19 pneumonia. *J Thorac Dis*. 2021;13(2):1215-1229. doi:10.21037/jtd-20-2580