Developing clinical prediction models for diabetes classification and progression

Anita Louise Lynam

PhD thesis 2019

Developing clinical prediction models for diabetes classification and progression

Submitted by Anita Louise Lynam to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies, in September 2019.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Abstract

Patients with type 1 and type 2 diabetes have very different treatment and care requirements. Overlapping phenotypes and lack of clear classification guidelines make it difficult for clinicians to differentiate between type 1 and type 2 diabetes at diagnosis. The rate of glycaemic deterioration is highly variable in patients with type 2 diabetes but there is no single test to accurately identify which patients will progress rapidly to requiring insulin therapy. Incorrect treatment and care decisions in diabetes can have life-threatening consequences.

The aim of this thesis is to develop clinical prediction models that can be incorporated into routine clinical practice to assist clinicians with the classification and care of patient diagnosed with diabetes. We addressed the problem first by integrating features previously associated with classification of type 1 and type 2 diabetes to develop a diagnostic model using logistic regression to identify, at diagnosis, patients with type 1 diabetes. The high performance achieved by this model was comparable to that of machine learning algorithms.

In patients diagnosed with type 2 diabetes, we found that patients who were GADA positive and had genetic susceptibility to type 1 diabetes progressed more rapidly to requiring insulin therapy. We built upon this finding to develop a prognostic model integrating predictive features of glycaemic deterioration to predict early insulin requirement in adults diagnosed with type 2 diabetes.

The three main findings of this thesis have the potential to change the way that patients with diabetes are managed in clinical practice.

Use of the diagnostic model developed to identify patients with type 1 diabetes has the potential to reduce misclassification. Classifying patients according to the model has the benefit of being more akin to the treatment needs of the patient rather than the aetiopathological definitions used in current clinical guidelines. The design of the model lends itself to implementing a triage-based approach to diabetes subtype diagnosis.

Our second main finding alters the clinical implications of a positive GADA test in patients diagnosed with type 2 diabetes. For identifying patients likely to progress rapidly to insulin, genetic testing is only beneficial in patients who test positive for GADA. In clinical practice, a two-step screening process could be implemented - only patients who test positive for GADA in the first step would go on for genetic testing.

The prognostic model can be used in clinical practice to predict a patient's rate of glycaemic deterioration leading to a requirement for insulin. The availability of this data will enable clinical practices to more effectively manage their patient lists, prioritising more intensive follow up for those patients who are at high risk of rapid progression. Patients are likely to benefit from tailored treatment. Another key clinical use of the prognostic model is the identification of patients who would benefit most from GADA testing saving both inconvenience to the patient and a cost-benefit to the health service.

Table of Contents

Acknowledgments

I am sincerely grateful to my supervisors Dr Angus Jones, Dr Beverley Shields and Professor Andrew Hattersley for giving me the wonderful and life changing opportunity to do this PhD. I would like to express my special appreciation to Dr Jones for sharing his diabetes expertise and guiding my statistical writing into something more interpretable for clinical audiences, and to Dr Shields for both sharing her knowledge so generously and providing mentorship on the statistical aspects of my PhD. I would also like to thank my pastoral tutor Professor David Richards for his time, encouragement and words of wisdom that kept me going during my low times.

I have been privileged to work with amazing people in the diabetes research team and thank you all for your collaboration and friendship. I would particularly like to thank Dr John Dennis for being a role model, sharing his statistical expertise and always knowing which Stata or R function I should use, and to Dr Lauric Ferrat for sharing his passion for mathematical modelling with me.

I am also very grateful to collaborators in Dundee, Oxford and The Netherlands for allowing me to use their data and being so helpful answering my queries, and to everyone in the Exeter NIHR Clinical Research Facility team for all their hard work helping me clean and understand the Exeter study data.

Last but not least, to my wonderful husband Neil I send a very special heartfelt thank you for all your emotional and financial support during these past three years. I will be eternally grateful for your patience and encouragement, and for being there to share my highs and lows. I could not have done this without you.

Abbreviations

ADOPT	A Diabetes Outcome Progression Trial
AIC	Akaike information criterion
ALT	Alanine Transaminase
AUPRC	Area Under the Precision Recall Curve
BIC	Bayesian information criterion
BMI	Body Mass Index
CRF	Clinical Research Facility
DARE	Diabetes Alliance for Research in England
DCS	Hoorn Diabetes Care System
DPP4 inhibitor	Dipeptidyl peptidase-4 inhibitor
eGFR:	estimated Glomerular Filtration Rate
GADA	GAD65 autoantibodies
GBM	Gradient Boosting Machine
GLP-1 receptor agonist	Glucagon-like peptide 1 receptor agonist
GoDarts	Genetics of Diabetes Audit and Research in Tayside Scotland
HbA_{1c}	Hemoglobin A1c
HDL	High-Density Lipoprotein
HLA	Human Leukocyte Antigen
HOMA	Homoeostatic Model Assessment
IA-2	Islet Antigen 2
IAA:	Insulin Autoantibodies
ICA	Islet Cell Antibodies
KNN	K-Nearest Neighbours
LADA	Latent Autoimmune Diabetes in Adults
LDL	Low-Density Lipoprotein
LR	Logistic Regression
ML	Machine Learning
MODY	Maturity-Onset Diabetes of the Young
MRC	Medical Research Council
NICE	National Institute for Health and Clinical Excellence
NN	Neural Network
NPV	Negative predictive value
PPV	Positive predictive value
PRIBA	Predicting Response to Incretin Based Agents in Type 2 Diabetes
PROMASTER	PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes
RETROMASTER	RETROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes
RF	Random Forest
ROC AUC	Area Under the Receiver Operating Characteristic Curve
RP	Royston-Parmar flexible parametric survival model
SAID	Severe AutoImmune Diabetes
SGLT2 inhibitors	Sodium-glucose co-transporter-2 inhibitors
SMOTE	Synthetic Minority Over-Sampling Technique

SNP	Single-Nucleotide Polymorphism
STARD	Standards for the Reporting of Diagnostic Accuracy Studies
StartRight	Getting the Right Classification and Treatment From Diagnosis in Adults With Diabetes
SVM	Support Vector Machine
T1D	Type 1 Diabetes
T1D GRS	Type 1 Diabetes Genetic Risk Score
T2D GRS	Type 2 Diabetes Genetic Risk Score
TRIG	Triglycerides
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UCPCR:	Urinary C-peptide Creatinine Ratio
UKPDS	United Kingdom Prospective Diabetes Study
VI	Variable Importance
WHO	World Health Organisation
YDX	Young Diabetes in Oxford
ZnT8	Zinc transporter 8

Chapter 1.

Introduction

1.1 Structure

This chapter is divided into four parts.

First the aims and structure of this thesis are stated. We then present the treatment and management challenges of type 1 and type 2 diabetes that can occur in clinical practice and review the current evidence on clinical features and biomarkers associated with classification and glycaemic progression. Next the challenges associated with the implementation of clinical prediction models into clinical practice are discussed and the key methodological approaches to diagnostic and prognostic model development and validation are reviewed. Finally we introduce the datasets used in subsequent chapters of this thesis.

1.2 Aims and structure of thesis

The overall aim of this thesis is to develop clinical prediction models that can assist with the classification and care of patient diagnosed with diabetes in clinical practice.

The thesis is divided into six chapters.

This chapter (chapter 1) presents an overview of the treatment and management challenges of type 1 and type 2 diabetes that can occur in clinical practice and the opportunity for the development of clinical prediction models.

In Chapter 2, we investigate whether patient clinical features and biomarkers can be used to differentiate between different diabetes subtypes (type 1 and type 2) at diagnosis by applying logistic regression modelling, and validate our model with an independent dataset.

In chapter 3, we compare the performance of five different supervised machine learning algorithms and logistic regression using the diabetes subtype classification example from chapter 2.

In chapter 4, we examine whether common type 1 diabetes genetic variants can predict rapid glycaemic deterioration (time to insulin therapy from diagnosis) over and above GADA testing in patients clinically diagnosed with type 2 diabetes.

In chapter 5, we apply flexible parametric survival analysis to investigate the use of clinical features and biomarkers of patients clinically diagnosed with type 2 diabetes to predict rapid glycaemic deterioration (time to insulin therapy from diagnosis), and validate our findings with an independent dataset.

Chapter 6 is a discussion of the main findings, conclusions, limitations and future work generated by each chapter.

1.3 Treatment and management challenges of type 1 and type 2 diabetes

1.3.1 Overview of diabetes

Diabetes is a disease in which the body's ability to regulate sugar in the blood (glucose) is impaired leading to excess sugar in the blood (hyperglycaemia) which is a cause of serious health conditions such as diabetic retinopathy, nephropathy and neuropathy (1). Type 1 and type 2 are the two major subtypes of diabetes with type 2 being more common.

1.3.2 Type 1 diabetes

Type 1 diabetes is characterised by beta-cell destruction leading to rapid development of near-absolute insulin deficiency (2-5). Patients are often first

diagnosed with diabetes when they present with symptoms such as thirst or tiredness, or in a metabolic crisis.

This severe insulin deficiency leads to acute glucose fluctuations which need to be controlled by continuous glycose monitoring and physiological insulin replacement administered by intensive insulin regimens (multiple daily injections or continuous subcutaneous insulin infusion therapy using an insulin pump) (3, 6). Patients with type 1 diabetes need insulin treatment in the early stages of the disease; without insulin treatment they are at risk of acids building up in the blood (ketoacidosis) which can be life-threatening (7).

1.3.3 Type 2 diabetes

In contrast, type 2 diabetes is a progressive metabolic disease (also called metabolic disorder). Patients with type 2 diabetes can still produce insulin (unlike type 1 diabetes) but their body is unable to use it effectively, the beta cells become exhausted leading to a gradual reduction in the capacity of the beta cells to make insulin (8, 9). Patients with type 2 diabetes do not develop the severe insulin deficiency that is seen in patients with type 1 diabetes, they can usually be successfully treated initially with lifestyle changes or oral agents for many years (10-12) or can even achieve remission by maintained weight loss (13). However, due to this characteristic progressive reduction, many patients will eventually require insulin therapy to maintain glucose control (11).

1.3.4 Importance of diabetes subtype classification for clinical management

Severe insulin deficiency is the fundamental difference between type 1 and type 2 diabetes and it is this deficiency that determines treatment requirements. To ensure that a patient receives the correct treatment it is therefore critical to correctly distinguish between patients with type 1 and type 2 diabetes, at time of

diagnosis, based on a definition that closely aligns to their treatment requirement. However, correctly classifying patients is challenging (14, 15) because of the overlapping phenotypes of these diabetes subtypes (16, 17) and a lack of clear classification guidelines which tend to focus on the aetiopathological definitions rather than ones that relate to treatment requirements (2, 5, 18, 19).

No clinical features and biomarkers provide perfect separation when used in isolation. Individual patients can have some features that would indicate type 1 whilst their other features indicate type 2. Many of the tests that can assist classification are not routinely indicated in clinical practice and the lack of a single diagnostic test that can be used to classify diabetes robustly at diagnosis also makes classification challenging in clinical practice. There are no existing prediction models that can be used for classification of type 1 and type 2 diabetes at diagnosis. All of these challenges combined contribute to the serious and common problem of misclassification in type 1 and type 2 diabetes (20-22). Incorrectly classifying patients with type 1 diabetes as type 2 can have life-threatening consequences; without insulin therapy, patients with type 1 diabetes are at risk of ketoacidosis which left untreated can be fatal (7). Patients with type 2 diabetes incorrectly classified as type 1 will be treated un-necessarily with insulin: as a result, patients may suffer unfounded negative quality of life impacts such as work restrictions and the health service treatment costs are un-necessarily increased.

1.3.5 *Clinical features and biomarkers associated with classification of type 1 and type 2 diabetes*

Clinical features

The use of age of diagnosis and BMI jointly predominates in clinical practice for diabetes classification, with younger age and lower BMI historically associated with type 1 diabetes; both of these features have strong evidence for utility at diagnosis (22) and are easily obtained. However these features are becoming less distinctive, type 2 diabetes is occurring in young patients as obesity levels increase (16) and type 1 diabetes can occur in adults (17).

Presentation of symptoms such as glycaemia, weight loss or ketosis at diagnosis, glucose metabolism and family history of type 1 or type 2 diabetes in first-degree relatives are often used in clinical practice to classify patients but classification based on these features has little or no evidence base (22).

Islet-autoantibodies

The presence of one or more islet-autoantibodies (GAD65 autoantibodies (GADA) (23, 24), Islet Antigen 2 (IA-2) (23, 25), Zinc transporter 8 (ZnT8) (26), insulin autoantibodies (IAA) (23, 27) and Islet Cell Antibodies (ICA) (27)) is a marker of type 1 diabetes. GADA, IA-2 and ZnT8 are the three islet-autoantibodies most often used in clinical practice. Testing for ICA has largely been superseded by testing for individual autoantibodies (GADA, IA-2 and ZnT8).

There are limitations associated with the use of islet-autoantibodies to classify diabetes subtype that mean that widespread testing would not solve misclassification alone. The sensitivity of the tests for these markers performed individually for classification is low: whilst islet-autoantibodies appear early in

life (27) and thus may have utility at diagnosis, islet-autoantibodies are not detectable at diagnosis in all patients with type 1 diabetes and longer durations post diagnosis are associated with higher negativity (28). The presence of multiple islet-autoantibodies increases the specificity of the tests (23, 27) but comprehensive testing is not routinely indicated in clinical practice.

Other potential limitations are that the frequency of islet-autoantibody positivity may differ by ethnicity and/or sex although the evidence for these findings is weak as they are based on small studies (29-31) and IAA testing is only useful at first diagnosis since IAA is increased with exogenous insulin use (32).

Whilst GADA is a marker of type 1 diabetes, many adult patients with type 2 diabetes appear GADA positive (33, 34) but do not have the characteristics associated with type 1 diabetes. GADA is an imperfect diagnostic test: the likelihood of false positive result when testing in an adult population with low prevalence of type 1 diabetes will be high (35). This means that consideration of the prior probability of type 1 diabetes is important when interpreting the results of a single positive autoantibody result. The interpretation may be very different for a patient with low likelihood of type 1 diabetes based on other features such as their age and BMI compared to that of a patient clinically likely to have type 1 diabetes. In the former situation, a positive results is more likely to be false positive result.

Another consideration associated with the use of islet-autoantibodies for diagnostic purposes is the variation in specificity of the test between different laboratories. For example, the range of GADA specificity for the laboratories participating in the 2010 Diabetes Autoantibody Standardisation Programme was from 68% to 100% (36). This variation arises from the use of different

assay formats and the use of different thresholds to define a positive result. Thresholds are usually defined using centiles of titres observed in a non-diabetic population with higher titres increasing the specificity of the test. The 97.5th or 99th centile is normally used but in some cases the assay lowest reportable value has been used.

The limitations and considerations discussed above, and the substantial cost that would be incurred in testing everyone with diabetes for islet-autoantibodies are reasons why routine testing is not currently recommended in clinical practice.

Genetics

There is a strong genetic component to type 1 diabetes which is measurable by single nucleotide polymorphism (SNP) genotyping (37). These SNPs are located in the Human Leukocyte Antigen (HLA) and non-HLA regions with DR3 and DR4-DQ8 alleles in the HLA region being the highest genetic determinants of type 1 diabetes (38, 39). A type 1 diabetes genetic risk score (T1D GRS) consisting of a combination of SNPs from both regions can discriminate between patients with type 1 and type 2 diabetes (37, 40). Advantages of using genotyping are that results do not change over time and susceptible genetic variants are common across ethnicities (41), but it is not currently routinely indicated in clinical practice. There are also common genetic variants associated with type 2 diabetes (42, 43) but a T2D GRS has far less discrimination power than the T1D GRS (37).

C-peptide

C-peptide is a substance made in the pancreas in equal amount to insulin and is a measure of how much insulin is being produced in the body. C-peptide is

measured instead of insulin because it has a longer half-life and is not affected by exogenous insulin. C-peptide measurement is reliable and is widely available in clinical practice, more so than it has been in the past (22).

C-peptide values measured in patients with longstanding diabetes provides a gold-standard test for classifying patients according to their treatment requirements (44). A low C-peptide value measured at any time post diagnosis (< 200 pmol/L (non-fasting)) confirms severe endogenous insulin deficiency (44, 45), the key feature of type 1 diabetes. Patients with low C-peptide (< 200 pmol/L) will have the treatment requirements of type 1 diabetes - high glucose variability and lack of glycaemic response to non-insulin therapies whilst patients with high C-peptide (>600 pmol/L) do not have absolute insulin deficiency but may still require insulin for glucose management. However some patients with type 1 diabetes can retain significant amounts of endogenous insulin for 3 – 5 years (44, 46) particularly if they are obese. This means that there will be some overlap which limits the utility of the test at diagnosis.

Suggested C-peptide threshold for classification of type 1 diabetes based on treatment requirement is < 200 pmol/L (non-fasting)) and > 600 pmol/L (non-fasting) for type 2 diabetes (44, 47, 48). There will be uncertainty in the classification of patients whose C-peptide value is in the intermediate range (>=200 pmol/L and <=600 pmol/L).

1.3.6 Glycaemic deterioration in patients with type 2 diabetes

The clinical course of glycaemic deterioration is highly variable in patients with type 2 diabetes; some patients can be successfully treated without insulin for many years or decades whilst others will need insulin within months of

diagnosis (8, 9, 49). This variability may reflect differences in underlying pathophysiology which is highly heterogeneous in type 2 diabetes (50-54).

In addition to correctly classifying patients, being able to correctly identify, at diagnosis, those patients with type 2 diabetes that are likely to have more rapid glycaemic deterioration may be helpful clinically. In clinical practice, the treatment and management of patients could then be personalised according to their individual risk. For example, patients likely to rapidly progress could be offered more frequent follow up, earlier treatment intensification or targeted treatment with interventions to delay glycaemic progression. In research, high risk patients could be targeted by clinical trials aimed at developing effective therapies to slow progression.

1.3.7 Clinical features and biomarkers independently associated with glycaemic deterioration in patients diagnosed with type 2 diabetes

A number of routinely indicated clinical features and biomarkers have been reported to be associated with glycaemic deterioration (Table 1) in patients with clinically diagnosed type 2 diabetes but some effect sizes are small and not all features investigated are independently associated (55). Despite differences in the definition of glycaemic deterioration, duration of diabetes at start of study, follow up times and cohorts between studies, the association findings for many features are consistent.

Table 1: Studies identifying clinical features and biomarkers independently associated with glycaemic deterioration

Study	Definition of glycaemic deterioration	Start point	Independently associated clinical features and biomarkers
Dennis et al. (51)	HbA_{1c} progression over time	Newly diagnosed	Age at diagnosis
Zhou et al. (55)	Time to insulin therapy	Diagnosis	Age at diagnosis, TRIG, HDL, BMI
Levy et al. (56)	Time to failure of dietary therapy	Newly diagnosed	Fasting glucose, age at diagnosis, beta cell function (measured by OGTT)
Turner et al. (34)	Insulin therapy within six years of diagnosis	Diagnosis	GADA and beta-cell function (measured using HOMA)
Matthews et al. (57)	Sulphonylurea failure within six years of diagnosis	Newly diagnosed	Age at diagnosis, beta cell function (measured using HOMA), fasting glucose, drug treatment
Ringborg et al. (58)	Time to insulin therapy	Initiation of Oral Anti-Diabetic (OAD) treatment (diabetes duration unknown)	Age < 65 years, type of OAD treatment, HbA_{1c}
Cook et al. (59)	Time until HbA_{1c} >= 64 mm/mol (8.0%) or glucose-lowering therapy intensified (insulin or adding a third oral agent).	Initiation of metformin/ sulphonylurea combination therapy (median diabetes duration 3.8 years)	Age, sex, serum creatinine, smoking, HbA_{1c}, diabetes duration

Study	Definition of glycaemic deterioration	Start point	Independently associated clinical features and biomarkers
Pani et al. (60)	HbA_{1c} >= 53 mm/mol (7%) or medical therapy initiation within 1 year.	Post diagnosis (diabetes duration unknown)	HbA_{1c}, age, weight gain
Waldman et al. (61)	Initiation of oral hypoglycaemic agents (OHAs) or insulin.	Post diagnosis (median diabetes duration 5 years)	HDL-C, HDL-C/apolipoprotein A-I
Donnelly et al. (49)	HbA_{1c} progression over time	Diagnosis	GADA, age at diagnosis, BMI, HDL, year of diagnosis
Pilla et al. (62)	Time to insulin initiation	Post diagnosis (mean diabetes duration 5.5 years)	Age, ethnicity, HbA_{1c}, number of drugs, BMI, smoking, hypertension, chronic kidney disease, cardiovascular disease and family history. Number of complications (cardiovascular disease, chronic kidney disease, diabetic neuropathy, and diabetic retinopathy), source of medical care.
Schrijnders et al. (63)	Time needed to treatment intensification with either insulin or oral triple therapy	Post diagnosis (mean diabetes duration 5.5 years)	HbA_{1c}, age at diagnosis.

Study	Definition of glycaemic deterioration	Start point	Independently associated clinical features and biomarkers
Kostev et al. (64)	Insulin initiation within 6 years	First-line prescriptions of metformin or sulfonylureas (mean diabetes duration 1 year)	First line drug In patients treated with metformin: eGFR, sex, source of medical care, history of stroke, prescription of diuretics and statins, age and diagnosed hyperlipidaemia In patients treated with sulfonylureas: high eGFR, diagnosed congestive heart failure and prescriptions of diuretics
Gentile et al. (65)	Insulin initiation within 5 years	First eGFR evaluation (mean diabetes duration 7 years)	Duration, HbA_{1c}, TRIG, HDL ,age, drug (diabetes and lipid-lowering (statins)), comorbidities (retinopathy), LDL, BMI, eGFR

Clinical features

Multiple studies have shown that age at diagnosis (49, 51, 55-57, 63), and BMI (49, 62, 65) are independently associated with faster glycaemic deterioration. Findings for smoking status are conflicting with an independent association found in some studies (59, 62) but not in others (55, 60). Findings for sex were also inconsistent, some found an association with glycaemic progression, (59, 64) but another did not (58). Black and Hispanic ethnicity have been associated with a lower risk of insulin initiation than white ethnicity whilst a family history of diabetes was associated with higher risk (62). All of the above features are easily obtained and have utility at diagnosis.

Biomarkers

There is strong evidence that high baseline HbA_{1c} (58-60, 62, 63, 65) is associated with increased glycaemic deterioration, HbA_{1c} is routinely measured in clinical practice and has utility at diagnosis. Hypertension (systolic blood pressure > 140 mmHg) (62) and higher fasting glucose (56, 57) are also associated with faster deterioration. Triglycerides (TRIG) (55, 65), Low-Density Lipoprotein (LDL) (65) and High-Density Lipoprotein (HDL) (49, 55, 61, 65) are independently associated with deterioration. Limitations of the use of these lipids tests are the strong collinearity between TRIG and HDL (55), and the requirement for patients to fast prior to TRIG measurement.

Islet-autoantibodies

The presence of islet-autoantibodies (GADA (34, 49) and IA-2 (66, 67)) have been independently associated with rapid glycaemic deterioration in participants with type 2 diabetes. In addition, the presence of GADA islet-autoantibodies in adult patients with a clinical diagnosis of type 2 diabetes has been used to define Latent Autoimmune Diabetes in Adults (LADA) (24, 68-71). The common

view is that patients classified as LADA have a homogenous intermediate phenotype but there has recently been an opposing view that LADA is not in fact homogenous, rather LADA is likely to reflect a combination of two heterogeneous populations with very different phenotypes – true positives (type 1 diabetes) and false positives (type 2 diabetes) (35).

A limitation of the use of GADA and IA-2 for identifying patients likely to have more rapid glycaemic deterioration is that testing is not currently recommended in clinical practice so not all patients will be tested. As discussed in section 1.3.5, the prior prevalence and specificity of the tests need to be considered when interpreting a positive result.

Genetics

It has previously been shown that genetic variants in the HLA region associated with type 1 diabetes may alter the risk of rapid progression to insulin in patients with type 2 diabetes who are GADA positive (72-74) and that high risk HLA is associated with low C-peptide in a type 2 diabetes cohort (74). There are no studies however that have specifically examined the association between T1D GRS, which includes HLA specific SNP's, and rapid progression in patients with type 2 diabetes. Two studies found no association between glycaemic deterioration and gene variants associated with type 2 diabetes (55, 75).

Other features that have been associated with glycaemic deterioration include: type of initial treatment (58), diabetes duration (65), drug therapy (62, 64, 65), diabetes complications (62, 64, 65), source of medical care (62, 64), estimated Glomerular Filtration Rate (eGFR) (64, 65) and serum creatinine (59) but these are not discussed further as they cannot be used and are not useful at diagnosis. Beta-cell function assessed by HOMA (34, 57) or oral glucose

tolerance tests (OGTT) (56) have been associated with glycaemic deterioration but the former test is not routinely available in clinical practice and the latter is now only measured for the diagnosis of gestational diabetes.

1.3.8 Conclusion

There is evidence that many clinical features and biomarkers are associated with classification of type 1 and type 2 diabetes and glycaemic progression in patients with type 2 diabetes. The diagnostic or prognostic accuracy of these clinical features and biomarkers used in isolation may be improved by the use of a holistic approach in which these clinical features and biomarkers are combined.

The most effective approach is to combine these features in multivariable prediction models as is now common in many areas of clinical practice. There are however, currently no diagnostic or prognostic prediction models to help clinicians distinguish between type 1 and type 2 diabetes subtypes or to predict rapid progression in patients with type 2 diabetes.

1.4 Clinical prediction model concepts, and methods used for development, validation and reporting

1.4.1 Overview of clinical prediction models

Clinical prediction models have a grounding in evidence-based medicine. They provide clinicians with external evidence of the probability of a particular outcome for an individual patient that can be taken into consideration when making treatment or testing choices. This outcome may be the presence or absence of a disease or condition (diagnostic) or the future development of a disease, event or complication (prognostic) (76).

Literature on prediction models has increased over time (77), in 2010 there were 101 publications listed on PubMed with the terms “prediction model” or “prognostic model” compared to 410 listed in 2018. In all areas of medicine, including diabetes, clinical prediction models have been implemented as websites and applications, and many incorporated into clinical guidelines (78-87). Models associated with diabetes include a diagnostic model to identify monogenic forms of diabetes in patients with young-onset diabetes prognostic models (88) and prognostic models to predict the risk of type 2 diabetes (89-91) and the risk of glycaemic deterioration (49, 65). A classification tool based on five diabetes clusters (50) has also been developed. A model to identify undiagnosed type 2 diabetes (92) has been incorporated into NICE guidance. A search for “Diabetes Mellitus” disease on MDCalc (93) returned 23 medical apps.

However, the number of models actually implemented into clinical practice is very low compared to the number of models developed and published. It has been suggested that between 1993 and 2011, models for diabetes were being published at a rate of about one every three weeks (89). There are several plausible explanations for this low implementation rate; the inclusion of predictors that are not routinely indicated in clinical practice render models unusable and a lack of clinical credibility and evidence reduces confidence in the model (94). Chapter 1 has already discussed the clinical utility of various features with prior evidence of an association with diabetes classification and progression. Clinical credibility is concerned with the validity of the model development including adherence to model assumptions and interpretability of the model whilst evidence is concerned with the performance and accuracy of the model.

1.4.2 Reporting guidelines for diagnostic and prognostic studies

There are a number of systematic reviews that have performed critical appraisals of the development, validation and reporting of clinical prediction models (95-102). The identification of shortcomings in many of the studies evaluated in these reviews has led to the introduction of reporting guidelines (103). Adherence to these guidelines is now included in the author instructions for most journals when submitting papers addressing development and/or validation of prediction models. The relevant reporting guidelines for diagnostic and prognostic models are Standards for the Reporting of Diagnostic Accuracy Studies (STARD) (104) and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) (105-107). STARD was first published in 2003 and updated in 2015; its objective is to "improve the completeness and transparency of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalisability (external validity)." (108). The TRIPOD statement was published in 2015 and has the reporting of model development, validation or updates as its focus.

1.4.3 Statistical model concepts

The focus of this thesis is multivariable models (multiple predictors) developed using clinical study data.

Logistic regression

In situations where the outcome of interest is binary, the most commonly applied statistical model is binary logistic regression (LR). The technical details of LR have been well documented (109-114). In brief, LR is a form of generalized linear model with a logit link function allowing data to be modelled

on the log odds scale. As this is a linear model, linearity (linear relationship between any continuous predictors and the logit of the outcome variable) is a strong assumption. There are various methods to deal with non-linearity (simple transformations, restricted cubic splines, fractional polynomials (77)) although, to reduce the risk of overfitting, dealing with non-linearity is only recommended for strong predictors or in predictors where non-linearity is known to be likely (77). Overfitting arises from either model or parameter uncertainty and is a significant issue in regression modelling (77). A model that is over-fitted will have lower performance when it is applied to subjects outside the study data.

Survival models

Cox proportional hazards regression (115) is the most well-known and commonly applied model for time-to-event (survival) outcomes in medical studies where there is censored data. There are however limitations of its use for individual patient predictions including; 1) it requires a strong proportional hazards assumption of no time/predictor interaction, 2) it requires a non-parametric estimate for the baseline hazard to obtain survival/failures probabilities over time for individual patients, 3) the non-parametric baseline hazard is a noisy step function and 4) the model is fitted very closely to the data so may not perform well in external data (116). An alternative model that deals with these limitations is a Royston-Parmar flexible parametric survival model (RP) (116, 117). The main feature of this model is the use of restricted cubic splines to flex the Weibull baseline hazard function allowing for complex baseline hazards (118). These models can also be extended to incorporate time-dependent effects (predictors whose regression coefficients vary over time). The parametric nature of these models make them preferable when individual predictions are required, however the use of these models in

prognostic studies is currently very limited, only 12 studies were identified in a systematic review covering the period 2001 to 2016 (119). Other considerations in survival models are the choice of endpoint, censoring, restricting follow-up times and dealing with competing risks (120, 121).

1.4.4 Machine learning applications

Logistic regression and Cox proportional hazards regression are well established statistical models with strong theoretical backgrounds but in recent years there has been an increased interest in the use of machine learning algorithms as an alternative for developing clinical prediction models.

Supervised machine learning algorithms have been used to develop diagnostic and prognostic clinical prediction models in medical applications such as predicting diabetes, assessing fracture risk, fibromyalgia diagnosis, genetics and cancer mortality (122-127). The advantages of using these machine learning algorithms over classic statistical models are their ability to process vast amounts of data such as medical images, biobank and electronic health care records, and their ability to deal with complex interactions and non-linearity. Estimation biases resulting from mis-specified statistical models are avoided by the use of machine learning algorithms since they are non-parametric (128) and they have greater modelling flexibility through the use of tuning parameters (129). However the main disadvantage with the use of machine learning is the lack of transparency which makes them difficult to interpret (130), limiting their clinical credibility.

There have been many applied studies comparing the performance of machine learning to traditional statistical models but their findings are inconsistent (131-

142). Many of the comparison studies have limitations that render them at high risk of bias (143).

1.4.5 Key methodological approaches to diagnostic and prognostic model development and validation

In an attempt to improve the methodological standard of the development and validation of prediction models, several studies and books have published frameworks (76, 144-147) or recommendations (77, 148) for best practice. This thesis provides a brief overview of the key methodological approaches to diagnostic and prognostic model development and validation.

Missing data

Many real world datasets used for developing prediction models will have missing data. The most common approach to missing data is to use a complete-case analysis whereby all observations that have missing values for the variables of interest are excluded from the analysis. This approach may be acceptable when the amount of missing data is either low (149) or when the data is assumed to be missing completely at random (MCAR) (77). The main disadvantage with this approach is that the sample size is reduced and bias is introduced if the assumption of MCAR is not valid. More complex approaches for dealing with missing data may be required in situations where bias is likely (150-153).

Effective sample size

An effective sample size is to some extent, governed by the complexity of the research question (77). To avoid overfitting, the minimum training data size for different model types is determined by the number of events rather than the number of observations (149). The general rule of thumb is 10 events per

variable (EPV) (154) but there is much debate over the rationale for this number (155, 156). An alternative approach for determining the EPV in binary outcome models based on the number of predictors, total sample size and events fraction has been proposed (155). To achieve the effective sample size, data reduction techniques can be applied to adjust the number of predictors.

Selection and coding of predictors

There are various strategies for selecting predictors for inclusion in the model (157). Univariate screening is often used for pre-selecting predictors prior to multivariable modelling but is generally not recommended (158) whilst selection of predictors for inclusion in the model based on previous evidence and expert clinical knowledge, regardless of their statistical significance in the model, has been recommended (149). Another strategy that is often used is a significance criteria strategy that is based on hypothesis testing; predictors are included or excluded from the model using iterative stepwise selection methods. An approach that uses a significance criteria strategy but forces predictors with prior evidence or expert knowledge into the model has also been previously used (159). An alternative approach is to use an information criteria strategy that is based on selecting the best model from a set of several models using Akaike information criterion (AIC) or Bayesian information criterion (BIC) where more complex models are penalized.

In the development of clinical prediction models, testing for non-linearity in continuous predictors has become standard practice. Dichotomisation or grouping of continuous predictors often occurs in medical research but leads to a loss of information (160). This practice is appropriate in some situations, for example to replicate use of test results in clinical practice. It is not however appropriate to deal with non-linearity, non-linearity is better modelled using

simple transformations, restricted cubic splines (77, 149) or fractional polynomials (77, 161). A comparison of AIC or BIC between models developed using linear and different non-linear functions can be used to determine the most appropriate relationship.

In multivariable models there is the possibility of interactions between the variables, for example age may have a stronger effect for males compared to females or an interaction including time in survival models. Interaction terms can be included in the model (time-dependent effects for survival models) but their use can lead to overfitting and overly complex models.

Model performance

There are several performance measures that can be used to assess the quality of the model, these fall into three main aspects; overall performance (distance between predictions and actual outcomes), discrimination (separation of patients with and without the outcome) and calibration (predictions versus observed outcomes). A summary of common performance measures used in medical research is included in Table 2 (77).

Table 2: Summary of performance measures (adapted from Steyerberg) (76)

Measure	Advantages	Disadvantages
Overall performance		
R^2	Commonly used to express amount of explained variation. Can be used to for model comparison. R^2_D available for RP models (162).	Cannot be used to compare models from different populations/datasets. Difficult to interpret. Many different calculations available (e.g. Cox-Snell and Nagelkerke's R^2). Can be used for survival models. Nagelkerke's R^2 severely penalizes false predictions close to 0% and 100%.
Brier score	Less severe in penalizing false predictions close to 0% and 100% than Nagelkerke's R^2.	Interpretation of score depends of the prevalence but can be scaled between 0% and 100%. Calibration component of the Brier score can be tested using Spiegelhalter's z-test. Cannot be calculated for survival models.
Discrimination		
Concordance (c) statistic (C-index) for logistic regression	Rank order statistic insensitive to prevalence. Can be visualized using ROC curve. Well established.	Related to variance of predictors (163). Interpretation varies by clinical area and is based on artificial concept.
Harrell's C-statistic for survival models	Indicates the rank order of the proportion of all pairs that can be ordered.	Some pairs cannot be ordered. Cannot be used if time-dependent effects are used in the model (164).

Measure	Advantages	Disadvantages
D-statistic for survival models (162)	Can be used to calculate a R^2_D which is easier to interpret.	Hard to interpret. Interpretation is based on two created groups (based on the model output) and the model scale. Not well established.
Calibration		
Calibration in the large	Can be visualized in a calibration plot. Indicates if predictions are systematically too low or too high. Statistical testing of the difference in log odds between predictions and observed outcomes is possible	Accurate by design in apparent validation
Calibration slope	Can be visualized in a calibration plot. Indicates under or over fitting. Statistical testing of the deviation of the slope from 1 (miscalibration) is possible.	Accurate by design in apparent validation
Hosmer and Lemeshow	Can be visualised. Goodness of fit test for logistic regression.	Sensitive to sample size and number of groups. Limited power in small samples. Interpretation is difficult. Cannot be calculated for survival models.
Ratio of expected and observed number of events (E/O)	Easy to calculate. Can be used for survival models using expected and observed event probabilities rather than number of events.	Ratio is affected by prevalence.

Assessment of overly influential observations

All observations used to develop a model will influence the fit to some extent but it is possible that some observations will overly influence the model. This may be related to the data quality such as insufficient observations, data errors and extreme predictor values or it may be the case that the data may contain unusual observations where the relationship between the predictors and outcome differ from that observed in the majority of observations (149). In regression models, diagnostic statistics can be used to identify influential observations (165, 166) however the values used to classify an observation as influential are subjective. Careful consideration is also needed on how to deal with influential observations; removing such observations is not generally recommended as this may artificially inflate the predictive accuracy of the model (149).

Clinical usefulness

Clinical usefulness has been defined as the improvement in classification derived from the use of a prediction model above some default position or rule that does not use the said model (77). Clinical usefulness measures fall into two main types; traditional methods using a set threshold selected using either an intuitive or optimal approach, or those that are derived using a decision-analytic approach (77). Measures using the former approach are often used to evaluate prediction modes and are well embedded in clinical use but there are no defined thresholds to indicate clinical usefulness (167). The latter approach involves incorporating the harm and benefit of a decision based on the prediction model into the assessment of clinical usefulness. An example of a harm is an un-necessary operation or medication and a benefit is correct diagnosis of a disease. This net benefit approach has advantages over the first approach

including capturing the clinical consequences of the prediction model in the assessment (167) and the use of decision curves to consider a range of thresholds but quantifying the harms and benefits is a significant limitation.

Model validation

The purpose of model validation is to provide evidence for the performance and accuracy of the model. Model validation comprises of two aspects; internal and external validation.

Internal validation

Internal validation is where the model performance is assessed using the same dataset that was used to develop the model. Several techniques exists for internal validation, the difference between them being the specification of the samples used to both develop and validate the model (77). Apparent validation is a technique where the entire dataset is used to develop the model, the same dataset used to develop the model is then used to assess the model performance. The advantages of this method is that the development sample size is maximised and the assessment of performance is stable but the performance estimate will be overly optimistic. Calibration in the large and calibration slope validation tests are not useful when using apparent validation as they will be accurate by design (77).

Another internal validation technique that is often used in medical research is split-sample validation. The model is developed using a random or stratified subset of the original dataset (classically 50 - 70%) and the model performance assessed using the remaining data. There are numerous issues with the use of this method related to variance and bias (149). With the availability of more efficient techniques such as bootstrapping, the use of split-sample validation is

not now generally recommended (77, 168). Cross-validation is an internal validation technique that is related to split-sample validation but has an advantage of using a larger subset of the original dataset for model development. It uses the same approach of developing the model on a random subset of the original dataset and evaluating the model on the remaining data but this process is repeated several times so that every patient in the original dataset is included at least once in the model assessment. To achieve stable results, the whole cross-validation process may have to be repeated as many as 50 times (77). Ten-fold cross-validation is the most common cross-validation method where the original data is divided into ten equal sized groups or folds; the first group is used to validate the model and the other nine groups are used to develop the model. This process is repeated ten times with a different group used each time for the validation, the performance estimate is an average of the estimates from each round of validation. Jack-knife cross validation is an extreme version of the ten-fold cross validation where only one patient at a time is left out of the development group, this method is not efficient with large number of patients and can underestimate model variability (77).

Bootstrap validation is much the preferred internal validation technique having many advantages over the other techniques such as dealing with model uncertainty and estimate stability. In bootstrap validation, samples are drawn with replacement from the original dataset. For each bootstrapped sample, a model is developed and then evaluated in both the same bootstrapped sample (apparent validation) and the original dataset (test validation). The difference between the two sets of results indicates the amount of optimism which can be used to derive optimism adjusted performance estimates.

External validation

External validation is considered the best quality validation technique in a suggested hierarchy of various validation techniques (149). External validation involves evaluating the model using a separate dataset from the dataset used to develop the model, in which the patients are different in some respect from the patients used to develop the model. The nature of the external validation can be temporal (model validated on new patients recruited to the study), geographic (model validated on patients from another study centre) or fully independent (model validation undertaken by independent researchers) (77).

1.4.6 Conclusion

Clinical prediction models are a valuable commodity in clinical practice. Many different diagnostic and prognostic models have been developed in all areas of medicine using both traditional statistical methods and machine learning but the number actually implemented into clinical use is comparatively low. The lack of clinical uptake can be due to the inclusion of predictors that are not routinely indicated in clinical practice but also a lack of clinical credibility and evidence. Reporting guidelines have been introduced to address the lack of clinical credibility and evidence in the development and validation of clinical prediction models. Several studies and books have published frameworks and recommendations for best methodological approaches to model development and validation, Frank Harrell and Ewout Steyerberg being key leaders in this field.

1.5 Data overview

This section provides an overview of the different datasets used in the subsequent chapters of this thesis. All of the datasets were obtained from existing diabetes studies that recruited adults with a clinical diagnosis of either type 1 or type 2 diabetes.

1.5.1 Datasets

Diabetes Alliance for Research in England (DARE)

DARE (2007 - 2017) was a cross-sectional study designed to explore the causes and complications of diabetes (169). Patients with any type of diabetes were recruited from primary and secondary care in eight diabetes research regions across England. Clinical measurements and blood were collected at recruitment and ongoing biochemical data collected from pathology laboratories. Within the dataset, data were accessible for approx. 6,000 Exeter-recruited participants.

Predicting Response to Incretin Based Agents in type 2 Diabetes (PRIBA)

PRIBA (2011 – 2013) was a prospective study of 957 adult participants with a clinical diagnosis of type 2 diabetes starting DPP4 inhibitors or GLP-1 receptor agonist treatment as part of their normal care. Patients were recruited from primary or secondary care in South West England. The primary analysis was the relationship between insulin secretion (measured by blood C-peptide or Urinary C-peptide Creatinine Ratio (UCPCR)) and glycaemic response (measured by HbA_{1c}) (170). Clinical measurements and blood were taken at the initial visit and follow up clinical measurements and blood collected at three and six months.

MRC PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (PROMASTER)

PROMASTER (2013 – 2015) was an observational study of 820 adult participants with clinically diagnosed type 2 diabetes starting second or third line glucose lowering treatment (Sulphonylurea, DPP-4 inhibitors, GLP-1R agonists, SGLT2 inhibitors, Glitazone or insulin) as part of their normal care (171). Patients were recruited from primary or secondary in South West England, Tayside, Oxford, Glasgow, London and Newcastle. The primary outcome of the study was a comparison of two groups of participants; those who showed a good response to the treatment and those who had a poor treatment response. Clinical measures, fasting blood and urine samples were taken at first visit and repeated at second visit approx. six months after starting the new treatment to measure response.

MRC Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (RetroMASTER)

RetroMASTER (2013 – 2015) is an observational study of 562 participants with clinically diagnosed type 2 diabetes that were being treated with a second or third line glucose lowering treatment (Sulphonylurea, DPP-4 inhibitors, GLP-1R agonists, SGLT2 inhibitors, Glitazone or insulin) as part of their normal care for at least four months (172). Participants were grouped according to their rate of diabetes progression (rapid or slow progression to insulin therapy (<7, >7 years respectively)). Participants were recruited from primary and secondary care in Exeter, Oxford and Dundee. The primary outcome of the study was to compare the clinical characteristic of the two groups of participants. Fasting blood, urine samples and standard biomarkers were collected, along with medical and prescribing history.

MRC Crossover

Crossover (2013 -2015) was an intervention study of 143 adult participants clinically diagnosed with type 2 diabetes and treated with sulphonylurea tablets as their normal care (173). Patients were recruited in Exeter and Tayside. The primary outcome of the study was to understand individual variation in altered glycaemic response to two different treatments. The study had a cross-over model where patients were randomised to be treated for periods of four weeks with Gliclazide (DPP-IV thera) or Sitagliptin (sulphonylurea) in a crossover fashion. Clinical measurements were collected and a mixed-meal test was performed at baseline and at each study drug visits. Fasting blood was collected at each cross-over.

Genetics of Diabetes Audit and Research Tayside Study (GoDarts)

GoDarts is an observational case-control study comprising of patients with a clinical diagnosis of type 2 diabetes recruited from primary and secondary care in the Tayside area of Scotland since 1998. The primary aim of GoDarts is to provide a database which can be used to investigate the genetics, complications and treatment of type 2 diabetes (174). Cross-sectional baseline data, including a blood sample and clinical and lifestyle factors, collected in the study is linked to individual electronic medical records which includes laboratory data, prescription history and hospital admissions making GoDarts a longitudinal cohort (175). Data is available on approx. 10,000 participants.

Young Diabetes in Oxford (YDX)

YDX is a cross-sectional study of participants diagnosed with diabetes (of any type) up to the age of 45 years. Participants were recruited from primary and secondary care in Oxfordshire. One of the aims of the study was to identify clinical features that could be used to pre-select patients at high risk of Maturity-

Onset Diabetes of the Young (MODY) for genetic testing. Data was accessible for 1,200 participants screened between 2005 and 2017.

Hoorn Diabetes Care System (DCS)

DCS is a prospective cohort study representing the data of over 12,000 participants clinically diagnosed with type 2 diabetes in West-Friesland, Netherlands since 1998 (176). The longitudinal dataset contains baseline clinical measurements and annual follow-up visit data. Additional health data, including prescription history and cause-specific mortality, is collected using electronic medical record linkage.

A Diabetes Outcome Progression Trial (ADOPT)

ADOPT (2000 – 2006) was an intention to treat randomised drug efficacy trial in adult patients who had been recently diagnosed with type 2 diabetes (177-179). Patients were recruited in 488 centres in the US, Canada and Europe, a total of 4,360 participants underwent randomisation (179). The study was designed to compare glycaemic control (long-term blood glucose) of participants treated with alternative therapies (thiazolidinedione (rosiglitazone), metformin and sulfonylurea (glibenclamide)). The primary outcome was time to monotherapy failure defined by confirmed level of fasting plasma glucose of more than 180 mg/dl (10.0 mmol/l) (177). Baseline data collected included biomarkers such as lipids and GADA (179).

1.5.2 Data preparation

Each dataset was supplied individually by the study and imported into Stata/SE 15.1 (StataCorp, College Station, TX) except for the ADOPT data which was accessed through the Clinical Trial Data Transparency Portal under approval from GSK (Proposal 930).

In DARE, additional data including C-peptide, HbA_{1c} and islet-autoantibodies was obtained from electronic patient medical records where available. Islet-autoantibody testing and genotyping were requested and performed for those DARE participants where these data were missing but whose blood serum had been stored.

Some participants had been recruited to more than one of the studies included in this thesis. Duplicate participant observations were removed where the analysis was performed on a merged dataset, DARE data took precedence when this occurred.

Potential data errors or inconsistencies identified in the Exeter-based datasets were checked against paper or electronic study records and amended accordingly.

1.6 References

1. Fowler MJ. Microvascular and Macrovascular Complications of Diabetes. Clinical Diabetes. 2008;26(2):77.
2. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes - 2019. Diabetes care. 2019;42(Supplement 1):S13.
3. DeWitt DE, Hirsch IB. Outpatient insulin therapy in type 1 and type 2 diabetes mellitus: Scientific review. JAMA. 2003;289(17):2254-64.
4. Oram RA, Jones AG, Besser REJ, Knight BA, Shields BM, Brown RJ, et al. The majority of patients with long-duration type 1 diabetes are insulin microsecretors and have functioning beta cells. Diabetologia. 2014;57(1):187-91.
5. World Health Organization. Classification of diabetes mellitus. World Health Organization 2019 [Available from: https://apps.who.int/iris/handle/10665/325182 License: CC BY-NC-SA 3.0 IGO.
6. American Diabetes Association. 8. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2018. Diabetes care. 2018;41(Supplement 1):S73 - S85.
7. American Diabetes Association. Standards of Medical Care in Diabetes—2013. Diabetes care. 2013;36(Supplement 1):S11.
8. U.K. Prospective Diabetes Study Group. U.K. Prospective Diabetes Study 16: Overview of 6 Years' therapy of Type II Diabetes: A Progressive Disease. Diabetes. 1995;44(11):1249.
9. Fonseca VA. Defining and Characterizing the Progression of Type 2 Diabetes. Diabetes care. 2009;32(suppl 2):S151.
10. Hope SV, Jones AG, Goodchild E, Shepherd M, Besser REJ, Shields B, et al. Urinary C-peptide creatinine ratio detects absolute insulin deficiency in Type 2 diabetes. Diabetic Medicine. 2013;30(11):1342-8.
11. Inzucchi SE, Bergenstal RM, Buse JB, Diamant M, Ferrannini E, Nauck M, et al. Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia. 2012;55(6):1577-96.
12. Yki-Järvinen H. Combination Therapies With Insulin in Type 2 Diabetes. Diabetes care. 2001;24(4):758.
13. Lean MEJ, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Durability of a primary care-led weight-management intervention for remission of type 2 diabetes: 2-year results of the DiRECT open-label, cluster-randomised trial. The Lancet Diabetes & Endocrinology. 2019;7(5):344-55.
14. Largay J. Case Study: New-Onset Diabetes: How to Tell the Difference Between Type 1 and Type 2 Diabetes. Clinical Diabetes. 2012;30(1):25.
15. Hope SV, Wienand-Barnett S, Shepherd M, King SM, Fox C, Khunti K, et al. Practical Classification Guidelines for Diabetes in patients treated with insulin: a cross-sectional study of the accuracy of diabetes diagnosis. The British journal of general practice : the journal of the Royal College of General Practitioners. 2016;66(646):e315-22.
16. Rosenbloom AL, Joe JR, Young RS, Winter WE. Emerging epidemic of type 2 diabetes in youth. Diabetes care. 1999;22(2):345.
17. 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. The Lancet Diabetes & Endocrinology. 2018;6(2):122-9.
18. Jones AG, Besser REJ, Shields BM, McDonald TJ, Hope SV, Knight BA, et al. Assessment of endogenous insulin secretion in insulin treated diabetes predicts postprandial glucose and treatment response to prandial insulin. BMC Endocrine Disorders. 2012;12(1):6.
19. National Institute for Health and Care Excellence. Type 1 diabetes in adults: diagnosis and management (NICE guideline NG17) 2015 [Cited 14/08/2018]. Available from: https://www.nice.org.uk/guidance/ng17.
20. Farmer A, Fox R. Diagnosis, classification, and treatment of diabetes. BMJ. 2011;342.
21. Stone MA, Camosso-Stefinovic J, Wilkinson J, De Lusignan S, Hattersley AT, Khunti K. Incorrect and incomplete coding and classification of diabetes: a systematic review. Diabetic Medicine. 2010;27(5):491-7.
22. Shields BM, Peters JL, Cooper C, Lowe J, Knight BA, Powell RJ, 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).
23. McDonald TJ, Colclough K, Brown R, Shields B, Shepherd M, Bingley P, et al. Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabetic medicine : a journal of the British Diabetic Association. 2011;28(9):1028-33.
24. Tuomi T, Groop LC, Zimmet PZ, Rowley MJ, Knowles W, Mackay IR. Antibodies to Glutamic Acid Decarboxylase Reveal Latent Autoimmune Diabetes Mellitus in Adults With a Non—Insulin-Dependent Onset of Disease. Diabetes. 1993;42(2):359.
25. Bonifacio E, Lampasona V, Bingley PJ. IA-2 (islet cell antigen 512) is the primary target of humoral autoimmunity against type 1 diabetes-associated tyrosine phosphatase autoantigens. Journal of immunology (Baltimore, Md : 1950). 1998;161(5):2648-54.
26. Wenzlau JM, Frisch LM, Gardner TJ, Sarkar S, Hutton JC, Davidson HW. Novel antigens in type 1 diabetes: the importance of ZnT8. Curr Diab Rep. 2009;9(2):105-12.
27. Bingley PJ. Clinical applications of diabetes antibody testing. The Journal of clinical endocrinology and metabolism. 2010;95(1):25-33.
28. Tridgell DM, Spiekerman C, Wang RS, Greenbaum CJ. Interaction of Onset and Duration of Diabetes on the Percent of GAD and IA-2 Antibody–Positive Subjects in the Type 1 Diabetes Genetics Consortium Database. Diabetes care. 2011;34(4):988.
29. Zimmet PZ. The Pathogenesis and Prevention of Diabetes in Adults: Genes, autoimmunity, and demography. Diabetes care. 1995;18(7):1050.
30. Zimmet PZ, Elliott RB, Mackay IR, Tuomi T, Rowley MJ, Pilcher CC, et al. Autoantibodies to glutamic acid decarboxylase and insulin in islet cell antibody positive presymptomatic type 1 diabetes mellitus: frequency and segregation by age and gender. Diabetic medicine : a journal of the British Diabetic Association. 1994;11(9):866-71.
31. Zimmet PZ, Rowley MJ, Mackay IR, Knowles WJ, Chen QY, Chapman LH, et al. The ethnic distribution of antibodies to glutamic acid decarboxylase: presence and levels of insulin-dependent diabetes mellitus in Europid and Asian subjects. Journal of diabetes and its complications. 1993;7(1):1-7.
32. Sutton M, Klaff LJ, Asplin CM, Clemons P, Tatpati O, Lyen K, et al. Insulin autoantibodies at diagnosis of insulin-dependent diabetes: effect on the

antibody response to insulin treatment. Metabolism: clinical and experimental. 1988;37(11):1005-7.
33. Niskanen LK, Tuomi T, Karjalainen J, Groop LC, Uusitupa MIJ. GAD Antibodies in NIDDM: Ten-year follow-up from the diagnosis. Diabetes care. 1995;18(12):1557.
34. Turner R, Stratton I, Horton V, Manley S, Zimmet P, Mackay IR, et al. UKPDS 25: autoantibodies to islet-cell cytoplasm and glutamic acid decarboxylase for prediction of insulin requirement in type 2 diabetes. The Lancet. 1997;350(9087):1288-93.
35. Jones A, McDonald T, Shields B, Hattersley A. Latent Autoimmune Diabetes of Adults (LADA) represents a mixed population of autoimmune (type 1) and non-autoimmune (type 2) diabetes rather than an intermediate phenotype. Currently in review.
36. Williams AJ, Lampasona V, Schlosser M, Mueller PW, Pittman DL, Winter WE, et al. Detection of Antibodies Directed to the N-Terminal Region of GAD Is Dependent on Assay Format and Contributes to Differences in the Specificity of GAD Autoantibody Assays for Type 1 Diabetes. Diabetes. 2015;64(9):3239-46.
37. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, et al. A Type 1 diabetes genetic risk score can aid discrimination between Type 1 and Type 2 diabetes in young adults. Diabetes care. 2016;39(3):337-44.
38. Noble JA, Valdes AM. Genetics of the HLA region in the prediction of type 1 diabetes. Curr Diab Rep. 2011;11(6):533-42.
39. Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, et al. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. Diabetes. 2008;57(4):1084-92.
40. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, et al. Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. Diabetes care. 2019;42(2):200-7.
41. Golden SH, Brown A, Cauley JA, Chin MH, Gary-Webb TL, Kim C, et al. Health disparities in endocrine disorders: biological, clinical, and nonclinical factors--an Endocrine Society scientific statement. The Journal of clinical endocrinology and metabolism. 2012;97(9):E1579-639.
42. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012;44(9):981-90.
43. Morris AP. Fine mapping of type 2 diabetes susceptibility loci. Curr Diab Rep. 2014;14(11):549.
44. Jones AG, Hattersley AT. The clinical utility of C-peptide measurement in the care of patients with diabetes. Diabetic Medicine. 2013;30(7):803-17.
45. Leighton E, Sainsbury CA, Jones GC. A Practical Review of C-Peptide Testing in Diabetes. Diabetes Ther. 2017;8(3):475-87.
46. Thunander M, Törn C, Petersson C, Ossiansson B, Fornander J, Landin-Olsson M. Levels of C-peptide, body mass index and age, and their usefulness in classification of diabetes in relation to autoimmunity, in adults with newly diagnosed diabetes in Kronoberg, Sweden. European journal of endocrinology. 2012;166(6):1021-9.
47. Hope SV, Knight BA, Shields BM, Hattersley AT, McDonald TJ, Jones AG. Random non-fasting C-peptide: bringing robust assessment of endogenous

insulin secretion to the clinic. Diabetic medicine : a journal of the British Diabetic Association. 2016;33(11):1554-8.
48. Hope SV, Knight BA, Shields BM, Hill AV, Choudhary P, Strain WD, et al. Random non-fasting C-peptide testing can identify patients with insulin-treated type 2 diabetes at high risk of hypoglycaemia. Diabetologia. 2018;61(1):66-74.
49. Donnelly LA, Zhou K, Doney ASF, Jennison C, Franks PW, Pearson ER. Rates of glycaemic deterioration in a real-world population with type 2 diabetes. Diabetologia. 2018;61(3):607-15.
50. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology.
51. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. The Lancet Diabetes & Endocrinology. 2019;7(6):442-51.
52. Tuomi T, Carlsson A, Li H, Isomaa B, Miettinen A, Nilsson A, et al. Clinical and genetic characteristics of type 2 diabetes with and without GAD antibodies. Diabetes. 1999;48(1):150.
53. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine. 2015;7(311):311ra174.
54. Udler MS. Type 2 Diabetes: Multiple Genes, Multiple Diseases. Current Diabetes Reports. 2019;19(8):55.
55. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CNA, et al. Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. Diabetes care. 2014;37(3):718.
56. Levy J, Atkinson AB, Bell PM, McCance DR, Hadden DR. Beta-cell deterioration determines the onset and rate of progression of secondary dietary failure in type 2 diabetes mellitus: the 10-year follow-up of the Belfast Diet Study. Diabetic medicine : a journal of the British Diabetic Association. 1998;15(4):290-6.
57. Matthews DR, Cull CA, Stratton IM, Holman RR, Turner RC. UKPDS 26: sulphonylurea failure in non-insulin-dependent diabetic patients over six years. Diabetic Medicine. 1998;15(4):297-303.
58. Ringborg A, Lindgren P, Yin DD, Martinell M, Stålhammar J. Time to insulin treatment and factors associated with insulin prescription in Swedish patients with type 2 diabetes. Diabetes & Metabolism. 2010;36(3):198-203.
59. Cook MN, Girman CJ, Stein PP, Alexander CM, Holman RR. Glycemic control continues to deteriorate after sulfonylureas are added to metformin among patients with type 2 diabetes. Diabetes care. 2005;28(5):995-1000.
60. Pani LN, Nathan DM, Grant RW. Clinical Predictors of Disease Progression and Medication Initiation in Untreated Patients With Type 2 Diabetes and A1C Less Than 7%. Diabetes care. 2008;31(3):386.
61. Waldman B, Jenkins AJ, Davis TME, Taskinen M-R, Scott R, O'Connell RL, et al. HDL-C and HDL-C/ApoA-I Predict Long-Term Progression of Glycemia in Established Type 2 Diabetes. Diabetes care. 2014;37(8):2351.
62. Pilla SJ, Yeh H-C, Juraschek SP, Clark JM, Maruthur NM. Predictors of Insulin Initiation in Patients with Type 2 Diabetes: An Analysis of the Look AHEAD Randomized Trial. Journal of general internal medicine. 2018;33(6):839-46.

63. Schrijnders D, Hartog LC, Kleefstra N, Groenier KH, Landman GWD, Bilo HJG. Within-Sulfonylurea-Class Evaluation of Time to Intensification with Insulin (ZODIAC-43). PloS one. 2016;11(6):e0157668-e.
64. Kostev K, Dippel F-W, Rathmann W. Predictors of insulin initiation in metformin and sulfonylurea users in primary care practices: the role of kidney function. Journal of diabetes science and technology. 2014;8(5):1023-8.
65. Gentile S, Strollo F, Viazzi F, Russo G, Piscitelli P, Ceriello A, et al. Five-Year Predictors of Insulin Initiation in People with Type 2 Diabetes under Real-Life Conditions. Journal of diabetes research. 2018;2018:7153087.
66. Bottazzo GF, Bosi E, Cull CA, Bonifacio E, Locatelli M, Zimmet P, et al. IA-2 antibody prevalence and risk assessment of early insulin requirement in subjects presenting with type 2 diabetes (UKPDS 71). Diabetologia. 2005;48(4):703-8.
67. Irvine WJ, Gray RS, McCallum CJ, Duncan LJP. CLINICAL AND PATHOGENIC SIGNIFICANCE OF PANCREATIC-ISLET-CELL ANTIBODIES IN DIABETICS TREATED WITH ORAL HYPOGLYCÆMIC AGENTS. The Lancet. 1977;309(8020):1025-7.
68. Pozzilli P, Di Mario U. Autoimmune Diabetes Not Requiring Insulin at Diagnosis (Latent Autoimmune Diabetes of the Adult). Diabetes care. 2001;24(8):1460.
69. Gale EAM. Latent autoimmune diabetes in adults: a guide for the perplexed. Diabetologia. 2005;48(11):2195-9.
70. Groop L, Tuomi T, Rowley M, Zimmet P, Mackay IR. Latent autoimmune diabetes in adults (LADA)—more than a name. Diabetologia. 2006;49(9):1996-8.
71. Buzzetti R, Zampetti S, Maddaloni E. Adult-onset autoimmune diabetes: current knowledge and implications for management. Nature Reviews Endocrinology. 2017;13:674.
72. Kobayashi T, Tamemoto K, Nakanishi K, Kato N, Okubo M, Kajio H, et al. Immunogenetic and Clinical Characterization of Slowly Progressive IDDM. Diabetes care. 1993;16(5):780.
73. Groop L, Miettinen A, Groop P-H, Meri S, Koskimies S, Bottazzo GF. Organ-Specific Autoimmunity and HLA-DR Antigens as Markers for β-Cell Destruction in Patients With Type II Diabetes. Diabetes. 1988;37(1):99-103.
74. Maioli M, Pes GM, Delitala G, Puddu L, Falorni A, Tolu F, et al. Number of autoantibodies and HLA genotype, more than high titers of glutamic acid decarboxylase autoantibodies, predict insulin dependence in latent autoimmune diabetes of adults. European journal of endocrinology. 2010;163(4):541-9.
75. Hornbak M, Allin KH, Jensen ML, Lau CJ, Witte D, Jørgensen ME, et al. A Combined Analysis of 48 Type 2 Diabetes Genetic Risk Variants Shows No Discriminative Value to Predict Time to First Prescription of a Glucose Lowering Drug in Danish Patients with Screen Detected Type 2 Diabetes. PLOS ONE. 2014;9(8):e104837.
76. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine. 2013;10(2):e1001381.
77. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2 ed. New York: Springer; 2009.
78. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. The BMJ. 2010;341:c6624.

79. Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower-Thomas K. The Nottingham Prognostic Index: five- and ten-year data for all-cause survival within a screened population. Ann R Coll Surg Engl. 2015;97(2):137-9.
80. Fox KA, Dabbous OH, Goldberg RJ, Pieper KS, Eagle KA, Van de Werf F, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). Bmj. 2006;333(7578):1091.
81. Apgar V. A proposal for a new method of evaluation of the newborn infant. Current researches in anesthesia & analgesia. 1953;32(4):260-7.
82. Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL, et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. Lancet. 2007;369(9558):283-92.
83. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical care medicine. 1985;13(10):818-29.
84. Maddrey WC, Boitnott JK, Bedine MS, Weber FL, Jr., Mezey E, White RI, Jr. Corticosteroid therapy of alcoholic hepatitis. Gastroenterology. 1978;75(2):193-9.
85. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax. 2003;58(5):377-82.
86. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81(24):1879-86.
87. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
88. Shields BM, McDonald TJ, Ellard S, Campbell MJ, Hyde C, Hattersley AT. The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. Diabetologia. 2012;55(5):1265-72.
89. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 2011;343.
90. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. Bmj. 2012;345:e5900.
91. Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. BMJ. 2017;359:j5019.
92. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabetic Medicine. 2010;27(8):887-95.
93. MDCalc. Diabetes Mellitus [Available from: https://www.mdcalc.com/.
94. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ. 1995;311(7019):1539.

95. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9(1):103.
96. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14(1):40.
97. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart. 2012;98(5):360-9.
98. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20.
99. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLOS Medicine. 2012;9(5):e1001221.
100. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology. 2015;68(1):25-34.
101. Strijker M, Chen JW, Mungroop TH, Jamieson NB, van Eijck CH, Steyerberg EW, et al. Systematic review of clinical prediction models for survival after surgery for resectable pancreatic cancer. The British journal of surgery. 2019;106(4):342-54.
102. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Medical Informatics and Decision Making. 2006;6(1):38.
103. Heus P, Damen J, Pajouheshnia R, Scholten R, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med. 2018;16(1):120.
104. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Bmj. 2015;351:h5527.
105. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Bmj. 2015;350:g7594.
106. Moons KG, Altman DG, Reitsma JB, Collins GS. New Guideline for the Reporting of Studies Developing, Validating, or Updating a Multivariable Clinical Prediction Model: The TRIPOD Statement. Advances in anatomic pathology. 2015;22(5):303-5.
107. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ open. 2019;9(4):e025611.
108. The EQUATOR Network. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies [Available from: http://www.equator-network.org/reporting-guidelines/stard/.
109. Cox DR. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society Series B (Methodological). 1958;20(2):215-42.
110. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54(1-2):167-79.

111. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.; 2001.
112. Menard SW. Applied logistic regression analysis. Thousand Oaks, CA: Sage Publications; 1995.
113. van Houwelingen JC, le Cessie S. Logistic Regression, a review. Statistica Neerlandica. 1988;42(4):215-32.
114. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical decision making : an international journal of the Society for Medical Decision Making. 2001;21(1):45-56.
115. Cox D. Regression models and life tables. Journal of the Royal Statistical Society, Series B (Methodological). 1972;34(2):187 - 220.
116. Royston P. Flexible Parametric Alternatives to the Cox Model, and more. The Stata Journal. 2001;1(1):1-28.
117. Royston P, Lambert P. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. USA: Stata Press; 2011.
118. Lambert PC, Royston P. Further Development of Flexible Parametric Models for Survival Analysis. The Stata Journal. 2009;9(2):265-90.
119. Ng R, Kornas K, Sutradhar R, Wodchis WP, Rosella LC. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. Diagnostic and Prognostic Research. 2018;2(1):4.
120. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555-61.
121. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine. 2007;26(11):2389-430.
122. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making. 2010;10:16-.
123. Atkinson EJ, Therneau TM, Melton LJ, 3rd, Camp JJ, Achenbach SJ, Amin S, et al. Assessing fracture risk using gradient boosting machine (GBM) models. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2012;27(6):1397-404.
124. Zhang J, Xu J, Hu X, Chen Q, Tu L, Huang J, et al. Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images. BioMed research international. 2017;2017:7961494-.
125. Emir B, Masters ET, Mardekian J, Clair A, Kuhn M, Silverman SL. Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. Journal of pain research. 2015;8:277-88.
126. Ban H-J, Heo JY, Oh K-S, Park K-J. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC genetics. 2010;11:26-.
127. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy. JAMA Network Open. 2018;1(3):e180926-e.
128. Kruppa J, Liu Y, Biau G, Kohler M, König IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. Biometrical Journal. 2014;56(4):534-63.
129. Boulesteix A-L, Schmid M. Machine learning versus statistical modeling. Biometrical Journal. 2014;56(4):588-93.

130. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
131. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. International journal of medical informatics. 2018;119:22-38.
132. van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW. Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Medical Research Methodology. 2011;11(1):143.
133. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses. PLOS ONE. 2014;9(6):e98587.
134. Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of Incident Diabetes in the Jackson Heart Study Using High-Dimensional Machine Learning. PloS one. 2016;11(10):e0163942-e.
135. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. JAMA Network Open. 2019;2(3):e190968-e.
136. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment JAMA Network Open. 2018;1(4):e181018-e.
137. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics. 2001;34(1):28-36.
138. Harrison RF, Kennedy RL. Artificial Neural Network Models for Prediction of Acute Coronary Syndromes Using Clinical Data From the Time of Presentation. Annals of Emergency Medicine. 2005;46(5):431-9.
139. Faisal M, Scally A, Howes R, Beatson K, Richardson D, Mohammed MA. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health informatics journal. 2018:1460458218813600.
140. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Statistics in medicine. 1998;17(21):2501-8.
141. Hsieh MH, Sun L-M, Lin C-L, Hsieh M-J, Hsu C-Y, Kao C-H. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer management and research. 2018;10:6317-24.
142. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiology. 2017;2(2):204-9.
143. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019.

144. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal. 2014;35(29):1925-31.
145. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ : British Medical Journal. 2013;346:e5595.
146. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLOS Medicine. 2013;10(2):e1001380.
147. Hingorani AD, Windt DAvd, Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ : British Medical Journal. 2013;346:e5793.
148. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. BJOG : an international journal of obstetrics and gynaecology. 2017;124(3):423-32.
149. Harrell F. Regression Modeling Strategies. 2 ed. Switzerland: Springer International Publishing; 2015. 582 p.
150. Carpenter JR, Kenward MG. Multiple Imputationand its Application: Wiley; 2012.
151. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63(2):205-14.
152. Little RJ, Rubin DB. Statistical Analysis with Missing Data (2nd Edition). New Jersey: Wiley; 2002.
153. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-91.
154. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
155. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research. 2018:0962280218784726.
156. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC medical research methodology. 2016;16(1):163-.
157. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biometrical journal Biometrische Zeitschrift. 2018;60(3):431-49.
158. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907-16.
159. Smits M, Dippel DW, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, et al. Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. Ann Intern Med. 2007;146(6):397-405.
160. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.

161. Royston P, Sauerbrie W. Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables: Wiley; 2008.
162. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in medicine. 2004;23(5):723-48.
163. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Medical Research Methodology. 2012;12(1):82.
164. Newson RB. Comparing the Predictive Powers of Survival Models Using Harrell's C or Somers' D. The Stata Journal. 2010;10(3):339-58.
165. Belsley DA, Kuh K, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley; 1980.
166. Hoaglin DC, Welsch RE. The Hat Matrix in Regression and ANOVA. The American Statistician. 1978;32(1):17-22.
167. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6.
168. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.
169. DiabetesGenes.org. Diabetes Alliance for Research in England (DARE) [Cited 15/11/2018]. Available from: https://www.diabetesgenes.org/current-research/dare/.
170. ClinicalTrials.gov. Predicting Response to Incretin Based Agents in Type 2 Diabetes (PRIBA) 2011 [Available from: https://clinicaltrials.gov/ct2/show/NCT01503112.
171. clinicaltrials.gov. PROMASTER - PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (PROMASTER) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02105792?term=promaster&rank=1.
172. ClinicalTrials.gov. RetroMASTER - Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes [Cited 15/11/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02109978.
173. ClinicalTrials.gov. MASTERMIND - Understanding Individual Variation in Treatment Response in Type 2 Diabetes (Mastermind) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT01847144?term=mastermind
174. clinicalTrials.gov. Genetics of Diabetes Audit and Research in Tayside Scotland (DOLORisk Dundee) (GoDARTS) [Available from: https://clinicaltrials.gov/ct2/show/NCT02783469.
175. Hebert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, et al. Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). International journal of epidemiology. 2017.
176. van der Heijden AA, Rauh SP, Dekker JM, Beulens JW, Elders P, t Hart LM, et al. The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. BMJ open. 2017;7(5):e015599.
177. Viberti G, Kahn SE, Greene DA, Herman WH, Zinman B, Holman RR, et al. A Diabetes Outcome Progression Trial (ADOPT). Diabetes care. 2002;25(10):1737.

178. Viberti G, Lachin J, Holman R, Zinman B, Haffner S, Kravitz B, et al. A Diabetes Outcome Progression Trial (ADOPT): baseline characteristics of Type 2 diabetic patients in North America and Europe. Diabetic medicine : a journal of the British Diabetic Association. 2006;23(12):1289-94.
179. Kahn SE, Haffner SM, Heise MA, Herman WH, Holman RR, Jones NP, et al. Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. N Engl J Med. 2006;355(23):2427-43.
180. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models 2018 [Available from: https://CRAN.R-project.org/package=gbm.
181. Meyer D, Dimitriadou E, Hornik J, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien 2018 [Available from: https://CRAN.R-project.org/package=e1071.
182. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.
183. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

Chapter 2.

Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18 to 50

Anita L Lynam, Timothy J McDonald, Anita V Hill, John M Dennis, Richard A Oram, Ewan R Pearson, Michael N Weedon, Andrew T Hattersley, Katharine R Owen, Beverley M Shields, Angus G Jones

BMJ Open 2019 10.1136/bmjopen-2019-031586

Acknowledgments of co-authors and contributions to paper

I was instrumental in formulation of the study concept, research methods and design. I performed the review of existing literature. I combined and cleaned the data. I designed and performed the statistical analysis and interpreted the results. I conceived the idea of and built the R-shiny app. I drafted the manuscript and responded to reviewer comments.

Beverley Shields and Angus Jones conceived the idea. Andrew Hattersley, Anita Hill, Angus Jones and Ewan Pearson researched the Exeter cohort data, Tim McDonald researched the pathology data, Mike Weedon and Richard Oram researched the genetic data and Katharine Owen researched the YDX data. Tim McDonald, John Dennis, Richard Oram, Andrew Hattersley and Katharine Owen discussed and contributed to study design. Beverley Shields and John Dennis provided support for the statistical analysis. All co-authors assisted me with the clinical interpretation of results. Beverley Shields and James Vine of Limited Edition Design assisted with implementing the app into the Diabetes Genes website. Angus Jones and Beverley Shields reviewed and gave feedback on the draft manuscripts and reviewer comments. All authors critically revised the manuscript and approved the final version.

Catherine Angwin, Anita Hill and Robert Bolt of the NIHR Exeter Clinical Research Facility (CRF) assisted me with data cleaning the Exeter cohorts. Steven Spaull of the NIHR Exeter CRF searched the DARE database and retrieved saved serum samples for additional antibody testing, Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital conducted the autoantibody analysis on these samples. Maarten van Smeden allowed me access to his Beyond EPV R-Shiny app (BETA version).

Abstract

Objective

To develop and validate multivariable clinical diagnostic models to assist distinguishing between type 1 and type 2 diabetes in adults aged 18 to 50.

Research design and methods

Multivariable logistic regression analysis was used to develop classification models integrating five pre-specified predictor variables, including clinical features (age of diagnosis, BMI) and clinical biomarkers (GADA and Islet Antigen 2 islet-autoantibodies, Type 1 Diabetes Genetic Risk Score), to identify type 1 diabetes with rapid insulin requirement using data from existing cohorts.

The study population consisted of 1,352 (model development) and 582 (external validation) participants diagnosed with diabetes between the age of 18 and 50 years of white European origin, recruited from primary and secondary care in the United Kingdom.

Type 1 diabetes was defined by rapid insulin requirement (within 3 years of diagnosis) and severe endogenous insulin deficiency (C-peptide <200pmol/L). Type 2 diabetes was defined by either a lack of rapid insulin requirement or, where insulin treated within 3 years, retained endogenous insulin secretion (C-peptide >600pmol/L at ≥5 years diabetes duration). Model performance was assessed using area under the receiver operating characteristic curve (ROC AUC), and internal and external validation.

Results

Type 1 diabetes was present in 13% of participants in the development cohort. All five predictor variables were discriminative and independent predictors of

type 1 diabetes ($p<0.001$ for all) with individual ROC AUC ranging from 0.82 to 0.85. Model performance was high: ROC AUC range 0.90 [95%CI 0.88, 0.93] (clinical features only) to 0.97 [0.96, 0.98] (all predictors) with low prediction error. Results were consistent in external validation (clinical features and GADA ROC AUC 0.93 [0.90, 0.96]).

Conclusions

Clinical diagnostic models integrating clinical features with biomarkers have high accuracy for identifying type 1 diabetes with rapid insulin requirement, and could assist clinicians and researchers in accurately identifying patients with type 1 diabetes.

Making the correct diagnosis of type 1 and type 2 diabetes is crucial for appropriate management, with guidelines for these conditions recommending very different glucose-lowering treatment and education (1-3). These differences are predominantly driven by the rapid development of severe endogenous insulin deficiency in type 1 diabetes (1). This means that patients with type 1 diabetes need rapid insulin treatment and are at risk of life-threatening ketoacidosis without insulin treatment. They develop a requirement for physiological insulin replacement (e.g. multiple injections, carbohydrate counting and pumps) due to the very high glycaemic variability associated with severe insulin deficiency (4, 5) and have poor glycaemic response to most adjuvant glucose-lowering therapies (6). In contrast, patients with type 2 diabetes continue to make substantial endogenous insulin even many decades after diagnosis (7). Glycaemia is therefore usually managed initially with lifestyle change or oral agents (4, 8) and, if insulin treatment is needed, a combination of simple insulin regimens and adjuvant non-insulin therapies (4, 5, 8, 9).

Correctly distinguishing between diabetes subtypes at diagnosis is often difficult and misclassification therefore common (10-12). Current guidelines focus on aetiopathological definitions without giving clear criteria for clinical use (1, 13). In clinical practice, clinical features are predominantly used to determine diabetes subtype but only age at diagnosis and BMI have evidence for utility at diabetes onset, whereas other features used by clinicians such as symptoms at diagnosis, weight loss or ketosis do not have an evidence base (14). Increasing obesity rates mean that many patients with type 1 diabetes will be obese and type 2 diabetes is occurring in the young (15). Type 1 diabetes has been recently shown to occur at similar rates in those aged above and below 30 (16). Therefore simple cut-offs based on age at diagnosis and BMI are unlikely to

accurately diagnose diabetes type for many patients (1, 10). Similarly, there is no single diagnostic test that can be used to classify diabetes robustly at diagnosis. While measurement of islet-autoantibodies can assist classification, many patients with type 1 diabetes are islet-autoantibody negative and many patients with the clinical phenotype of type 2 diabetes, without rapid insulin requirement, are islet-autoantibody-positive (17). A type 1 genetic risk score has been recently shown to assist diagnosis of diabetes type but this provides imperfect discrimination in isolation (18).

In order to classify diabetes a suitable "gold standard" is necessary. As the key factor driving differences in treatment decisions between the two subtypes is the lack of endogenous insulin secretion, direct measurement of endogenous insulin secretion in longstanding insulin-treated diabetes (>3-5 years), using C-peptide, provides a robust classification that closely relates to treatment requirements (19); patients with severe endogenous insulin deficiency (low C-peptide) have the high glucose variability, absolute insulin requirement, and lack of response to non-insulin glucose-lowering therapies that are characteristic of type 1 diabetes, regardless of their clinical characteristics and clinician's diagnosis (7, 11, 19-23). However, this test may have limited utility at diagnosis, as patients with recent onset type 1 diabetes may have retained endogenous insulin secretion (21, 24).

Clinical prediction models offer a way of combining multiple patient features and biomarkers to improve accuracy of diagnosis or prognosis. In diabetes, diagnostic models combining clinical features are available to predict the risk of prevalent or incident type 2 diabetes (25) and there is a model to identify monogenic forms of diabetes in patients with young-onset diabetes (26). However there are no statistical prediction models to help distinguish type 1 and

type 2 diabetes at diagnosis. We therefore aimed to develop and validate multivariable clinical diagnostic models that combine clinical features and biomarkers to identify type 1 diabetes (defined by rapid insulin requirement and severe endogenous insulin deficiency) in patients aged between 18 and 50 years at diabetes diagnosis.

Methods

We used logistic regression to model the relationship between each of clinical features and biomarkers, and type 1 diabetes defined by rapid insulin requirement and severe endogenous insulin deficiency (see below). We assessed the performance of the models using both internal validation and external validation.

Study population – development cohort

To maximise the sample size and to create a development cohort reflecting the general population prevalence of type 1 diabetes, participants were identified from four Exeter, UK-based cohorts (27-30) and combined into a single dataset. Combining the four Exeter cohorts was considered appropriate given that the assessment of both their clinical features and laboratory measurements were consistent across them.

These cohorts comprised of participants with clinically diagnosed diabetes recruited from primary and secondary care. Summaries of the cohorts including recruitment and data collection methods, and the number of type 1 diabetes in each cohort are shown in Supplementary Table 1.

Participants were eligible for the study (model development or validation) if they had a clinical diagnosis of type 1 or type 2 diabetes between the ages of 18 and 50 years. Participants with known secondary or monogenic diabetes (31), or a

known disorder of the exocrine pancreas (32), were excluded. All participants included in this study were of white European origin.

Study population - external validation cohort

Participants meeting the study inclusion criteria were identified in the Young Diabetes in Oxford (YDX) study (33). YDX is a cross-sectional study of participants diagnosed with diabetes (of any type) up to the age of 45 years, recruited from primary and secondary care in the Thames Valley region, UK. Participants with known secondary, pancreatic or monogenic diabetes were excluded.

Model outcome: type 1 and type 2 diabetes definition

Type of diabetes was defined by the presence or absence of rapid insulin requirement and severe endogenous insulin deficiency after a diagnosis of diabetes, as follows:

Type 1 diabetes: Insulin treatment within <= 3 years of diabetes diagnosis and severe insulin deficiency (non–fasting C-peptide < 200pmol/L) (21).

Type 2 diabetes: Either 1) no insulin requirement for 3 years from diabetes diagnosis or 2) where insulin was started within 3 years of diagnosis, substantial retained endogenous insulin secretion (C-peptide >600pmol/L) at >=5 years diabetes duration.

Cohort participants not meeting the above criteria or with insufficient information were excluded from analysis, as type of diabetes and rapid insulin requirement could not be robustly defined.

Model predictors

Five pre-specified predictor variables were assessed, based on prior evidence and availability: age at diagnosis (14), BMI (14), GADA and IA-2 islet-autoantibodies (17, 34), and a Type 1 diabetes Genetic Risk Score (T1D GRS) (18).

Assessment of clinical features

At study recruitment visit, clinical history including time to insulin and age at diagnosis were self-reported by participants in an interview with a research nurse. Height and weight were measured for calculation of BMI.

Laboratory Measurement

C-peptide

In the development cohort, C-peptide was measured on stored EDTA taken at study visits (non-fasting random (35), fasting, or at 90 minutes in a post-mixed-meal tolerance test (majority 87% non-fasting)). With specific additional consent, C-peptide was also measured on post-recruitment non-fasting EDTA samples collected as part of routine clinical care. Fasting C-peptide values were multiplied by 2.5 to non-fasting equivalent (21). The median C-peptide value was used where more than one eligible C-peptide value was available (62% of participants requiring this measure for outcome definition). C-peptide was measured using an electrochemiluminescence immunoassay on a Roche Diagnostics E170 analyser (Roche, Mannheim, Germany) by the Academic Department of Blood Sciences at the Royal Devon and Exeter Hospital. In the external validation cohort, C-peptide measurement was performed in the Biochemistry Laboratory of the Oxford University Hospitals NHS Trust using a

chemiluminescence immunoassay on an ADVIA Centaur analyser (Siemens Healthcare Diagnostics Ltd).

Islet-autoantibodies

In the development cohort, GADA and IA-2 were measured on EDTA taken at recruitment or obtained from local laboratory records. Both islet-autoantibodies were measured using the RSR Ltd ELISA assays (RSR Ltd, Cardiff, UK) on the Dynex DS2 ELISA Robot (Dynex Technologics, Worthing, UK) by the Academic Department of Blood Sciences at the Royal Devon and Exeter Hospital. The department participates in the International Autoantibody Standardization Programme. The cut-off for positivity for GADA was ≥11 units/ml and IA-2 was ≥15 units/ml, based on the 97.5th centile of 1,559 controls without diabetes (34).

In the external validation cohort, GADA was measured by a radioimmunoassay using 35S-labeled full-length GAD65 by the Department of Clinical Science, University of Bristol, Bristol, U.K. Results were expressed in World Health Organization (WHO) units per millilitre derived from a standard curve calibrated from international reference material (National Institute for Biological Standards and Control code 97/550). The cut-off for positivity for GADA was 13 WHO Units/mL initially, using a local assay (samples measured n=218, DASP2010 sensitivity 88% at 93% specificity) and changed to 33 DK Units/mL later in the study (standard assay, DASP2010 sensitivity 80%, specificity 97%).

Type 1 Diabetes Genetic Risk Score (T1D GRS)

The T1D GRS was calculated on the development cohort as previously described (18). In brief, T1D GRS consists of 30 common type 1 diabetes genetic variants (single nucleotide polymorphisms (SNPs)) from HLA and non-

HLA loci; each variant is weighted by its effect size on type 1 diabetes risk from previously published literature, with weights for DR3/DR4-DQ8 assigned based on imputed haplotypes (Supplementary Table 2). All SNPs had an INFO > 0.8. The combined score represents an individual's genetic susceptibility to type 1 diabetes. T1D GRS calculation was not performed if genotyping results were missing for either of the two alleles with the greatest weighting (DR3/DR4-DQ8 or HLA_DRB1_15) or if more than two of any other SNPs were missing. For ease of clinical interpretation the score is presented in this article as the score and centile position of the distribution in the Wellcome Trust Case Control Consortium type 1 diabetes population (36).

Missing data

Models were developed using complete case analysis. The percentage of participants in the development data meeting our inclusion criteria but excluded due to missing data was 10% (Supplementary Table 2). The missing data for the majority of these participants was related to the model outcome, 11 participants were excluded due to missing BMI. These missing data were never collected (not by design). The nature of the missing data (missing data mechanism) was not investigated for these data due to the low amount of missing data (BMI), the sample size was considered sufficient to give unbiased estimates using complete-case analysis and the missing outcome data is highly unlikely to depend on the values of the predicted variables.

Missing data for the remaining predictor variables (GADA, IA-2 and T1D GRS) were never collected (by design). These missing data were handled by use of a staged model development sequence which was considered a suitable method of analysis in this situation and makes best use of the available data. To assess

the appropriateness of this approach, we first looked at the missing data patterns to describe the missing data. 70% of the participants had complete data and only 4% had missing data for all three predictor variables. The missing data mechanism for these variables was investigated by regressing a binary missing variable on the other variables. If no variables predict whether a given variable is missing, then it is plausible that the data is missing completely at random (MCAR) and a complete cases analysis is appropriate. If the data is not MCAR, then the complete case may not be a random sample and may produce biased estimates. Postive IA-2 was a significant predictor of missing GADA and vice versa, GADA was also a significant predictor of missing T1D GRS. Although these results suggest that the data may not be MCAR, there is no reason to assume that the missing values are distributed significantly differently from the non-missing values i.e. the data appears to be missing at random and multiple imputation was not considered.

Statistical analysis

Model development

We used logistic regression analysis to develop the models.

Clustering of data by cohort origin was not adjusted for in the models since cohort origin was inherently associated with type of diabetes (Supplementary Table 1).

Age at diagnosis, BMI and T1D GRS were modelled as continuous variables and transformations used to ensure linearity on the logit scale (37) (Supplementary Figures 1A and 1B). GADA and IA-2 were both dichotomized into negative or positive based on the cut-off for positivity in line with how the results are reported clinically (2). Sample sizes were checked using both

minimal Events Per Variable (EPV) criteria (>=10) (38) and square root of the mean squared prediction error (rMPSE) (39) and were considered sufficient for reliable diagnostic modelling.

Models were built and validated in four stages, this staged development sequence was selected in order of clinical availability of the predictors and, as some participants had missing diagnostic test data, to maximise the sample size at each stage: 1) model including only clinical features (age at diagnosis and BMI); 2) Addition of GADA to the linear predictor from model 1; 3) Addition of both GADA and IA-2 to the linear predictor from model 1; 4) Addition of T1D GRS to model 3 linear predictor.

Evaluation of model performance: Internal validation

Three internal validation techniques were used to assess the discrimination and calibration performance of the models: 1) directly using the data used to develop the model (apparent validation, ROC AUC); 2) Jack-knife cross-validation; 3) Bootstrapping (with replacement method) (37).

Evaluation of model performance: External validation

Performances of model 1 (clinical features) and model 2 (clinical features + GADA), were evaluated in the YDX study cohort. We were unable to externally evaluate models 3 and 4 as IA-2 autoantibodies and T1D GRS were not available in the YDX study.

Model comparisons

Four nested replica models were built on the subset of participants with complete data on all predictor variables (n = 943). The predictive information of each additional predictor on the model performance was assessed using the

Unitless Index of Adequacy (37), log likelihood ratio test (37), Net Reclassification Improvement and Integrated Discrimination Improvement (40).

Sensitivity analysis

Model development of all 4 models was repeated on 943 participants with complete data. To assess performance of biomarker models in those difficult to classify on clinical features alone model AUC ROC was repeated for each model in participants with intermediate age of diagnosis (range 25-35 years (inclusive)) and BMI (range 25-35 kg/m2 (inclusive)).

All statistical analyses were performed using STATA version 15, STATA Corp, Texas, USA (unless otherwise stated).

Results

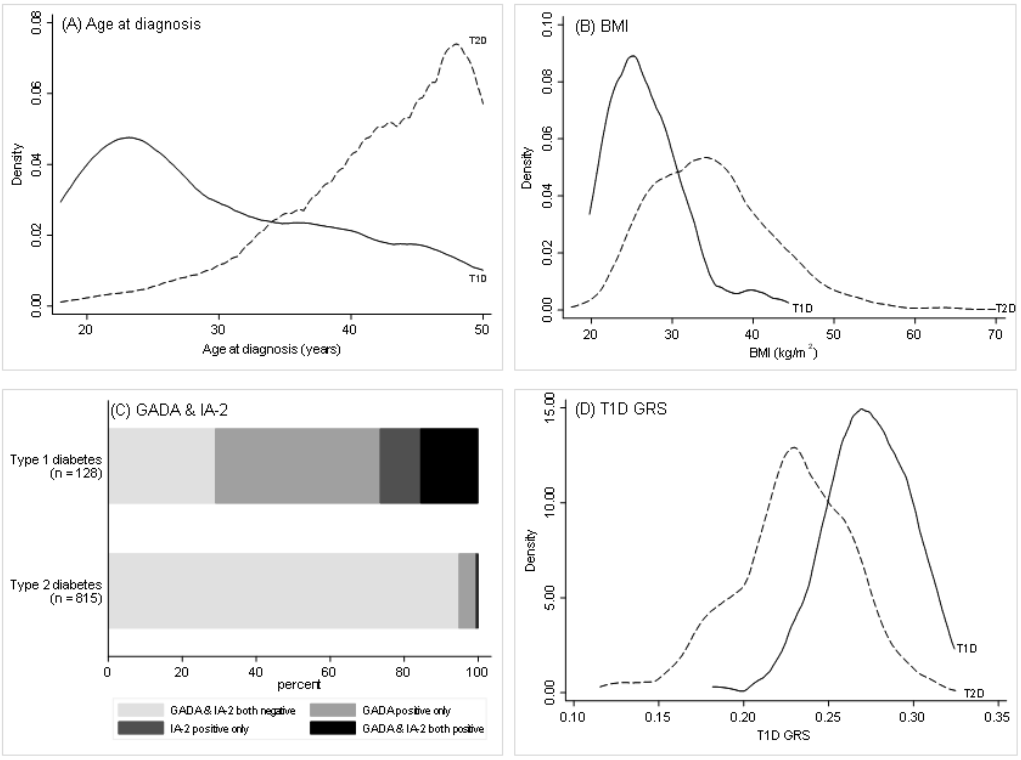
1,352 (type 1 diabetes n = 179) participants met analysis inclusion criteria for the clinical features model with 943 participants having all predictor variables measured. A flow diagram describing the flow of participants through the study is shown in Supplementary Figure 2. The majority of participants (n = 904 (67%)) were identified from DARE which is an unselected cohort enriched for type 1 diabetes due to some secondary care recruitment (type 1 diabetes prevalence 19.6%). The inclusion of 448 participants from the other three Exeter cohorts, which are type 2 diabetes focused, resulted in an overall cohort prevalence of type 1 diabetes very similar to that in published population cohorts (41) meaning that model probabilities are more likely to be relevant to the general population than those obtained using DARE alone.

Only 37 (2.7% of the cohort) had an undefinable outcome due to intermediate C-peptide levels (200-600pmol/L when insulin-treated within 3 years of

diagnosis). The remaining exclusions were due to either missing data or short duration of diabetes. The characteristics and type 1 diabetes outcome prevalence of the included participants were similar in all four development samples (Supplementary Table 3). There were no clinically relevant differences in the characteristics of the participants who were excluded from the fourth model development stage (n = 409) (Supplementary Table 4). Islet-autoantibodies and C-peptide were measured at median 13 years and 16 years post-diagnosis respectively.

Clinical features or biomarkers in isolation overlap substantially between diabetes types (Figure 1). Participants with type 1 diabetes and rapid insulin requirement were diagnosed younger compared to the participants with type 2 diabetes (median 27 vs 44 years, $p < 0.001$) and had a lower BMI (median 26 vs 34 kg/m^2, $p < 0.001$). Positive autoantibodies (GADA, IA-2 or both) were more common in the participants with type 1 diabetes (71% of participants with type 1 diabetes vs 5% of participants with type 2 diabetes, $p < 0.001$). Patients with type 1 diabetes had a higher T1D GRS (median 0.27 vs 0.23 (equivalent to 40th and 4th centile of the Wellcome Trust Case Control Consortium population with type 1 diabetes (36), $p < 0.001$). These features overlapped substantially between participants meeting criteria for type 1 and type 2 diabetes (Figure 1 (A – D)) with AUC ROC for these features in isolation: 0.82 (age at diagnosis), 0.83 (BMI), 0.83 (islet-autoantibodies) and 0.85 (T1D GRS).

Figure 1: Density plots for (A) age at diagnosis, (B) BMI and (D) T1D GRS. Stacked bar chart (C) showing percentages of participants (total n = 943 (stage 4 model development sample)) by actual type 1 diabetes outcome and GADA/IA-2 status. Dashed line shows the distribution for type 2 diabetes (T2D) (n = 815), solid line shows the distribution for type 1 diabetes (T1D) (n = 128) of participants included in the stage 4 model development.



Combining clinical features using a diagnostic model improves model discrimination

In model 1, age at diagnosis and BMI were both significant independent predictors of type 1 diabetes, with the odds of having type 1 diabetes increasing with younger age at diagnosis and lower BMI. Combined, these features provided excellent discrimination (ROC AUC=0.904, perfect test = 1) (Figure 2a), with low probabilities capturing the majority of participants with type 2 diabetes and type 1 diabetes being very unlikely (Figure 2b; sensitivity, specificity, and positive and negative predictive values at various probability cut-offs are reported in Table 1). In successive models adding in GADA (model 2 (figures 2c and 2d)), then IA-2 (model 3 (figures 2e and 2f)) and then T1D GRS (model 4 (figures 2g and 2h)), the addition of each predictor to the previous

model resulted in significant improvements in discrimination (Supplementary Table 5) and model fit (Supplementary Tables 6 and 7). In sensitivity analysis, results were similar when restricting all models to only the 943 participants with complete data on all predictor variables (Supplementary Table 8).

Figure 2: Development sample validation results. Plots are the results from the validation of the models. First row (a and b): clinical features logistic regression model (n = 1,315). Second row (c and d): clinical features + GADA logistic regression model (n = 1,036). Third row (e and f): clinical features + GADA + IA-2 logistic regression model (n = 1,025). Fourth row (g and h): clinical features + GADA + IA-2 + T1D GRS logistic regression model (n =943). Plots (a), (c), (e), & (g) are ROC curves showing discrimination ability of the models. Plots (b), (d), (f) & (h) are boxplots of fitted model probabilities grouped by actual diabetes outcome.

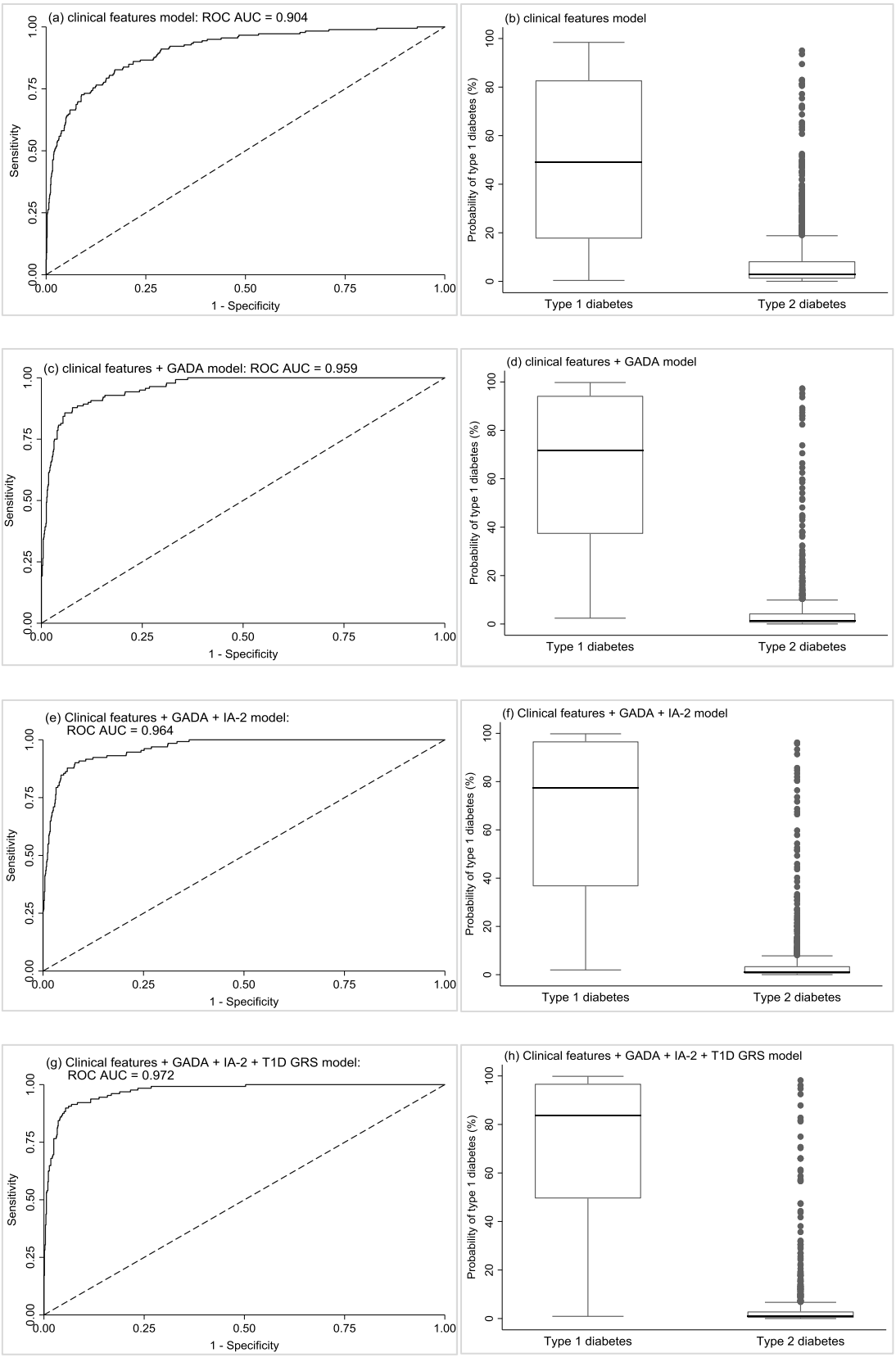


Table 1: Model performance at different cut-offs for classifying type 1 diabetes for all four logistic regression models (development cohort). Positive and negative predictive values assume prevalence for type 1 diabetes: Model 1 = 13%, Model 2 = 14%, Model 3 = 13%, Model 4 = 14%

* Youden's Index - best trade-off between sensitivity and specificity (sensitivity+specificity – 1).

Model 1: Clinical features (n = 1,352)						
	Probability (%) cut-off					
	10	30	50	70	90	12 *
Sensitivity/specificity (%)	85/79	64/95	49/98	35/99	15/100	83/83
Accuracy (%)	80	90	91	90	89	83
PPV (%)	38	64	79	83	90	42
NPV (%)	97	95	93	91	89	97
Model 2: Clinical features + GADA (n = 1,036)						
	Probability (%) cut-off					
	10	30	50	70	90	16 *
Sensitivity/specificity (%)	90/88	80/96	66/97	52/99	31/100	86/92
Accuracy (%)	89	94	93	92	90	92
PPV (%)	55	75	80	85	92	64
NPV (%)	98	97	95	93	90	98
Model 3: Clinical features + GADA + IA-2 (n = 1,025)						
	Probability (%) cut-off					
	10	30	50	70	90	12 *
Sensitivity/specificity (%)	91/91	80/96	69/98	57/99	37/100	90/92
Accuracy (%)	91	94	94	93	92	92
PPV (%)	59	75	81	85	92	62
NPV (%)	99	97	96	94	92	98
Model 4: Clinical features + GADA + IA-2 + T1D GRS (n = 943)						
	Probability (%) cut-off					
	10	30	50	70	90	14 *
Sensitivity/specificity (%)	92/90	84/96	74/98	63/99	41/100	91/93
Accuracy (%)	90	95	94	94	92	93
PPV (%)	59	78	83	88	93	67
NPV (%)	99	98	96	94	92	99

In further sensitivity analysis restricting analysis to those most difficult to classify on clinical features alone due to both intermediate BMI (range 25-35 kg/m2 (inclusive)) and age of diagnosis (range 25-35 years (inclusive)), model performance remained high for models incorporating biomarker measurement (clinical features + islet-autoantibodies AUC ROC 0.89, clinical features + islet-autoantibodies + T1D GRS AUC ROC 0.95) (Supplementary Table 9). This compares to AUC ROC of 0.72 for GADA and IA-2 measurement alone, and 0.89 for T1D GRS measurement alone in this sub population (n = 71).

Internal validation suggests robust model performance

Results of the internal validation bootstrap (Supplementary Table 5) indicate good model discrimination, with very similar model performance in bootstrapped samples (near identical ROC AUC for all models (max decrease = 0.0018)), high calibration indicating the predicted probabilities closely fit the observed probabilities (calibration slope range 0.98 - 1.00 (0.9 – 1.1 is indicative of good calibration)), and very low levels of optimism suggesting little error due to overfitting.

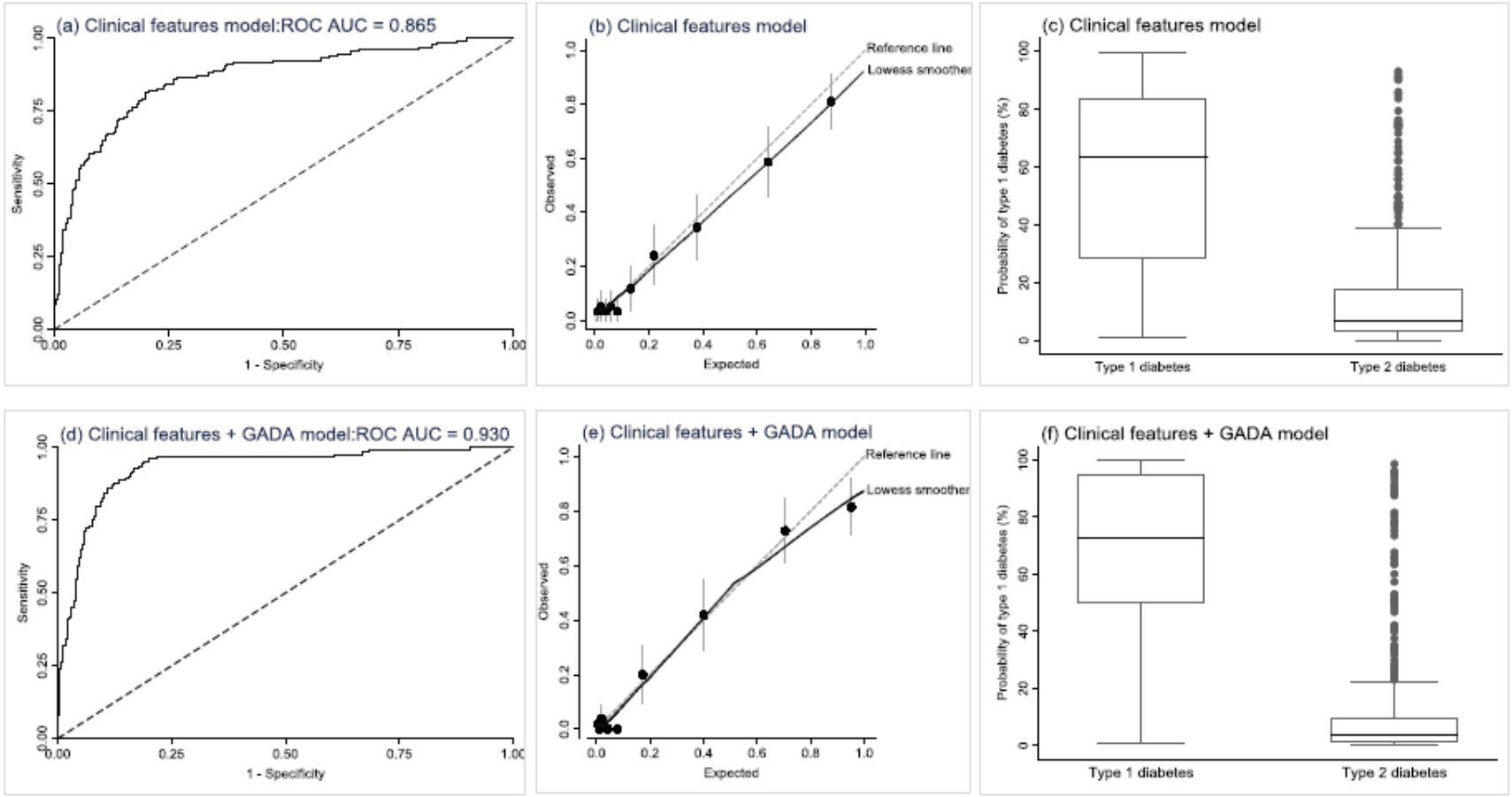
Model performance remains high in an external validation cohort with different characteristics

582 participants in the YDX study met criteria for external validation (Supplementary Figure 3). Compared to the participants in the Exeter model development cohort, the participants in the YDX study were younger at diagnosis (consistent with the narrower age range in YDX (18-45y) (median 37 years vs 43 years, $p < 0.001$)), had a lower BMI (median 31 kg/m^2 vs 33 kg/m^2, $p < 0.001$), had a higher percentage of GADA (20% versus 12%, $p < 0.001$) and

a higher prevalence of type 1 diabetes by study definition (22% vs 14%, $p < 0.001$) (see Supplementary Table 10 for participant characteristics).

There was a small decrease in performance of the model 1 (clinical features) and model 2 (clinical features and GADA) when they were applied to the external validation samples but both still showed high levels of discrimination despite differences in the two cohorts (ROC AUC = 0.865 and 0.930 for models 1 (Figures 3a, 3b and 3c) and 2 (Figures 3d, 3e and 3f) respectively (Supplementary Table 11). Both models slightly over estimated type 1 diabetes prevalence but there was no evidence of miscalibration (Figures 3b and 3e, Supplementary Table 11). Sensitivity and specificity in the validation cohort are shown in Supplementary Table 12.

Figure 3: External validation results. Plots on the first row (a, b, c) are the results from the external validation of the clinical features logistic regression model applied to participants in the YDX study (n = 582). The second row of plots (d, e, f) are the results from the external validation of the clinical features + GADA logistic regression model applied to participants in the YDX study (n = 549). Plots (a) & (d) are ROC curves showing discrimination ability of the models, dashed line represents the reference line. Plots (b) & (e) are calibration plots. Plots (c) & (f) are boxplots of fitted model probabilities grouped by actual diabetes outcome.



Participants with high model probability type 1 diabetes but type 2 diabetes outcome have the characteristics of type 1 diabetes but took > 3 years to commence insulin therapy.

Supplementary Table 13 shows the characteristics of 12 participants in the external validation cohort with >80% model type 1 diabetes probability, but an actual model outcome of type 2 diabetes. These participants had the clinical characteristics associated with type 1 diabetes with GADA positivity and low C-peptide in the majority of cases (median C-peptide 120 pmol/L). However the time to insulin was > 3 years in GADA positive cases, suggesting slow onset autoimmune diabetes. In contrast, the 6 participants who had a low model type 1 diabetes probability (< 16%) but an actual model outcome of type 1 diabetes (Supplementary Table 14) had features associated with type 2 diabetes.

Online calculator

The four models have been incorporated into an online calculator (beta version available at https://www.diabetesgenes.org/t1dt2d-prediction-model/). An additional four models with different combinations of the five predictor variables were also developed for the online calculator, to allow every combination of clinical features plus the other biomarkers as optional. As expected, ROC AUC and prediction error results for these four additional models were intermediate between the basic clinical features model and the full model with all features (Supplementary Table 15).

Supplementary Tables 16 - 23 inclusive show the β coefficients and odds ratios for all models. The regression equations for the online calculator are shown in Supplementary Table 24.

Conclusions

We have developed, evaluated and validated clinical diagnostic models combining age at diagnosis, BMI, GADA, IA-2, and T1D GRS to provide estimates of a patient's risk of having type 1 diabetes requiring rapid insulin therapy from diagnosis. These models show high performance, and could potentially assist classification of diabetes in clinical practice and provide a tool for evidence based classification in research cohorts.

Model performance was optimised in the model combining all five predictors (ROC AUC 0.97). However, all models performed well with ROC AUC > 0.9 and low cross-validated prediction errors in development. The results of the external validation provide additional confidence in model performance. This was undertaken in a distinct dataset with different type 1 diabetes prevalence and biochemical assays.

This is the first study developing clinical diagnostic models for classification of type 1 and 2 diabetes. Key strengths of this study include our systematic approach to model development including robust internal and external validation (42). Our staged approach to model development means that we have maximised the information gained from each predictor. Our model is parsimonious, we have used only five predictors previously shown to be associated with type 1 diabetes. This, in combination with large datasets, mean we have a high number of events per variable and very low risk of overfitting, a common problem with diagnostic models of this nature. Our use of predominantly population-based cohorts recruited largely from a primary care setting (for model development) means our results are likely to reflect true associations in patients seen in clinical practice. The overall prevalence of study

defined type 1 diabetes of 13% in our development dataset is close to the 11% reported type 1 diabetes prevalence at diagnosis in a UK population aged 20-50 (41).

A limitation of our study is the cross-sectional nature of our cohorts meaning that age at diagnosis and time to insulin were self-reported at a single visit. Insulin commencement was also based on clinical decision-making rather than a trial protocol. BMI and antibodies were measured at median 13 years after diagnosis. BMI, and GADA and IA-2 antibodies change modestly over time in adult onset diabetes, with previous research suggesting an approximately 18% lower combined GADA and IA-2 prevalence after 13.5 years diabetes duration in this age group (43), and BMI having higher discrimination for diabetes classification when measured at diagnosis (44). The potential impact on the results of BMI and islet-autoantibodies having been measured some years post diagnosis is that the predictions may be under-estimated. The lack of information at diagnosis also meant we were unable to assess whether other features available at diagnosis may assist classification, such as presentation glycaemia, ketosis, or weight loss. A prospective study to validate these models, and assess whether other features may assist classification is therefore ongoing (https://clinicaltrials.gov/ct2/show/NCT03737799).

A further limitation is that this model has been developed and tested in a white European population with young onset diabetes, extension of this work to non-white populations and older age groups is therefore a priority for future research.

These models have the potential to help robustly classify diabetes in research cohorts, and may have particular utility where genetic but not antibody data is

available, a common situation in many biobanks. They may also assist clinical decision making, with the important caveats that this evidence can only be applied to patients aged 18-50, of white ethnicity, and that these models are intended to act as a decision aid in conjunction with other information which a clinician may use to inform treatment decisions (for example severity of hyperglycaemia): they do not replace expert clinical opinion. A web-based calculator and smartphone app could be used to display the estimate of the patient's probability of having type 1 diabetes based on the predictor variable values entered. The models can be used with age of diagnosis and BMI as a minimum; users will then have a choice to add results of GADA, IA-2 and T1D GRS in any combination. This could therefore be used by clinicians as a triage-based approach to diabetes subtype diagnosis. For example, probabilities calculated on clinical features could be used as the basis for antibody testing, or the additional value likely to be gained from antibody or genetic testing could be assessed by inputting dummy results into the model. We propose providing the continuous probability outcome of the models rather than giving a threshold. This is because the decision made on whether to commence insulin for a given probability of type 1 diabetes will vary enormously due to other factors. For example temporary insulin treatment may be appropriate regardless of likely classification where hyperglycaemia is severe, and in some circumstances it may be appropriate to trial oral therapy even where type 1 diabetes has a high probability, for example where a person's occupation would be affected by insulin treatment and they can be carefully monitored for glycaemic deterioration.

In conclusion clinical diagnostic models integrating clinical features with biomarkers have high accuracy for identifying type 1 diabetes with rapid insulin

requirement in white participants aged 18 to 50 at diabetes diagnosis, and may assist clinicians in identifying patients with type 1 diabetes in clinical practice.

Acknowledgments

The authors thank participants who took part in these studies and the research teams who undertook cohort recruitment. We thank Catherine Angwin of the NIHR Exeter Clinical Research Facility for assistance with data preparation, and Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital for assistance with sample analysis. We are grateful to Maarten van Smeden for allowing us to access to his Beyond EPV R-Shiny app (BETA version).

References

1. American Diabetes Association (2018) 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. Diabetes care 41: S13 - S27
2. National Institute for Health and Care Excellence (2015) Type 1 diabetes in adults: diagnosis and management (NICE guideline NG17). Available from https://www.nice.org.uk/guidance/ng17, accessed 14/08/2018
3. National Institute for Health and Care Excellence (2015) Type 2 diabetes in adults: management (NICE guideline NG28). Available from https://www.nice.org.uk/guidance/ng28, accessed 14/08/2018
4. American Diabetes Association (2018) 8. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2018. Diabetes care 41: S73 - S85
5. DeWitt DE, Hirsch IB (2003) Outpatient insulin therapy in type 1 and type 2 diabetes mellitus: Scientific review. JAMA 289: 2254-2264
6. Frandsen CS, Dejgaard TF, Madsbad S (2016) Non-insulin drugs to treat hyperglycaemia in type 1 diabetes mellitus. The Lancet Diabetes & Endocrinology 4: 766-780
7. (1998) Effect of intensive therapy on residual β-cell function in patients with type 1 diabetes in the diabetes control and complications trial: A randomized, controlled trial. Annals of Internal Medicine 128: 517-523
8. Inzucchi SE, Bergenstal RM, Buse JB, et al. (2012) Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia 55: 1577-1596
9. Riddle MC (2008) Combined Therapy With Insulin Plus Oral Agents: Is There Any Advantage? Diabetes care 31: S125
10. Farmer A, Fox R (2011) Diagnosis, classification, and treatment of diabetes. BMJ 342
11. Hope SV, Knight BA, Shields BM, et al. (2018) Random non-fasting C-peptide testing can identify patients with insulin-treated type 2 diabetes at high risk of hypoglycaemia. Diabetologia 61: 66-74
12. Stone MA, Camosso-Stefinovic J, Wilkinson J, De Lusignan S, Hattersley AT, Khunti K (2010) Incorrect and incomplete coding and classification of diabetes: a systematic review. Diabetic Medicine 27: 491-497
13. World Health Organization (2006) Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. In:
14. Shields BM, Peters JL, Cooper C, et al. (2015) Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature. BMJ open 5
15. Rosenbloom AL, Joe JR, Young RS, Winter WE (1999) Emerging epidemic of type 2 diabetes in youth. Diabetes care 22: 345
16. Thomas NJ, Jones SE, Weedon MN, Shields BM, Oram RA, Hattersley AT (2018) Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. The Lancet Diabetes & Endocrinology 6: 122-129
17. Niskanen LK, Tuomi T, Karjalainen J, Groop LC, Uusitupa MIJ (1995) GAD Antibodies in NIDDM: Ten-year follow-up from the diagnosis. Diabetes care 18: 1557

18. Oram RA, Patel K, Hill A, et al. (2016) A Type 1 diabetes genetic risk score can aid discrimination between Type 1 and Type 2 diabetes in young adults. Diabetes care 39: 337-344
19. Jones AG, Besser REJ, Shields BM, et al. (2012) Assessment of endogenous insulin secretion in insulin treated diabetes predicts postprandial glucose and treatment response to prandial insulin. BMC Endocrine Disorders 12: 6
20. Steffes MW, Sibley S, Jackson M, Thomas W (2003) Beta-cell function and the development of diabetes-related complications in the diabetes control and complications trial. Diabetes care 26: 832-836
21. Jones AG, Hattersley AT (2013) The clinical utility of C-peptide measurement in the care of patients with diabetes. Diabetic Medicine 30: 803-817
22. Jones AG, McDonald TJ, Shields BM, et al. (2016) Markers of beta cell failure predict poor glycemic response to GLP-1 receptor agonist therapy in type 2 diabetes. Diabetes care 39: 250-257
23. Chow LS, Chen H, Miller ME, Marcovina SM, Seaquist ER (2015) Biomarkers related to severe hypoglycaemia and lack of good glycaemic control in ACCORD. Diabetologia 58: 1160-1166
24. Thunander M, Törn C, Petersson C, Ossiansson B, Fornander J, Landin-Olsson M (2012) Levels of C-peptide, body mass index and age, and their usefulness in classification of diabetes in relation to autoimmunity, in adults with newly diagnosed diabetes in Kronoberg, Sweden. European journal of endocrinology 166: 1021-1029
25. Collins GS, Mallett S, Omar O, Yu L-M (2011) Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine 9: 103-103
26. Shields BM, McDonald TJ, Ellard S, Campbell MJ, Hyde C, Hattersley AT (2012) The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. Diabetologia 55: 1265-1272
27. DiabetesGenes.org Diabetes Alliance for Research in England (DARE) Available from https://www.diabetesgenes.org/current-research/dare/, accessed 23 November 2017
28. ClinicalTrials.gov RetroMASTER - Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes. Available from https://www.clinicaltrials.gov/ct2/show/NCT02109978
29. ClinicalTrials.gov MASTERMIND - Understanding Individual Variation in Treatment Response in Type 2 Diabetes (Mastermind). Available from https://www.clinicaltrials.gov/ct2/show/NCT01847144?term=mastermind accessed 31 July 2018
30. clinicaltrials.gov PROMASTER - PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (PROMASTER). Available from https://www.clinicaltrials.gov/ct2/show/NCT02105792?term=promaster&rank=1, accessed 31 July 2018
31. Shields BM, Shepherd M, Hudson M, et al. (2017) Population-Based Assessment of a Biomarker-Based Screening Pathway to Aid Diagnosis of Monogenic Diabetes in Young-Onset Patients. Diabetes care 40: 1017-1025
32. Woodmansey C, McGovern AP, McCullough KA, et al. (2017) Incidence, Demographics, and Clinical Characteristics of Diabetes of the Exocrine Pancreas (Type 3c): A Retrospective Cohort Study. Diabetes care

33. Thanabalasingham G, Pal A, Selwood MP, et al. (2012) Systematic Assessment of Etiology in Adults With a Clinical Diagnosis of Young-Onset Type 2 Diabetes Is a Successful Strategy for Identifying Maturity-Onset Diabetes of the Young. Diabetes care 35: 1206-1212
34. McDonald TJ, Colclough K, Brown R, et al. (2011) Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabetic medicine : a journal of the British Diabetic Association 28: 1028-1033
35. Hope SV, Knight BA, Shields BM, Hattersley AT, McDonald TJ, Jones AG (2016) Random non-fasting C-peptide: bringing robust assessment of endogenous insulin secretion to the clinic. Diabetic medicine : a journal of the British Diabetic Association 33: 1554-1558
36. The Wellcome Trust Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678
37. Harrell F (2015) Regression Modeling Strategies. Springer International Publishing, Switzerland
38. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49: 1373-1379
39. van Smeden M, Moons KGM, de Groot JAH, et al. (2018) Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research: 0962280218784726
40. Steyerberg EW, Vickers AJ, Cook NR, et al. (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass) 21: 128-138
41. Scottish Diabetes Survey 2017 http://www.diabetesinscotland.org.uk/Publications/SDS%202017.pdf . Accessed 03/05/2019
42. Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal 35: 1925-1931
43. Tridgell DM, Spiekerman C, Wang RS, Greenbaum CJ (2011) Interaction of Onset and Duration of Diabetes on the Percent of GAD and IA-2 Antibody–Positive Subjects in the Type 1 Diabetes Genetics Consortium Database. Diabetes care 34: 988
44. Hope SV, Wienand-Barnett S, Shepherd M, et al. (2016) Practical Classification Guidelines for Diabetes in patients treated with insulin: a cross-sectional study of the accuracy of diabetes diagnosis. The British journal of general practice : the journal of the Royal College of General Practitioners 66: e315-322

Supplementary material

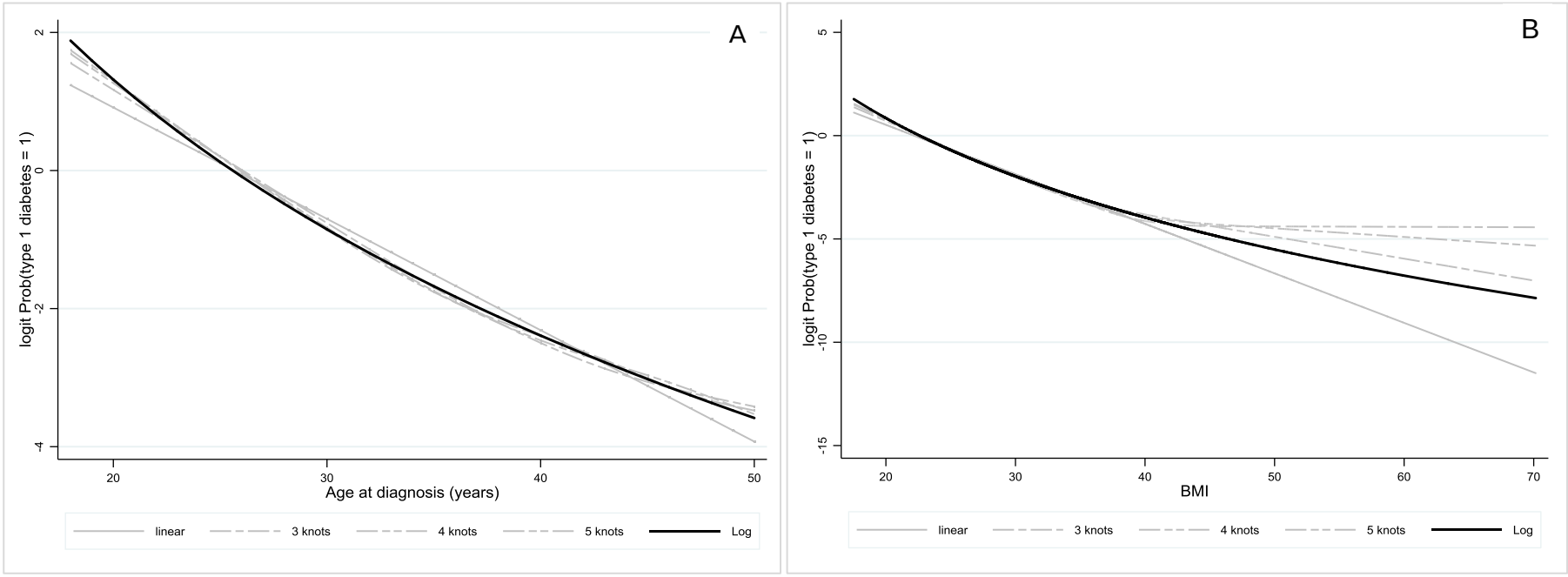
Supplementary Table 1: Cohort recruitment and data collection methods summary. *Included in the clinical features model stage 1 development.

	DARE	PRIBA	MRC Pro/RetroMaster	MRC crossover
Included participants*	904	368	72	8
Type 1/Type 2 diabetes n (%)	177 (19.6%)/727 (80.4%)	2 (0.5%)/366 (99.5%)	0 (0.0%)/72 (100%)	0 (0.0%)/8 (100%)
Data collection period	2007 to 2017	2011 to 2013	2013 to 2015	2013 to 2015
Study design	Cross-sectional	Longitudinal	Cross-sectional	Interventional Crossover
Setting	Primary and secondary care in eight diabetes research regions, England and retinal screening clinics.	Primary and secondary care in South West England	Primary and secondary care sites South West England, Tayside, Oxford, Glasgow, KCL and Newcastle, U.K.	Exeter and Tayside,U.K.
Inclusion criteria	Clinical diagnosis of diabetes (any type).	Clinical diagnosis of type 2 diabetes. Clinician determined requirement for DPP-IV inhibitor or GLP-1 analogue (HbA_{1c} >7.5%)	Clinical diagnosis of type 2 diabetes non-insulin treated within 6 months of diagnosis. Participants were selected on the basis of rapid or slow progression to insulin therapy (<7, >7 years). Age 18-90 inclusive.	Clinical diagnosis of type 2 diabetes, currently treated with sulphonylurea tablets and no change in treatment in previous 3 months, Last HbA_{1c} (within previous 12 months) ≥42 and ≤75 mmol/mol (6-9%). Age 19-79 inclusive.
Data collection	Clinical measurements and blood sample collected at visit. Ongoing biochemical data collected from pathology laboratories.	Clinical measurements and blood taken at initial visit. Follow up clinical measurements and blood collected at three and six months.	Clinical measures and fasting blood sample taken at visit.	MMT at baseline & MMT on each study drug visits. Three fasting blood collected at crossovers.

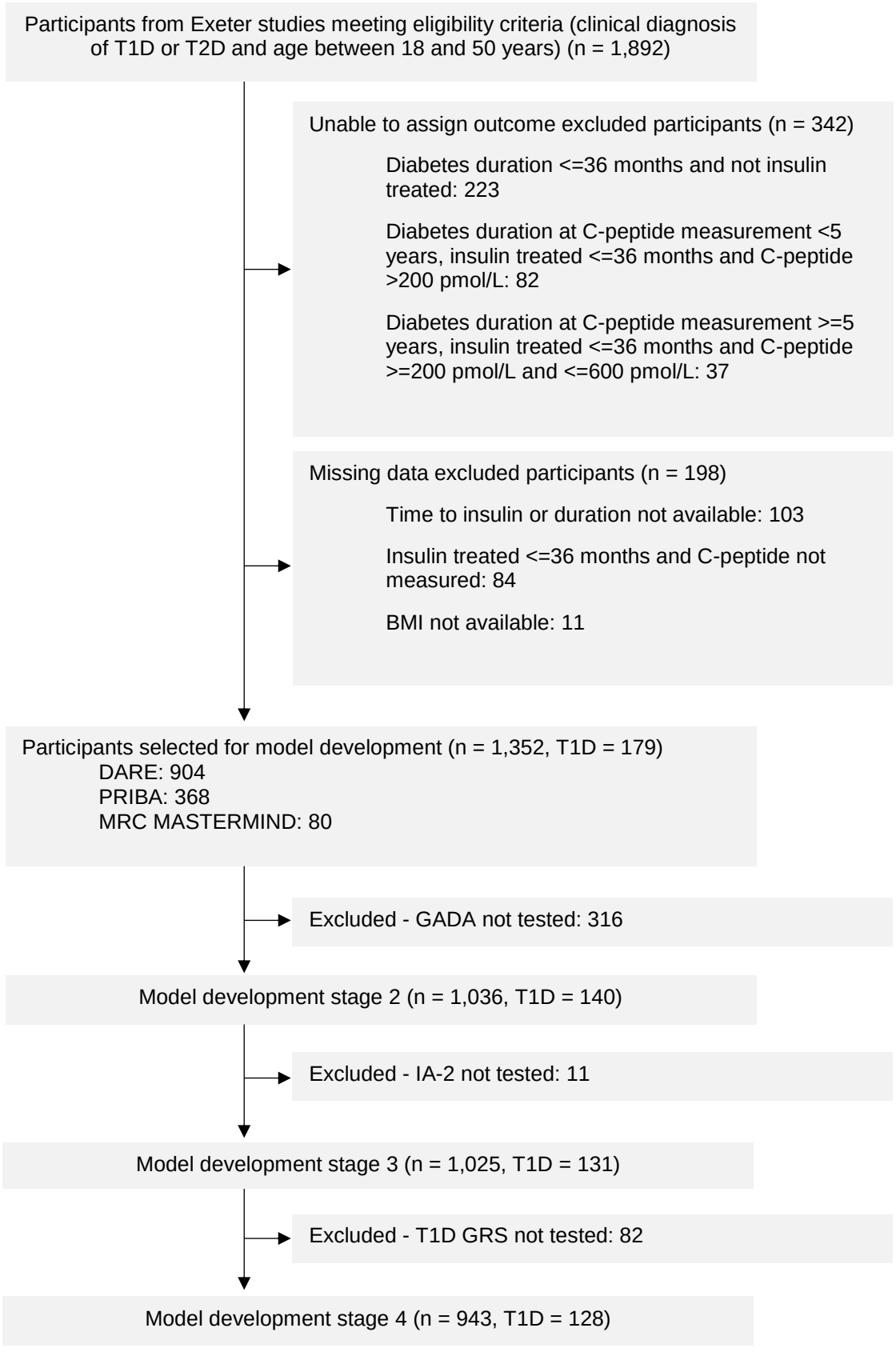
Supplementary Table 2: Type 1 diabetes SNPs included in the genetic risk score with weights. Effect allele is the risk increasing allele on the positive strand.

SNP	Gene	Odds Ratio	Weight	Effect Allele
rs2187668, rs7454108	DR3/DR4	48.18	3.87	
	DR3/DR3	21.12	3.05	
	DR4/DR4	21.98	3.09	
	DR4/X	7.03	1.95	
	DR3/X	4.53	1.51	
rs1264813	HLA_A_24	1.54	0.43	T
rs2395029	HLA_B_5701	2.50	0.92	T
rs3129889	HLA_DRB1_15	14.88	2.70	A
rs2476601	PTPN22	1.96	0.67	A
rs689	INS	1.75	0.56	T
rs12722495	IL2RA	1.58	0.46	T
rs2292239	ERBB3	1.35	0.30	T
rs10509540	C10orf59	1.33	0.29	T
rs4948088	COBL	1.30	0.26	C
rs7202877		1.28	0.25	G
rs12708716	CLEC16A	1.23	0.21	A
rs3087243	CTLA4	1.22	0.20	G
rs1893217	PTPN2	1.20	0.18	G
rs11594656	IL2RA	1.19	0.17	T
rs3024505	IL10	1.19	0.17	G
rs9388489	C6orf173	1.17	0.16	G
rs1465788		1.16	0.15	C
rs1990760	IFIH1	1.16	0.15	T
rs3825932	CTSH	1.16	0.15	C
rs425105		1.16	0.15	T
rs763361	CD226	1.16	0.15	T
rs4788084	IL27	1.16	0.15	C
rs17574546		1.14	0.13	C
rs11755527	BACH2	1.13	0.12	G
rs3788013	UBASH3A	1.13	0.12	A
rs2069762	IL2	1.12	0.11	A
rs2281808		1.11	0.10	C
rs5753037		1.10	0.10	T

Supplementary Figure 1: Relationship between age at diagnosis (A) and BMI (B) and response modelled using restricted cubic splines (k = 3, 4 and 5) and a simple log transformation. Age at diagnosis and BMI did not predict linearly, the graphs of fitted splines and log transformation suggested that a simple log transformation was sufficient to induce linearity in both variables.



Supplementary Figure 2: Flow diagram of participants through the model development stages. T1D: type 1 diabetes, T2D: type 2 diabetes



Supplementary Table 3: Characteristics of the Exeter, U.K. study participants included at each model development stage. Model 1 – Clinical features (Age at diagnosis & BMI), Model 2 – Clinical features + GADA, Model 3 - Clinical features + GADA + IA-2, Model 4 - Clinical features + GADA + IA-2 + T1D GRS. Median (IQR) or % or *Geometric mean [95% CI] for transformed variables. †Measured at recruitment (median 13 years post diagnosis). Minimum and maximum values for each continuous predictor variable used in the models

	Model 1 development	Model 2 development n = 1,036	Model 3 development n = 1,025	Model 4 development n = 943
Characteristic				
Sex (% Male)	59%	59%	59%	59%
Age at diagnosis (years)*	40 [39, 41]	40 [39, 40]	40 [39, 40]	40 [39, 40]
Age at diagnosis (years) min, max	18, 50	18, 50	18, 50	18, 50
BMI (kg/m^2)*†	33 [32, 33]	33 [32, 33]	33 [32, 33]	33 [32, 33]
BMI (kg/m^2)*† min, max	17.5, 70.2	17.5, 70.2	17.5, 70.2	17.5, 70.2
Duration of diabetes (years)	13 (8, 20)	13 (8, 20)	13 (8, 20)	13 (8, 20)
Type 1 diabetes	13%	14%	13%	14%
HbA_{1c} (%)†	8.2 (7.1, 9.6)	8.3 (7.3, 9.8)	8.3 (7.3, 9.8)	8.2 (7.2, 9.7)
HbA_{1c} (mmol/mol)†	66 (54, 81)	67 (56, 84)	67 (56, 84)	66 (55, 83)
GADA positive (%)	-	12%	12%	12%
IA-2 positive (%)	-	-	4%	4%
T1D GRS	-	-	-	0.24 (0.22, 0.26)
T1D GRS centile	-	-	-	5.8 (1.2, 23.7)
T1D GRS min, max	-	-	-	0.12, 0.32

Supplementary Table 4: Comparison of characteristics for participants included in the model 4 development and participants included in model 1 development but excluded from model 4. Median (IQR) or % or *Geometric mean [95% CI] for transformed variables. †Measured at recruitment (median 13 years post diagnosis).

	Model 4 development n = 943	Model 4 development exclusions n = 409	p value for comparison
Characteristic			
Sex (% Male)	59%	60%	>0.1
Age at diagnosis (years)*	40 [39, 40]	41 [40, 42]	0.04
BMI (kg/m^2)*†	33 [32, 33]	33 [32, 33]	> 0.1
Duration of diabetes (years)	13 (8, 20)	13 (7, 20)	> 0.1
Type 1 diabetes	14%	12%	> 0.1
HbA_{1c} (%)†	8.2 (7.2, 9.7)	8.0 (6.9, 9.3)	0.009
HbA_{1c} (mmol/mol)†	66 (55, 83)	64 (52, 78)	0.009

Supplementary Table 5: Model performance results for the internal validation performed at each development stage. * P value for Brier score is Spiegelhalter's z-test used to evaluate the calibration component of the Brier score, significant p-values indicate poor calibration. †Result reported as raw cross-validation estimate of prediction error with misclassification cost function (cut-off 0.5). cv.glm function in R version 3.3.3.

Performance parameter	Development sample validation	Internal validation (bootstrap 500)		Optimism
		Apparent (SD)	test (SD)	
Clinical features model (n = 1,352)				
ROC [95% CI]	0.90 [0.88, 0.93]	0.9056 (0.013)	0.9038 (0.0005)	0.0018
Calibration-in-the-large	0	0.0000 (0.000)	0.0003 (0.1072)	-0.0003
Calibration slope (b_L)	1	1.0000 (0.000)	0.9977 (0.0678)	0.0023
Brier Score	0.07 (p = 0.50)	-	-	-
Hosmer-Lemeshow	p = 0.95	-	-	-
Jack-knife cross validation†	0.09	-	-	-
Clinical features + GADA model (n = 1,036)				
ROC [95% CI]	0.96 [0.95, 0.97]	0.9595 (0.0070)	0.9586 (0.0010)	0.0009
Calibration-in-the-large	0	0.0000 (0.0000)	-0.0019 (0.1472)	0.0019
Calibration slope (b_L)	1	1.0000 (0.0000)	0.9850 (0.0787)	0.015
Brier Score	0.05 (p = 0.35)	-	-	-
Hosmer-Lemeshow	p = 0.39	-	-	-
Jack-knife cross validation†	0.07	-	-	-
Clinical features + GADA + IA-2 model (n = 1,025)				
ROC [95% CI]	0.96 [0.95, 0.98]	0.9622 (0.007)	0.9633 (0.0015)	0.0011
Calibration-in-the-large	0	0.0000 (0.000)	0.0055 (0.1567)	-0.0055
Calibration slope (b_L)	1	1.0000 (0.000)	0.9780 (0.0707)	0.022
Brier Score	0.04 (p = 0.31)	-	-	-
Hosmer-Lemeshow	p = 0.14	-	-	-
Jack-knife cross validation †	0.06	-	-	-
Clinical features + GADA + IA-2 + T1D GRS model (n = 943)				
ROC [95% CI]	0.97 [0.96, 0.98]	0.9718 (0.0060)	0.9710 (0.0006)	0.0008
Calibration-in-the-large	0	0.0000 (0.0000)	0.0084 (0.1675)	-0.0084
Calibration slope (b_L)	1	1.0000 (0.0000)	0.9880 (0.0810)	0.0124
Brier Score	0.04 (p = 0.35)	-	-	-
Hosmer-Lemeshow	p = 0.84	-	-	-
Jack-knife cross validation †	0.06	-	-	-

Supplementary Table 6: Unitless index of adequacy is the proportion of log likelihood explained by each model stage with reference to the end model containing all predictors. Based on replica models developed using stage 4 development sample (n = 943).

Model	LR χ^2	Adequacy
Clinical features	324.7 (df 2)	0.67
Clinical features + GADA	418.7 (df 3)	0.87
Clinical features + GADA + IA-2	447.6 (df 5)	0.93
Clinical features + GADA + IA-2 + T1D GRS	481.8 (df 6)	1.00

Supplementary Table 7: Model fit comparisons of nested models developed using stage 4 development sample (n = 943). Null hypothesis for Likelihood Ratio test: Additional predictor(s) has no predictive information. Net Reclassification Improvement (NRI) calculated using 50% classification cut-off. IDI = Integrated Discrimination Improvement

Model comparison	Likelihood Ratio test	NRI	IDI
Adding GADA to Clinical features model	LR χ^2 (1) = 94.02 p <0.001	0.12, p = 0.01	0.13, p < 0.001
Adding IA-2 to Clinical features + GADA model	LR χ^2 (2) = 28.82 p < 0.001	0.14, p = 0.004	0.15, p < 0.001
Adding T1D GRS to Clinical features + GADA + IA-2 model	LR χ^2 (3) = 34.20 p < 0.001	0.06, p = 0.04	0.06, p < 0.001

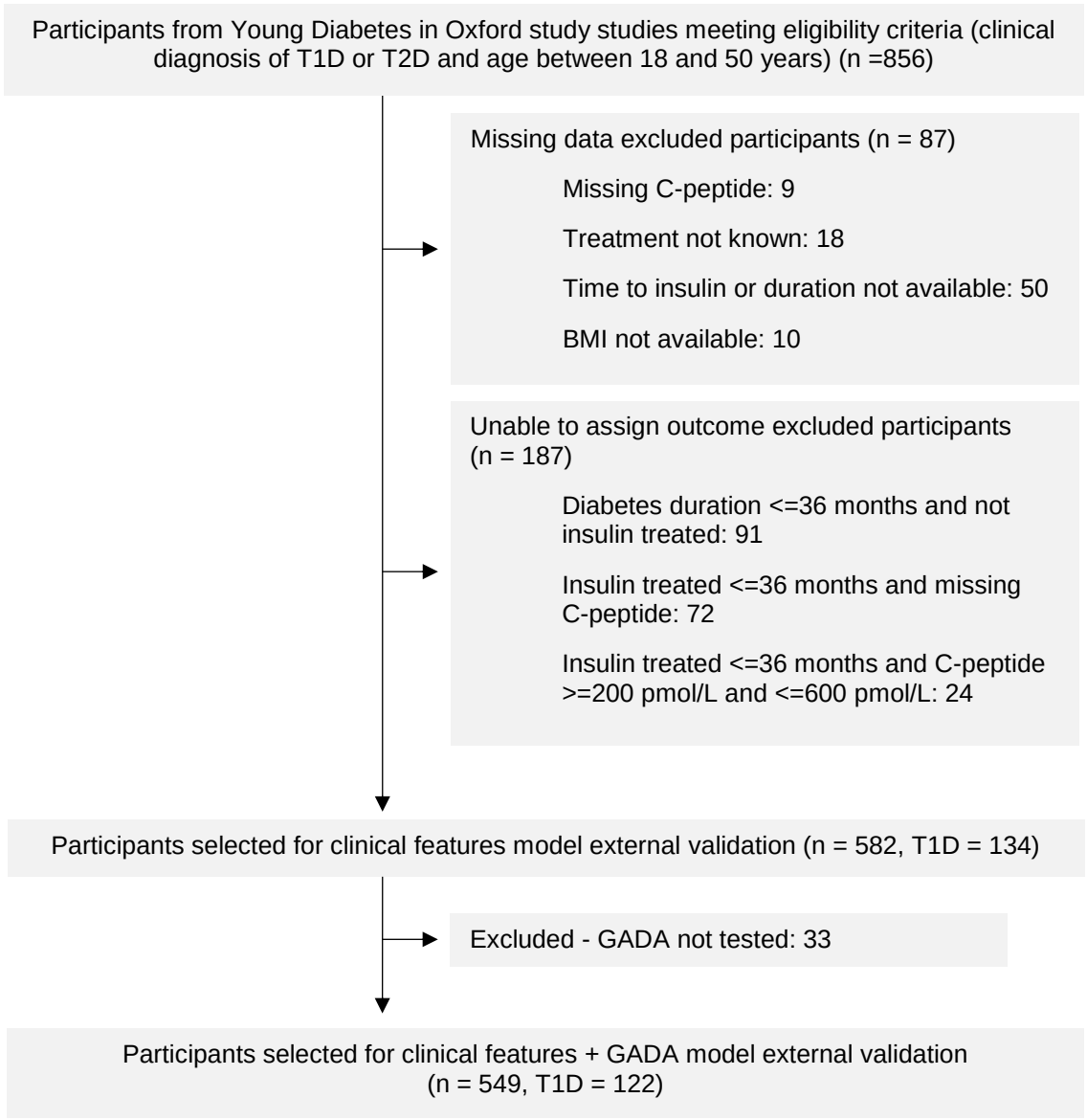
Supplementary Table 8: Model performance comparison with replica models developed using stage 4 development sample (n = 943).

Model	Clinical features ROC AUC	Clinical features + GADA ROC AUC	Clinical features + GADA + IA-2 ROC AUC
Development sample 1 (n = 1,352)	0.90 [0.88, 0.93]	-	-
Development sample 2 (n = 1,036)	-	0.96 [0.95, 0.97]	-
Development sample 3 (n = 1,025)	-	-	0.96 [0.95, 0.98]
Development sample 4 (n = 943)	**0.91 [0.89, 0.94]**	**0.96 [0.94, 0.97]**	**0.96 [0.95, 0.98]**

Supplementary Table 9: ROC AUC calculated including only patients aged 25-35 years (inclusive) at diagnosis and with BMI 25-35 kg/m^2 (inclusive).

Model	ROC AUC [95% CI]	n
Clinical Features	0.72 [0.61, 0.83]	104
Clinical Features + GADA	0.89 [0.80, 0.98]	78
Clinical Features + GADA + IA2	0.89 [0.80, 0.98]	77
Clinical Features + GADA + IA2 + T1D GRS	0.95 [0.90, 1.00]	71

Supplementary Figure 3: Flow diagram of participants through the model external validation stages. T1D: type 1 diabetes, T2D: type 2 diabetes



Supplementary Table 10: Baseline characteristics comparison of the development and validation data sets for: Model 1 – Clinical features (Age at diagnosis & BMI) and Model 2 – Clinical features + GADA. *Measured at recruitment (median 13 years and 14 years post diagnosis in development data sets and validation data sets). Kruskal-Wallis used for comparison testing continuous variables, chi-square for categorical variables.

	Model 1 development n = 1,352	**Model 1 validation n = 582**	**comparison p value**	**Model 2 development n = 1,036**	**Model 2 validation n = 549**	**comparison p value**
Characteristic						
Sex (% Male)	59%	61%	>0.1	59%	61%	> 0.1
Age at diagnosis (years)	43 (36, 48)	37 (30, 41)	<0.001	43 (36, 48)	37 (30, 41)	< 0.001
BMI (kg/m^2)*	33 (28, 38)	31 (27, 36)	<0.001	33 (28, 38)	31 (27, 36)	< 0.001
Duration of diabetes (years)*	13 (8, 20)	14 (8, 23)	0.03	13 (8, 20)	13 (8, 23)	> 0.1
Type 1 diabetes	13%	23%	<0.001	14%	22%	< 0.001
HbA_{1c} (%)*	8.2 (7.1, 9.6)	8.1 (7.2, 9.3)	>0.1	8.3 (7.3, 9.8)	8.1 (7.2, 9.4)	0.08
HbA_{1c} (mmol/mol)*	66 (54, 81)	65 (55, 78)	>0.1	67 (56, 84)	65 (55, 79)	0.08
GADA (% positive)	-	-	-	12%	20%	< 0.001

Supplementary Table 11: Model performance results for the external validation of the clinical features and clinical features+ GADA models. * P value for Brier score is Spiegelhalter's z-test used to evaluate the calibration component of the Brier score, significant p-values indicate poor calibration.

Performance parameter	External validation
Clinical features model (n = 582)	
ROC [95% CI]	0.86 [0.83, 0.90]
Expected/Observed	1.06
Calibration-in-the-large ($a|b_L$=1)	-0.14
Calibration slope (b_L)	0.85
Overall misclassification	-0.14 p = 0.05
Brier Score*	0.11 (p = 0.14)
Clinical features + GADA model (n = 549)	
ROC [95% CI]	0.93 [0.90, 0.96]
Expected/Observed	1.08
Calibration-in-the-large ($a|b_L$=1)	-0.23
Calibration slope (b_L)	0.90
Overall misclassification	-0.10 p > 0.1
Brier Score*	0.08 (p = 0.29)

Supplementary Table 12: Classification table comparing the development and validation samples at different cut-offs for probability of type 1 diabetes using the clinical features and clinical features + GADA logistic regression models.
PPV and NPV assume prevalence for type 1 diabetes: Clinical features model – 13% (development) and 23% (validation), Clinical features + GADA model - 14% (development) and 22% (validation).

Clinical features	Development (n = 1,352)					Validation (n = 582)				
	Probability cut-off					Probability cut-off				
	10	30	50	70	90	10	30	50	70	90
Sensitivity/specificity (%)	85/79	64/95	49/98	35/99	15/100	91/62	73/85	59/93	45/96	13/99
Accuracy (%)	80	90	91	90	89	69	82	85	84	79
Positive predictive value (PPV) (%)	38	64	79	83	90	42	59	71	77	77
Negative predictive value (NPV) (%)	97	95	93	91	89	96	91	88	85	79

Clinical features + GADA	Development (n = 1,036)					Validation (n = 549)				
	Probability cut-off					Probability cut-off				
	10	30	50	70	90	10	30	50	70	90
Sensitivity/specificity (%)	90/88	80/96	66/97	52/99	31/100	97/75	86/89	75/93	55/96	42/97
Accuracy (%)	89	94	93	92	90	80	88	88	87	85
Positive predictive value (PPV) (%)	55	75	80	85	92	53	69	73	80	81
Negative predictive value (NPV) (%)	98	97	95	93	90	99	96	93	88	85

Supplementary table 13: Characteristics of participants with probability of Type 1 diabetes > 80% but with type 2 diabetes actual outcome *Non fasting equivalent, measured > 5 years post diagnosis (unless < 200 pmol/L prior to 5 years). † C-peptide measured at single screening visit. ‡Clinical features + GADA model applied to participants in the YDX study.

Age at diagnosis (years)	BMI (kg/m^2)	GADA positive	C-Peptide (pmol/L)*	Insulin Treated	Time to insulin (months)	Duration at screening (years)†	Actual diabetes outcome	Probability of type 1 diabetes‡ (%)
18	26	0	775	1	Immediate	15	Type 2 diabetes	80
21	23	0	868	1	Immediate	10	Type 2 diabetes	82
27	29	1	-	0	-	3	Type 2 diabetes	88
38	22	1	550	1	48	10	Type 2 diabetes	88
36	22	1	175	1	72	12	Type 2 diabetes	89
23	32	1	25	1	48	29	Type 2 diabetes	90
30	25	1	25	1	36	30	Type 2 diabetes	91
29	25	1	225	1	48	12	Type 2 diabetes	93
23	28	1	50	1	120	28	Type 2 diabetes	95
33	21	1	65	1	96	47	Type 2 diabetes	95
34	20	1	25	1	120	22	Type 2 diabetes	96
23	22	1	-	0	-	3	Type 2 diabetes	99

Supplementary table 14: Characteristics of participants with probability of Type 1 diabetes < 16% (Youden's Index cut-off) but with type 1 diabetes actual outcome *Non-fasting equivalent, measured > 5 years post diagnosis (unless < 200 pmol/L prior to 5 years). † C-peptide measured at single screening visit. ‡Clinical features + GADA model applied to participants in the YDX study.

Age at diagnosis (years)	BMI (kg/m^2)	GADA positive	C-Peptide (pmol/L)*	Insulin Treated	Time to insulin (months)	Duration at screening (years)†	Actual diabetes outcome	Probability of type 1 diabetes (%)‡
41	40	0	50	1	12	41	Type 1 diabetes	0.6
40	34	0	198	1	12	34	Type 1 diabetes	1.8
43	31	0	125	1	3	1	Type 1 diabetes	2.1
39	33	0	25	1	24	17	Type 1 diabetes	2.5
38	25	0	68	1	Immediate	19	Type 1 diabetes	12.7
39	40	1	50	1	Immediate	16	Type 1 diabetes	14.9

Supplementary table 15: Model performance results for the four additional models in the online calculator. * Result reported as raw cross-validation estimate of prediction error with misclassification cost function (cut-off 0.5). cv.glm function in R version 3.3.3

Model	ROC [95% CI]	Jack-knife cross validation *
Clinical features + IA-2	0.93 [0.90, 0.95]	0.07
Clinical features + T1D GRS	0.93 [0.90, 0.95]	0.08
Clinical features + IA-2 + T1D GRS	0.95 [0.93, 0.97]	0.06
Clinical features + GADA + T1D GRS	0.97 [0.96, 0.98]	0.07

Supplementary Table 16: Clinical features logistic regression model (model 1). * Log transformed. Linear Predictor mean -2.96, sd 1.98

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	37.94 (2.67)	-	-
Age at diagnosis (years) *	-5.09 (0.41)	0.006 [0.003, 0.014]	<0.001
BMI (kg/m^2) *	-6.34 (0. 60)	0.002 [0.001, 0.005]	<0.001

Supplementary Table 17: Clinical features + GADA logistic regression model (model 2). Linear Predictor mean -3.37, sd 2.53

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-0.98 (0.19)	-	-
Model 1 linear predictor	0.94 (0.08)	2.57 (2.18, 3.03)	< 0.001
GADA positive	3.11 (0.32)	22.50 (12.13, 41.76)	< 0.001

Supplementary Table 18: Clinical features + GADA + IA-2 logistic regression model (model 3).
Linear Predictor mean -3.55, sd 2.58

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-1.28 (0.21)	-	
Model 1 linear predictor	0.92 (0.09)	2.50 [2.10, 2.98]	< 0.001
Antibody status - GADA positive only	3.08 (0.35)	21.81 [11.06, 43.02]	< 0.001
Antibody status - IA-2 positive only	3.49 (0.78)	32.93 [7.11, 152.64]	< 0.001
Antibody status - GADA & IA-2 both positive	4.35 (0.75)	77.53 [17.74, 338.84]	< 0.001

Supplementary Table 19: Clinical features + GADA + IA-2 + T1D GRS logistic regression model (model 4). T1D GRS standardized using mean 0.2356997, sd 0.0363499. Linear Predictor mean -3.74, sd 2.89.

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-0.67 (0.24)	-	-
Model 3 linear predictor	0.88 (0.08)	2.40 [2.06, 2.80]	< 0.001
T1D GRS (per 1 SD change)	1.08 (0.21)	2.93 [1.96, 4.39]	< 0.001

Supplementary Table 20: Clinical features + IA-2 logistic regression model. Linear Predictor mean -3.17, SD 2.28

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-0.36 (0.17)	-	-
Model 1 linear predictor	0.99 (0.08)	2.70 [2.30, 3.16]	< 0.001
IA-2 positive	3.19 (0.55)	24.39 [8.27, 71.92]	< 0.001

Supplementary Table 21: Clinical features + T1D GRS logistic regression model. T1D GRS standardized using mean 0.2360879, sd 0.0358468. Linear Predictor mean -3.180108, sd 2.401089.

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-0.65 (0.18)	-	-
Model 1 linear predictor	0.87 (0.07)	2.39 [2.09, 2.74]	< 0.001
T1D GRS (per 1 SD change)	1.22 (0.15)	3.38 [2.51, 4.54]	< 0.001

Supplementary Table 22: Clinical features + IA-2 + T1D GRS logistic regression model. T1D GRS standardized using mean 0.235673, sd 0.0363399. Linear Predictor mean -3.537275, sd 2.79395.

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-1.12 (0.23)	-	-
Model 1 linear predictor	0.87 (0.09)	2.40 [2.02, 2.84]	< 0.001
T1D GRS (per 1 SD change)	1.36 (0.20)	3.89 [2.64, 5.74]	< 0.001
IA-2 positive	2.95 (0.65)	19.17 [5.33, 68.81]	< 0.001

Supplementary Table 23: Clinical features + GADA + T1D GRS logistic regression model. T1D GRS standardized using mean 0.2359649, sd 0.0363407. Linear Predictor mean - 3.596086, sd 2.868552.

Included	β (SE)	Odds Ratio [95% CI]	p value
Constant (intercept)	-1.50 (0.24)	-	-
Model 1 linear predictor	0.85 (0.09)	2.33 [1.97, 2.76]	< 0.001
T1D GRS (per 1 SD change)	1.12 (0.20)	3.05 [2.09, 4.46]	< 0.001
GADA positive	2.63 (0.34)	13.89 [7.17, 26.90]	< 0.001

Supplementary Table 24: *To convert to probability use exp(lp)/(1+exp(lp)). †Dummy variable: negative = 0, positive = 1 ‡Dummy variables: false = 0, true = 1, AntiStatus1 = GADA positive only, AntiStatus2 = IA-2 positive only, AntiStatus3 = Both GADA and IA-2 positive.

Model	Linear predictor (lp) regression equation*
Clinical features	37.94 + (-5.09 * log(age)) + (-6.34 * log(BMI))
Clinical features + GADA†	34.8057844720 + (-4.801441792 * log (Age)) + (-5.980577792 * log(BMI)) + (2.937107976 * GADA†)
Clinical features + GADA + IA-2	33.49649577 + (-4.665598345 * Log(Age)) + (-5.81137397 * Log(BMI)) + (3.082366 * AntiStatus1‡) + (3.494462 * AntiStatus2‡) + (4.350717 * AntiStatus3‡)
Clinical features + GADA + IA-2 + T1D GRS	21.57649882 + (-4.086215772 * Log(Age)) + (-5.096252172 * Log(BMI)) + (2.702010666 * AntiStatus1‡) + (3.063255174 * AntiStatus2‡) + (3.813850704 * AntiStatus3‡) + (30.11052 * T1D GRS)
Clinical features + IA-2	37.26905033 + (3.194096 * IA-2†) + (-5.047657308 * Log(Age)) + (-6.287258808 * Log(BMI))
Clinical features + T1D GRS	24.46138054 + (-4.443506884 * Log(Age)) + (-5.534741384 *Log(BMI)) + (33.93968 * T1D GRS)
Clinical features + IA-2 + T1D GRS	23.2151829 +(2.953142 * IA-2†) + (-4.446784844 *Log(Age))+(-5.538824344 * Log(BMI)) + (37.40205 * T1D GRS)
Clinical features + GADA + T1D GRS	23.20924904 + (2.63093 * GADA†) + (-4.303557843 * Log(Age)) + (-5.360423718 *Log(BMI)) + (31.22606 * T1D GRS)

Chapter 3.

Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the classification of type 1 and type 2 diabetes in young adults

Anita L Lynam, John M Dennis, Katharine R Owen, Richard A Oram,

Angus G Jones, Beverley M Shield and Lauric A Ferrat

Submitted to Diagnostic and Prognostic Research August 2019

Acknowledgments of co-authors and contributions to paper

I conceived the idea and designed the study. I performed a review of the existing literature. I adapted an existing R script for this study including amending the code for additional machine learning algorithms, and writing new code for the hyperparameter tuning and validation tests. I performed the analysis and interpreted the results. I drafted the manuscript.

Lauric Ferrat assisted with the study design. Angus Jones researched the Exeter data and Kathrine Owen researched the YDX data. Lauric Ferrat is the original author of the R script. Lauric Ferrat, John Dennis and Beverley Shields assisted with analysing the data. Richard Oram and Angus Jones discussed and contributed to study design, and provided support for the clinical interpretation of results. Lauric Ferrat and Beverley Shields assisted with drafting the manuscript. All authors critically revised the manuscript and approved the final version.

Catherine Angwin of the NIHR Exeter Clinical Research Facility assisted with the data preparation. Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital conducted the autoantibody analysis on the GADA islet-autoantibody samples.

Abstract

Objective

There is much interest in the use of prognostic and diagnostic prediction models in all areas of clinical medicine. The use of machine learning to improve prognostic and diagnostic accuracy in this area has been increasing at the expense of classic statistical models. Previous studies have compared performance between these two approaches but their findings are inconsistent and many have limitations. We aimed to compare the discrimination and calibration of six models built using logistic regression and optimised machine learning algorithms in a clinical setting, where the number of potential predictors is often limited, and externally validate the models.

Research design and methods

We trained models using logistic regression and five commonly used machine learning algorithms to classify diabetes (type 1 versus type 2) based on three pre-specified predictor variables (Age, BMI and GADA islet-autoantibodies) using a UK cohort of adult participants (aged 18–50 years) with clinically diagnosed diabetes recruited from primary and secondary care (n = 1,036). Discrimination performance (ROC AUC and AUPRC) and calibration of each approach was compared in a separate external test dataset (n = 549).

Results

Average performance obtained in model training was similar in all models (ROC AUC >= 0.94). In external validation, decreases in performance were observed in all models. Calibration tests showed that all models overstated predicted risk and most had evidence of miscalibration.

Conclusions

Logistic regression performed as well as optimised machine algorithms to classify patients with type 1 and type 2 diabetes. This study highlights the utility of comparing traditional regression modelling to machine learning, particularly when using a small number of well understood, strong predictor variables.

There is much interest in the use of prognostic and diagnostic prediction models in all areas of clinical medicine including cancers (1, 2), cardiovascular disease (3, 4) and diabetes (5, 6). These models are increasingly being used as web-calculators (7-9) and medical apps for smartphones (10-12), and many have been incorporated into clinical guidelines (13-22).

There are many different approaches that can be used for developing these models. Classic statistical models such as logistic regression are commonly applied but there is increasing interest in the application of machine learning to improve prognostic and diagnostic accuracy in clinical research (23-26) with many examples of their use (27-33). Machine learning (ML) is a data science field dealing with algorithms in which computers (the machines) adapt and learn from experience (data), these algorithms have the ability to process the vast amounts of data, complex interactions and non-linearity. Supervised Learning is the most widely employed category of machine learning. In Supervised Learning, the machine predicts the value of an outcome (either binary or continuous) trained on a set of predictor variables.

There are many applied studies comparing the performance of classic models to different machine learning algorithms (34-45) but their findings are inconsistent. Many such comparison studies have limitations; not all use non-default parameter settings (hyperparameter tuning) or have validated performance on external data (46). Discrimination, as measured by area under the receiver operating characteristic curve, is almost always provided but studies have rarely assessed whether risk predictions are reliable (calibration) (46).

We aimed to use a methodological approach to explore and compare performance of machine learning and a classic statistical modelling approach using an example of a diabetes classification model. Classification of diabetes offers an interesting case study as it is an area where there is considerable misclassification in clinical practice. Type 1 diabetes and type 2 diabetes can be hard to distinguish between, particularly in adults aged between 18 and 50.

Methods

We focus on the capacity of each machine learning algorithm in a specific context using real data as the basis for our comparisons. An alternative method of comparing machine learning algorithms is to use simulation. Whilst simulation studies are interesting, the choice of model used to generate the simulation data can introduce bias. Our use of real data avoids this potential bias. In addition, our use of a real data allows us to test the algorithms in an external dataset with different data collection methods, simulated data is unable to capture such differences. In summary, our decision to use real data ensures that we are comparing the performance of the algorithms in a setting representative of clinical practice.

Sample size was checked using events per variable. For machine algorithms it has been suggested that over ten times as many events per variable is required to achieve stable results compared to traditional statistical modelling. For three predictors, this means that 300 events are required.

We selected a classic model and five supervised machine learning algorithms that 1) were appropriate for classification problems and 2) had been used previously in medical applications: Logistic Regression, Gradient Boosting Machine, Support Vector Machine (with Radial Basis Function Kernel), K-

Nearest Neighbours, Neural Network and Random Forest machine learning algorithms. We trained models using each algorithm, incorporating hyperparameter tuning, and compared the performance of the optimised models on a separate external test dataset.

Study population – training

Participants with clinically diagnosed diabetes were identified from Exeter, UK-based cohorts (47-50). Summaries of the cohorts including recruitment and data collection methods are shown in Supplementary Table 1. Only participants that had a clinical diagnosis of type 1 or type 2 diabetes between the ages of 18 and 50 years were eligible.

Study population – external test dataset

Participants were identified from the Young Diabetes in Oxford (YDX) study (51). Participants were recruited in the Thames Valley region, UK, and diagnosed with diabetes up to the age of 45 years. The same eligibility criteria were applied to this cohort.

All participants included in this study (training and test datasets) were of white European origin.

Model outcome (dependent variable): type 1 and type 2 diabetes definition

We used a binary outcome with values type 1 or type 2 diabetes. Type 1 diabetes was defined as having insulin treatment within <= 3 years of diabetes diagnosis and severe insulin deficiency (non–fasting C peptide < 200pmol/L). Type 2 diabetes was defined as either 1) no insulin requirement for 3 years from diabetes diagnosis or 2) where insulin was started within 3 years of diagnosis, substantial retained endogenous insulin secretion (C-peptide >600pmol/L) at

>=5 years diabetes duration. Participants not meeting the above criteria or with insufficient information were excluded from analysis, as type of diabetes and rapid insulin requirement could not be robustly defined (n = 342).

Predictor variables

We used three pre-specified predictor variables, age at diagnosis, BMI and GADA islet-autoantibodies. All three predictor variables have evidence for utility at diabetes diagnosis (52-54). Age at diagnosis was self-reported by the participant. Height and weight was measured at study recruitment by a research nurse to calculate BMI. Age at diagnosis and BMI were modelled as continuous variables and were standardised (55). GADA islet-autoantibodies were dichotomized into negative or positive based on clinically defined cut-offs, in accordance with clinical guidelines (56).

We removed all observations with missing predictor values (complete-case analysis).

Statistical analysis

Model training

All models were trained using the entire training dataset. We evaluated six classification algorithms; Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Neural Network (NN), K-Nearest Neighbours (KNN) and Random Forest (RF). For SVM we used the Radial Basis Function kernel parameter (55) and for NN we used the most commonly used single-hidden-layer neural network (55) trained using Quasi-Newton back propagation (BFGS) (57) optimisation method. There are no clear guidelines regarding either the choice of algorithms or the advantages and disadvantages

of each in specific clinical settings. A brief summary of each algorithm is shown in Table 1.

Table 1: Algorithm description and references

Algorithm	Description	References
Logistic Regression	A classic statistical algorithm for binary outcomes that uses maximum likelihood estimation. It is fully parametric but has a number of assumptions that need to be satisfied such as weak collinearity between the variables. There are no model parameters to be set. Coefficients are adjusted to allow for dependence between the characteristics. Is useful for inference, estimation, interpretation and prediction.	(55, 58-60)
Support Vector Machine	An artificial intelligence based method. It is a quadratic optimisation problem involving minimising penalties and maximizing margin width, the two classes are separated by constructing nonlinear decision boundaries (hyperplanes) using kernel trick that maximise the margin between them. It is non-parametric and requires penalty and kernel function parameters to be set.	(55, 61, 62)
Gradient Boosting Machine	An ensemble learning technique similar to random forest in the sense they average a large number of decision trees to make prediction. The difference between the two is the application of gradient boosting. In gradient boosting, the decision trees are trained sequentially with the weights of each successive model adjusted based on reducing the errors of the previous model. After few steps of the algorithm, the new decision trees are able to handle hard to fit data. Finally, the predicted class is determined from the average estimated class probability (or majority vote of predicted class) calculated over the ensemble of trees.	(55, 63, 64)
Neural Network	An artificial intelligence based method using an adaptive and non-sequential approach to learning that mimics a biological neural network. It is a non-parametric technique, it uses all the predictor variables resulting in complex models.	(55, 65-68)
K-Nearest Neighbours	A model-free method; it is a type of instance-based learning or lazy learning in which there is no training phase, instead the algorithm memorises the training data. Based on the principle that observations located close together in n-dimensional space will have the same outcome, the classification process involves a search the entire dataset for the k training points closest in Euclidean distance (k-neighbours), the predicted class is determined based on a majority vote of the actual class among these k-neighbours.	(55, 66, 69, 70)

Algorithm	Description	References
Random Forest	A popular artificial intelligence based algorithm that grows a large ensemble of classification trees on bootstrapped samples using a random selection of the predictor variables and performs bagging for class selection; after all the trees have been grown, the predicted class is determined from the average estimated class probability (or majority vote of predicted class) calculated over the ensemble of trees.	(55, 71, 72)

All models were trained using 5 repeats of 10-fold cross validation resampling method. We applied Synthetic Minority Over-Sampling Technique (SMOTE) inside of cross-validation to deal with imbalanced data (73). While real-world data medical applications are likely to be unbalanced, the use of sampling methods such as SMOTE can improve model prediction performance. We used a grid search to tune the model parameters (hyperparameter tuning) (74), i.e. optimize the performance of the machine learning algorithm. The hyperparameter metrics applied in the grid searches are shown in Supplementary Table 2. Optimal models were selected using the maximum mean area under the receiver operating characteristic curve (ROC AUC) calculated in the cross-validation.

Model performance measures

We used ROC AUC (75) and precision recall curve (AUPRC) as the summary metrics to evaluate model discrimination. The ROC AUC quantifies the probability that the risk scores from a randomly selected pair of individuals with and without this condition are correctly ordered. AUPRC is a more sensitive performance metric when dealing with strongly imbalanced data (unequal percentage in each class); it evaluates the performance of the model in regard of only one class and does not take into the account the ability of the model to

identify the second class (76-78). For both measures, a value of 1 indicates a perfect test.

We assessed calibration visually using calibration plots and statistically using calibration tests (calibration slope).

External testing

For each optimal model developed in the training dataset, external performance was evaluated in the YDX study cohort and compared to the internal (cross-validation resampling) performance. Calibration was investigated using calibration curves. We also checked for correlation in the predictions from each model.

Variable Importance

We assessed and compared the predictor variable importance (VI) in the optimal models (79). The VI model-specific metrics were scaled to derive values proportional to the most important predictor having value 100. VI metrics were not available for the SVM or KNN models.

Software

All analysis was performed using R software (version 3.5.2). Model training, internal evaluation and variable importance were performed using the Caret R package (79-83). VI model specific metrics were obtained using the Caret VarImp function.

Code

In Appendix 2 we share the code to allow reproduction of similar comparisons of machine learning algorithms with any number of predictor variables.

Results

1,036 participants in the Exeter cohort met inclusion criteria and were included in the training dataset, of whom 140 (14%) were classified as having type 1 diabetes. 549 participants (type 1 diabetes n = 122 (22%)) in the YDX cohort met criteria and were included in the external validation test dataset. Compared to the participants in the Exeter cohort, the participants in the YDX cohort were younger at diagnosis (consistent with the narrower age range in YDX (18-45y) (median 37 years vs 43 years, $p < 0.001$)), had a lower BMI (median 31 kg/m2 vs 33 kg/m2, $p < 0.001$), had a higher percentage of GADA (20% versus 12%, $p < 0.001$) and a higher prevalence of type 1 diabetes by study definition (22% vs 14%, $p < 0.001$) (Supplementary Table 3 for participant characteristics).

The average (mean) performance ROC AUC for the optimal models obtained in the resampling was high in all six models (ROC AUC >= 0.94) (Table 2 (resampling ROC AUC column)) with no difference in performance between models. Supplementary Table 2 includes the final model tuning parameters selected for the optimal models in the cross validation resampling.

Table 2: ROC AUC [95% CI] and AUPRC performance comparison of the six optimal models applied to the resampling and test datasets.

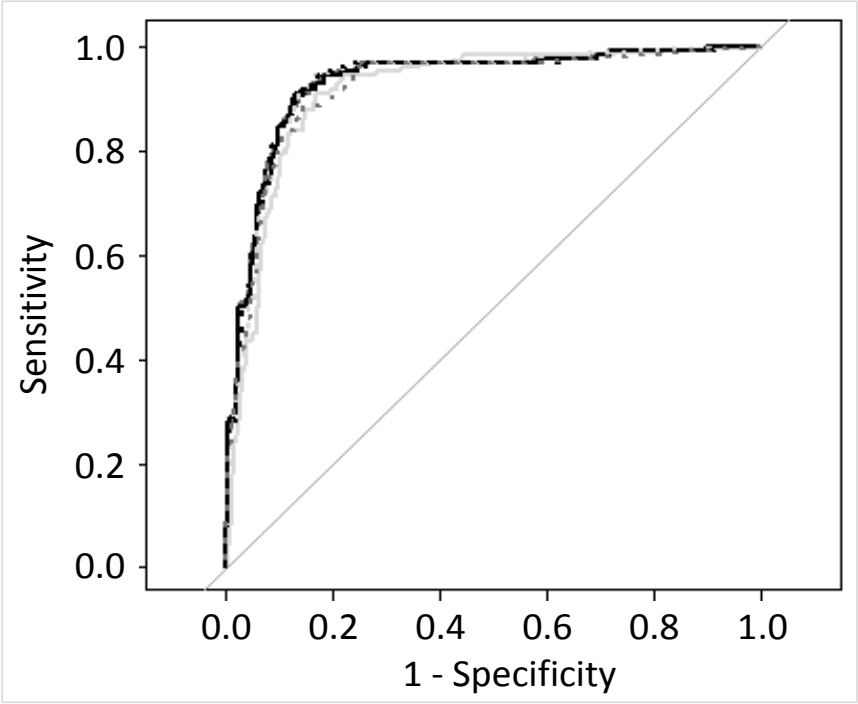
Model	Resampling ROC AUC	Test ROC AUC	Test AUPRC
Gradient Boosting Machine	0.96 [0.92, 1.00]	0.93 [0.90, 0.95]	0.88
Logistic regression	0.96 [0.90, 1.00]	0.93 [0.90, 0.96]	0.76
Support Vector Machine	0.96 [0.91, 1.00]	0.93 [0.90, 0.96]	0.75
Neural Network	0.96 [0.90, 1.00]	0.93 [0.90, 0.96]	0.84
Random Forest	0.94 [0.89, 0.99]	0.91 [0.89, 0.94]	0.87
K-Nearest Neighbours	0.95 [0.89, 1.00]	0.92 [0.89, 0.95]	0.88

Table 3: Calibration test results on test dataset.

Model	Calibration slope (b_L)	Calibration-in-the-large ($a \mid b_L=1$)	Overall misclassification
Gradient Boosting Machine	1.328	-0.738	0.328, p = 0.003
Logistic regression	0.808	-0.784	-0.192, p = 0.008
Support Vector Machine	0.776	-0.845	-0.224, p = 0.001
Neural Network	0.886	-0.746	-0.114, p = 0.138
Random Forest	0.359	-0.783	-0.641, p < 0.001
K-Nearest Neighbours	0.586	-0.260	-0.414, p < 0.001

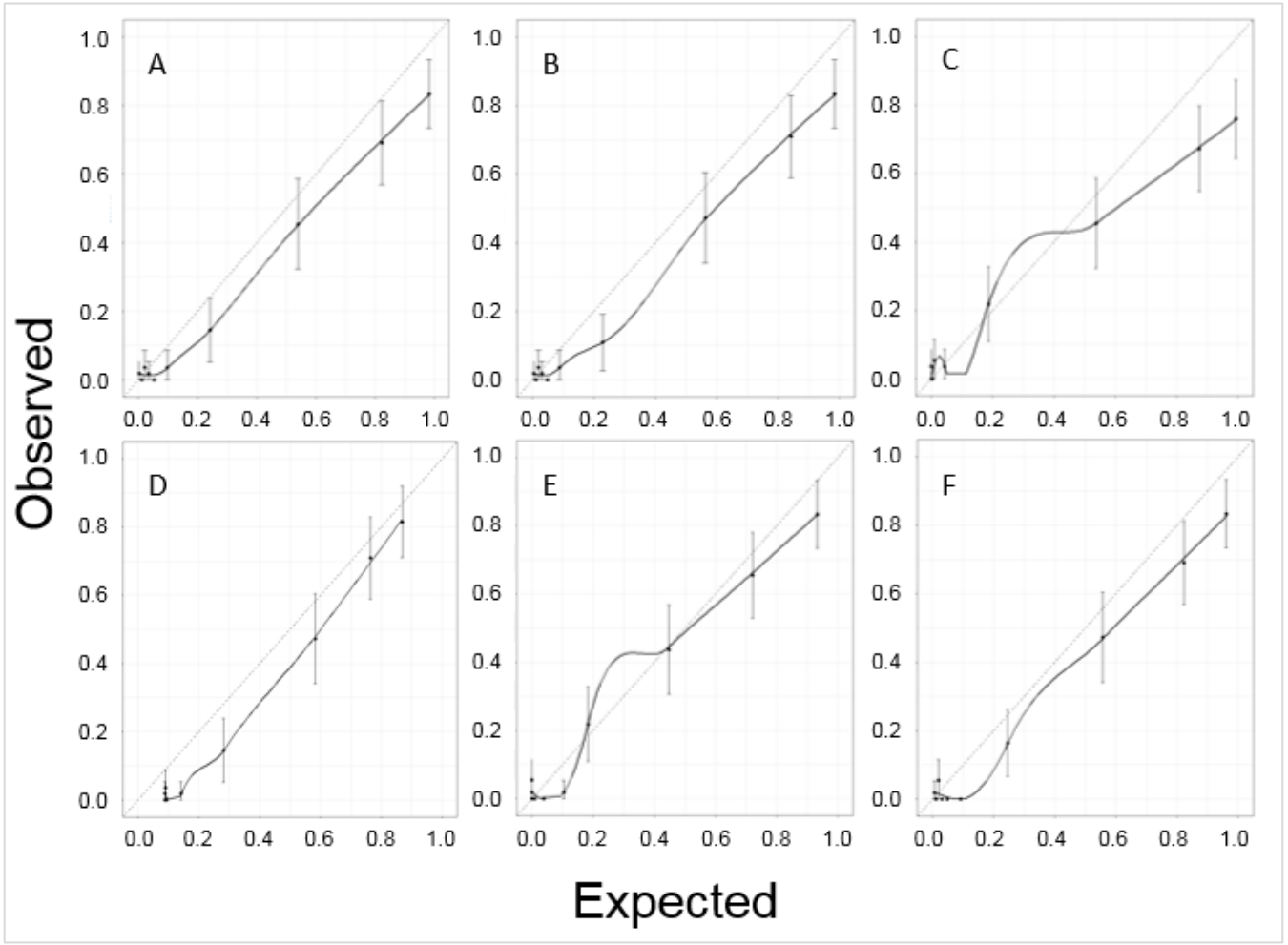
There was a decrease in the ROC AUC of all models when they were applied to the external test dataset (Table 2 (test ROC AUC column)) but all still showed high levels of performance (ROC AUC >= 0.90, Figure 1). When model performance on the external test dataset was assessed using AUPRC, there was a clear difference in performance of LR and SVM, and the other models (Supplementary Figure 1 and Table 2 (test AUPRC column)). Model predictions were highly correlated across models (Supplementary Table 4).

Figure 1: ROC AUC plots obtained using external validation dataset for six prediction models
Legend:
Solid lines: black = Support Vector Machine, dark grey = Logistic Regression, light grey = Random Forest
Dotted lines: black = Neural Network, dark grey = K-Nearest Neighbours, light grey = Gradient Boosting Machine



In the calibration tests performed on the external test dataset, all models over-estimated type 1 diabetes prevalence (Figure 2 and Table 3 (calibration in the large values < 0 indicate over-estimating risk)) and there was evidence of miscalibration (significant overall misclassification p values indicate miscalibration) in all models (often due to an underestimation of type 1) except for NN.

Figure 2: Calibration plots obtained using external validation dataset for prediction models: A: Logistic Regression B: Support Vector Machine C: Random Forest D: Gradient Boosting Machine E: K-Nearest Neighbours F: Neural Network.
Legend: Dashed line = reference line, Solid black line = loess smoother, 95% CI on observed data



Although performance was similar, variable importance differed by model in the training dataset (Supplementary Figure 2). Relative to other predictors, Age at diagnosis was more important in the LR, GBM and RF models than the neural network model, BMI was very important only in the GBM. GADA had similar high importance in all the four models assessed.

Conclusions

We found similar performance when applying logistic regression and five optimised machine learning algorithms to classify type 1 and type 2 diabetes, in both training and test datasets. Performance was high for all models. In calibration tests, all models overstated predicted risk and most had evidence of miscalibration. The choice of algorithm in this study made very little difference to the discrimination performance of the models.

Strengths of our study include the use of a systematic approach to model comparison dealing with limitations from previous studies (46, 84) including: 1) use of different datasets to train and test models, 2) use of default tuning parameters (35, 41) and 3) calibration (23). We have used the same dataset to train all our models; since model performance will differ between settings, use of the same dataset is crucial for valid model comparisons. The choice of tuning parameters will affect the performance of the model (74), we have optimised our models by applying hyperparameter tuning using a recognised grid search approach. We have increased the validity of our results by using an external test dataset.

We have compared several machine learning algorithms that have been selected for their suitability to our setting. The use of only three predictor variables means that we have a very low risk of overfitting. The use of only

three predictors may also be considered as a limitation of our study since these machine learning algorithms are designed to deal with larger datasets and more variables. Working with a few meaningful predictors is common in clinical settings and knowing the performance of machine learning models using low numbers of predictors is important. It is possible with more variables, machine learning approaches may prove more discriminative. However, we have achieved excellent performance using just these three predictors. Another limitation of our study is that we judge the model only on its performance. In real practice we would want to consider ease of implementation and interpretation when selecting the 'best' model.

For machine learning algorithms it has been suggested that over ten times as many events per variable is required to achieve stable results compared to traditional statistical modelling (85). Although we did not have the sufficient number of events per variable to meet this criteria (140 actual events compared to 300 suggested), the results of the external validation suggest stability was achieved.

The performance ranking of the models differed when ordering by each of the two discrimination performance measures (ROC AUC and AUPRC), it is therefore important that the performance measure being reported is the most appropriate for the individual clinical setting. In our study, ROC AUC is appropriate as we place equal weight on each type of misclassification error. In this setting LR, SVM and NN are the best models. If accuracy of estimated probability were an importance factor NN would come at the best approach. If wrongly identifying type 1 diabetes for type 2 diabetes was important then KNN and GBM with the highest AUPRC would be the best models. Overall the notion

of best model is context dependent but in this study the models perform similarly.

The observed decrease in ROC AUC when assessed in the external test data highlights the importance of external validation to test the transportability of models. Indeed, all of the algorithms underperformed in the test set. The models fit on the training data set might be over-fitted and their performance could be overestimated despite a rigorous internal validation. Other reasons might be that the test dataset used different GADA and C-peptide assays, and the different populations – this may diminish performance and does not necessarily mean over-fitting.

The performance of LR on both training and test datasets shows that classic algorithms can perform as well as more advanced algorithms even when disadvantaged by assuming linearity in the predictors. LR models are relatively easy to use and understand compared to machine learning algorithms where usage is limited by the difficultly of interpreting the model, often referred to as a “black boxes”. LR models also have a strong theoretical background which lead to the possibility of using well defined statistical tests to explore the statistical significant of the variables. There is an increasing number of studies demonstrating that LR can perform as well if not better, in a large number of settings (46). However we could not find a study that compared machine learning algorithms with optimised hyperparameters versus LR on an external dataset as we have done in this study which shows again that LR performs as well as more complex approaches.

We have shown through this study that machine learning performs similarly, however some differences subsist. However as previously described (86), each

database is unique and there is no ‘free lunch’, i.e. if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of other problems (46, 84). It is thus important to test different algorithms benchmarked against logistic regression to identify if one algorithm outperforms the other; if performance is similar then the simplest and most interpretable model can be used.

In a diabetes classification setting with three strongly predictive variables, a classic logistic regression algorithm performed as well as more advanced machine algorithms. This study highlights the utility of comparing traditional regression modelling to machine learning, particularly when using a small number of well understood, strong predictor variables. Furthermore, this article highlights once again the need to perform external validation when selecting models as we demonstrate that all algorithms can underperform on external data.

Acknowledgments

The authors thank participants who took part in these studies and the research teams who undertook cohort recruitment. We thank Catherine Angwin of the NIHR Exeter Clinical Research Facility for assistance with data preparation, and Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital for assistance with sample analysis.

References

1. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer. 2008;113(11):3075-99.
2. Amir E, Freedman OC, Seruga B, Evans DG. Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. JNCI: Journal of the National Cancer Institute. 2010;102(10):680-91.
3. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. Bmj. 2016;353:i2416.
4. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circ Cardiovasc Qual Outcomes. 2015;8(4):368-75.
5. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 2011;343.
6. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. Bmj. 2012;345:e5900.
7. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ open. 2015;5(3):e007825.
8. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabetic Medicine. 2010;27(8):887-95.
9. Rabin BA, Gaglio B, Sanders T, Nekhlyudov L, Dearing JW, Bull S, et al. Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. Cancer Epidemiol Biomarkers Prev. 2013;22(10):1645-56.
10. Watson HA, Carter J, Seed PT, Tribe RM, Shennan AH. The QUiPP App: a safe alternative to a treat-all strategy for threatened preterm labor. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2017;50(3):342-6.
11. Shields BM, McDonald TJ, Ellard S, Campbell MJ, Hyde C, Hattersley AT. The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. Diabetologia. 2012;55(5):1265-72.
12. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.
13. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. The BMJ. 2010;341:c6624.
14. Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower-Thomas K. The Nottingham Prognostic Index: five- and ten-year data for all-cause survival within a screened population. Ann R Coll Surg Engl. 2015;97(2):137-9.
15. Fox KA, Dabbous OH, Goldberg RJ, Pieper KS, Eagle KA, Van de Werf F, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). Bmj. 2006;333(7578):1091.

16. Apgar V. A proposal for a new method of evaluation of the newborn infant. Current researches in anesthesia & analgesia. 1953;32(4):260-7.
17. Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL, et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. Lancet. 2007;369(9558):283-92.
18. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical care medicine. 1985;13(10):818-29.
19. Maddrey WC, Boitnott JK, Bedine MS, Weber FL, Jr., Mezey E, White RI, Jr. Corticosteroid therapy of alcoholic hepatitis. Gastroenterology. 1978;75(2):193-9.
20. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax. 2003;58(5):377-82.
21. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81(24):1879-86.
22. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
23. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA. 2018;320(1):27-8.
24. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317-8.
25. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. 2017;15:104-16.
26. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015;13:8-17.
27. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making. 2010;10:16-.
28. Atkinson EJ, Therneau TM, Melton LJ, 3rd, Camp JJ, Achenbach SJ, Amin S, et al. Assessing fracture risk using gradient boosting machine (GBM) models. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2012;27(6):1397-404.
29. Zhang J, Xu J, Hu X, Chen Q, Tu L, Huang J, et al. Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images. BioMed research international. 2017;2017:7961494-.
30. Emir B, Masters ET, Mardekian J, Clair A, Kuhn M, Silverman SL. Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. Journal of pain research. 2015;8:277-88.
31. Ban H-J, Heo JY, Oh K-S, Park K-J. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC genetics. 2010;11:26-.

32. Wang F, Casalino LP, Khullar D. Deep Learning in Medicine—Promise, Progress, and Challenges. JAMA Internal Medicine. 2019;179(3):293-4.
33. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy. JAMA Network Open. 2018;1(3):e180926-e.
34. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. International journal of medical informatics. 2018;119:22-38.
35. van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW. Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Medical Research Methodology. 2011;11(1):143.
36. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses. PLOS ONE. 2014;9(6):e98587.
37. Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of Incident Diabetes in the Jackson Heart Study Using High-Dimensional Machine Learning. PloS one. 2016;11(10):e0163942-e.
38. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. JAMA Network Open. 2019;2(3):e190968-e.
39. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment JAMA Network Open. 2018;1(4):e181018-e.
40. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics. 2001;34(1):28-36.
41. Harrison RF, Kennedy RL. Artificial Neural Network Models for Prediction of Acute Coronary Syndromes Using Clinical Data From the Time of Presentation. Annals of Emergency Medicine. 2005;46(5):431-9.
42. Faisal M, Scally A, Howes R, Beatson K, Richardson D, Mohammed MA. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health informatics journal. 2018:1460458218813600.
43. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Statistics in medicine. 1998;17(21):2501-8.
44. Hsieh MH, Sun L-M, Lin C-L, Hsieh M-J, Hsu C-Y, Kao C-H. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer management and research. 2018;10:6317-24.
45. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiology. 2017;2(2):204-9.
46. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine

learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019.
47. DiabetesGenes.org. Diabetes Alliance for Research in England (DARE) [Cited 15/11/2018]. Available from: https://www.diabetesgenes.org/current-research/dare/.
48. ClinicalTrials.gov. RetroMASTER - Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes [Cited 15/11/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02109978.
49. ClinicalTrials.gov. MASTERMIND - Understanding Individual Variation in Treatment Response in Type 2 Diabetes (Mastermind) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT01847144?term=mastermind
50. clinicaltrials.gov. PROMASTER - PROspective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes (PROMASTER) [Cited 31/07/2018]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT02105792?term=promaster&rank=1.
51. Thanabalasingham G, Pal A, Selwood MP, Dudley C, Fisher K, Bingley PJ, et al. Systematic Assessment of Etiology in Adults With a Clinical Diagnosis of Young-Onset Type 2 Diabetes Is a Successful Strategy for Identifying Maturity-Onset Diabetes of the Young. Diabetes care. 2012;35(6):1206-12.
52. Shields BM, Peters JL, Cooper C, Lowe J, Knight BA, Powell RJ, 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).
53. Niskanen LK, Tuomi T, Karjalainen J, Groop LC, Uusitupa MIJ. GAD Antibodies in NIDDM: Ten-year follow-up from the diagnosis. Diabetes care. 1995;18(12):1557.
54. McDonald TJ, Colclough K, Brown R, Shields B, Shepherd M, Bingley P, et al. Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabetic medicine : a journal of the British Diabetic Association. 2011;28(9):1028-33.
55. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.; 2001.
56. National Institute for Health and Care Excellence. Type 1 diabetes in adults: diagnosis and management (NICE guideline NG17) 2015 [Cited 14/08/2018]. Available from: https://www.nice.org.uk/guidance/ng17.
57. Setiono R, Hui LCK. Use of a quasi-Newton method in a feedforward neural network construction algorithm. IEEE Transactions on Neural Networks. 1995;6(1):273-7.
58. Menard SW. Applied logistic regression analysis. Thousand Oaks, CA: Sage Publications; 1995.
59. van Houwelingen JC, le Cessie S. Logistic Regression, a review. Statistica Neerlandica. 1988;42(4):215-32.
60. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical decision making : an international journal of the Society for Medical Decision Making. 2001;21(1):45-56.
61. Vapnik VN. The nature of statistical learning theory: Springer-Verlag; 1995. 188 p.
62. Moguerza JM, Munoz A. Support Vector Machines with Applications. Statist Sci. 2006;21(3):322-36.

63. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001;29(5):1189-232.
64. Ridgeway G. Generalized boosted models: A guide to the gbm package. 2007(21/06/2019).
65. Goodfellow I, Bengio Y, Courville A. Deep Learning: The MIT Press; 2016. 800 p.
66. Ripley BD. Pattern Recognition and Neural Networks. New York: Cambridge University Press; 1996.
67. Hertz J, Krogh A, Palmer R. Introduction To The Theory Of Neural Computation. Redwood City, CA: Addison-Wesley; 1991.
68. Bishop C. Neural Networks for Pattern Recognition. New York: Oxford University Press; 1995.
69. Kotsiantis S, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Informatica. 2007;31:249-68.
70. Dasarathy B. Nearest Neighbor: Pattern Classification Techniques Los Alamitos, CA: IEEE Computer Society Press; 1991.
71. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
72. Ho TK, editor Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 14-16 Aug. 1995. New York: IEEE Computer society press, pp. 278-282.
73. Kuhn M, Johnson K. Applied Predictive Modeling: Springer, New York, NY.
74. Claesen M, Moor BD. Hyperparameter search in machine learning. MIC 2015: The XI Metaheuristics International Conference. 2015.
75. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145-59.
76. Wicked Good Data - r. Handling class imbalance with r and caret caveats when using the AUC 2017 [Available from: https://www.r-bloggers.com/handling-class-imbalance-with-r-and-caret-caveats-when-using-the-auc/.
77. Cook JA, Ramadas V. When to Consult Precision-Recall Curves SSRN. (01 March 2019). Retrieved 30 May 2019, from https://ssrn.com/abstract=3350582.
78. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; Pittsburgh, Pennsylvania, USA. 1143874: ACM; 2006. p. 233-40.
79. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28(5):1-26.
80. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models 2018 [Available from: https://CRAN.R-project.org/package=gbm.
81. Meyer D, Dimitriadou E, Hornik J, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien 2018 [Available from: https://CRAN.R-project.org/package=e1071.
82. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.
83. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.
84. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(1):3133-81.

85. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014;14(1):137.
86. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation. 1997;1(1):67-82.

Supplementary material

Supplementary Table 1: Cohort recruitment and data collection methods summary (training dataset).

	DARE	PRIBA	MRC Pro/RetroMaster	MRC crossover
Included participants	614	353	61	8
Data collection period	2007 to 2017	2011 to 2013	2013 to 2015	2013 to 2015
Study design	Cross-sectional	Longitudinal	Cross-sectional	Interventional Crossover
Setting	Primary and secondary care in eight diabetes research regions, England and retinal screening clinics.	Primary and secondary care in South West England	Primary and secondary care sites South West England, Tayside, Oxford, Glasgow, KCL and Newcastle, U.K.	Exeter and Tayside,U.K.
Inclusion criteria	Clinical diagnosis of diabetes (any type).	Clinical diagnosis of type 2 diabetes. Clinician determined requirement for DPP-IV inhibitor or GLP-1 analogue (HbA1C >7.5%)	Clinical diagnosis of type 2 diabetes non-insulin treated within 6 months of diagnosis. Participants were selected on the basis of rapid or slow progression to insulin therapy (<7, >7 years). Age 18-90 inclusive.	Clinical diagnosis of type 2 diabetes, currently treated with sulphonylurea tablets and no change in treatment in previous 3 months, Last HbA1c (within previous 12 months) ≥42 and ≤75 mmol/mol (6-9%). Age 19-79 inclusive.
Data collection	Clinical measurements and blood sample collected at visit. Ongoing biochemical data collected from pathology laboratories.	Clinical measurements and blood taken at initial visit. Follow up clinical measurements and blood collected at three and six months.	Clinical measures and fasting blood sample taken at visit.	MMT at baseline & MMT on each study drug visits. Three fasting blood collected at crossovers.

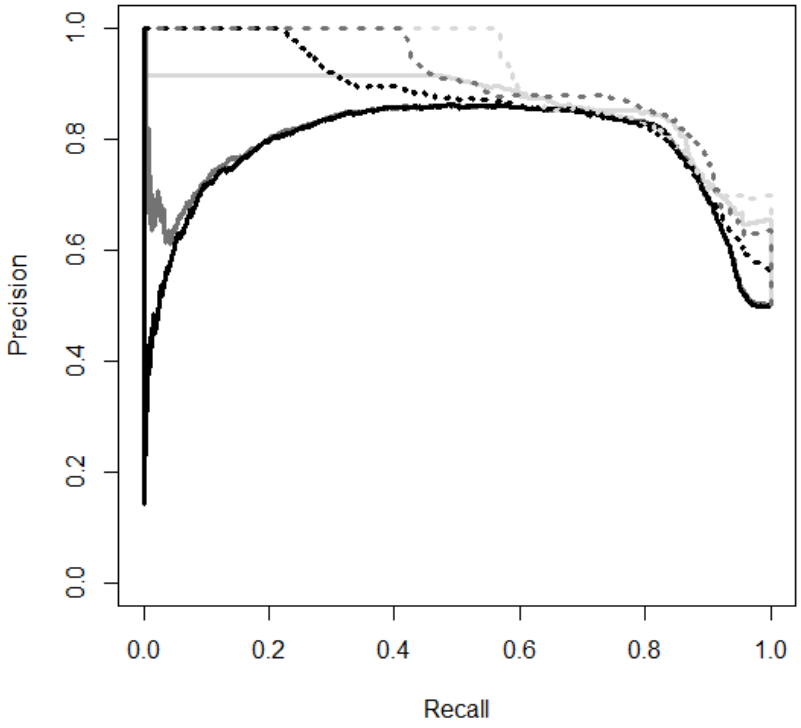
Supplementary Table 2: Model training details including the R training method used and grid search parameters applied in hyperparameter tuning, and model parameters for the optimal model selected using largest ROC AUC value. There are no model parameters for logistic regression. Hyperparameter tuning was not used for Random Forest due to the low number of predictor variables. Descriptions for search parameters are available in reference. Seed choice was set to 7 in model training.

Model	**R train method**	**Grid Search parameter values**	**Final values used for the optimal model**
Logistic Regression	glm	N/A	N/A
Gradient Boosting machine	gbm (180)	n.trees = (50,100,150,500,2000) interaction.depth = (1, 3, 6, 9, 10) shrinkage = (from 0.0005 to 0.1 by 0.001) n.minobsinnode = (5,10,15,20)	n.trees = 50, interaction.depth = 3 shrinkage = 0.0515 n.minobsinnode = 20
Support Vector Machine (with Radial Basis Function Kernel)	svmRadial (181)	sigma = (0.01, 0.1, 1, 10, 100) C = (from 0.1 to 1 by 0.05)	sigma = 0.01 C = 0.7
K-Nearest Neighbours	knn (182)	k = (from 1 to 100 by 1)	k = 99
Neural Network	nnet (182)	size = (from 1 to 10 by 1) decay = (0.5, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001)	size = 7 decay = 0.5
Random Forest	rf (183)	Not applied	mtry = 2

Supplementary Table 3: Characteristics of the Exeter, U.K. study participants included in the model training and Young Diabetes in Oxford participants included in the model external testing. Median (IQR) or %. *Measured at recruitment (median 13 years post diagnosis). Minimum and maximum values for each continuous predictor variable used in the models.

Characteristic	Training dataset n = 1,036	External validation dataset n = 549	comparison p value
Sex (% Male)	59%	61%	> 0.1
Age at diagnosis (years)	40 [39, 40]	37 [30, 41]	< 0.001
Age at diagnosis (years) min, max	18, 50	18, 49	NA
BMI (kg/m2)*	33 [32, 33]	31 [27, 36]	< 0.001
BMI (kg/m2)* min, max	17.5, 70.2	15.3, 87.7	NA
Duration of diabetes (years)	13 (8, 20)	13 (8, 23)	> 0.1
Type 1 diabetes	14%	22%	< 0.001
HbA1c (%)*	8.3 (7.3, 9.8)	8.1 (7.2, 9.4)	0.08
HbA1c (mmol/mol)*	67 (56, 84)	65 (55, 79)	0.08
GADA positive (%)	12%	20%	< 0.001

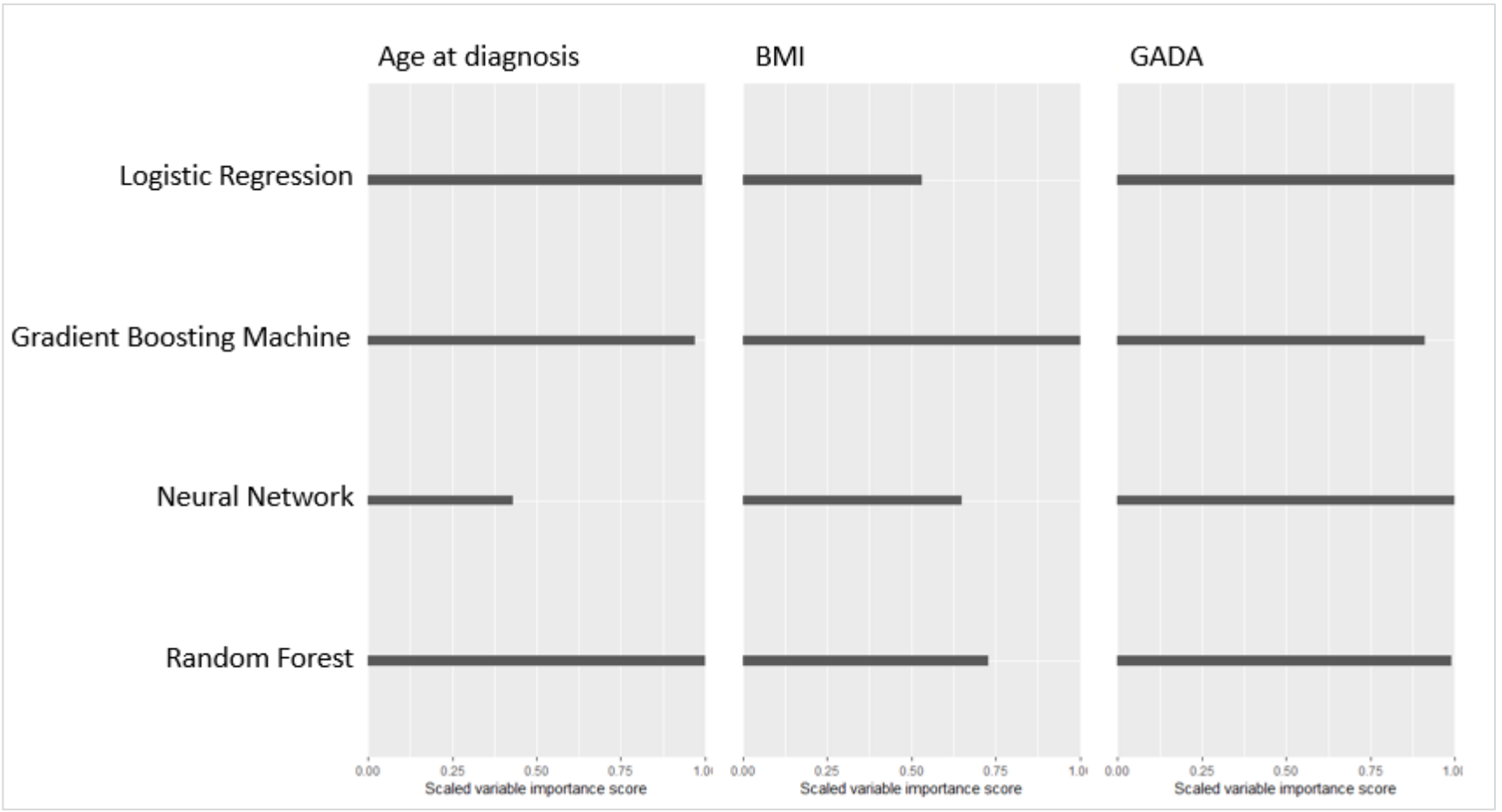
Supplementary Figure 1: Precision-Recall curves derived from test dataset.
Legend:
Solid lines: black = Support Vector Machine, dark grey = Logistic Regression, light grey = Random Forest
Dotted lines: black = Neural Network, dark grey = K-Nearest Neighbours, light grey = Gradient Boosting Machine



Supplementary Table 4: Correlation coefficient matrix of model predictions obtained from external test validation data

	Gradient Boosting Machine	Support Vector Machine	K-Nearest Neighbours	Neural Network	Random Forest	Logistic Regression
Gradient Boosting Machine	1.00					
Support Vector Machine	0.97	1.00				
K-Nearest Neighbours	0.94	0.96	1.00			
Neural Network	0.97	0.99	0.97	1.00		
Random Forest	0.94	0.93	0.90	0.93	1.00	
Logistic Regression	0.96	1.00	0.97	1.00	0.92	1.00

Supplementary Figure 2: Scaled variable importance by model



Supplementary material references

1. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models 2018 [Available from: https://CRAN.R-project.org/package=gbm.
2. Meyer D, Dimitriadou E, Hornik J, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien 2018 [Available from: https://CRAN.R-project.org/package=e1071.
3. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.
4. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

Chapter 4.

A Type 1 Diabetes Genetic Risk Score can identify patients with GAD65 autoantibody positive type 2 diabetes that rapidly progress to insulin therapy

Anita L Lynam , Timothy J McDonald, Femke Rutters, Louise A Donnelly, Andrew T Hattersley, Richard A Oram, Colin N A Palmer, Amber A van der Heijden, Fiona Carr, Petra J M Elders, Mike N Weedon, Roderick C Slieker, Leen M 't Hart, Ewan R Pearson, Beverley M Shields Angus G Jones

Diabetes Care, 2019 February 42(2)

Acknowledgments of co-authors and contributions to paper

I was instrumental in formulation of the study concept, research methods and design. I performed the review of existing literature. I combined and cleaned the data. I designed and performed the statistical analysis, and interpreted the results. I drafted the manuscript and responded to reviewer comments.

Timothy McDonald, Beverley Shields and Angus Jones conceived the idea. Louise Donnelly, Colin Palmer and Ewan Pearson researched the Genetics of Diabetes Audit and Research in Tayside Scotland (GoDarts) data. Femke Rutter, Amber van der Heijden, Petra Elders, Roderick Slieker and Leen M 't Hart researched the Hoorn Diabetes Care System (DCS) data. Timothy McDonald researched the pathology data. Mike Weedon and Richard Oram researched the genetic data. Angus Jones and Andrew Hattersley researched the Exeter cohort data. Beverley Shields, Timothy McDonald, Andrew Hattersley, Mike Weedon and Angus Jones assisted with data analysis and clinical interpretation. Beverley Shields and Angus Jones reviewed and gave feedback on the draft manuscripts and reviewer comments. All authors critically revised the manuscript and approved the final version.

Abstract

Objective

Progression to insulin therapy in clinically diagnosed type 2 diabetes is highly variable. The presence of GADA is associated with faster progression, but its predictive value is limited. We aimed to determine if a Type 1 Diabetes Genetic Risk Score (T1D GRS) could predict rapid progression to insulin treatment over and above GADA testing.

Research Design and Methods

We examined the relationship between T1D GRS, GADA (negative or positive) and rapid insulin requirement (within 5 years) using Kaplan-Meier survival analysis and Cox regression in 8,608 participants with clinical type 2 diabetes (onset >35 years, treated without insulin for ≥6 months). T1D GRS was analysed both continuously (as standardized scores) and categorized based on previously reported centiles of a type 1 diabetes population (<5th (low), 5th-50th (medium), >50th (high)).

Results

In GADA positive participants (3.3%), those with higher T1D GRS progressed to insulin more quickly: Probability of insulin requirement at five years [95% CI]: 47.9% [35.0%, 62.78%] (high T1D GRS) vs 27.6% [20.5%, 36.5%] (medium T1D GRS) vs 17.6% [11.2%, 27.2%] (low T1D GRS), p=0.001. In contrast T1D GRS did not predict rapid insulin requirement in GADA negative participants (p=0.4). In Cox regression analysis with adjustment for age of diagnosis, BMI and cohort, T1D GRS was independently associated with time to insulin only in the presence of GADA: hazard ratio per SD increase 1.48 [1.15, 1.90], p=0.002.

Conclusions

A Type 1 Diabetes Genetic Risk Score alters the clinical implications of a positive GADA test in patients with clinical type 2 diabetes, and is independent of and additive to clinical features.

Type 2 diabetes is a progressive disease due to a gradual reduction in the capacity of the pancreatic islet cells (beta cells) to produce insulin (1). The clinical course of this progression is highly variable with some patients progressing very rapidly to requiring insulin treatment, whilst others can be successfully treated with lifestyle changes or oral agents for many years (1; 2). Being able to identify patients likely to rapidly progress may have clinical utility in prioritization monitoring and treatment escalation, and in choice of therapy.

It has previously been shown that many patients with clinical features of type 2 diabetes have positive GADA and that the presence of this autoantibody is associated with faster progression to insulin (3; 4). This is often termed Latent Autoimmune Diabetes in Adults (LADA) (5; 6). However the predictive value of GADA testing is limited in a clinical type 2 diabetes population, with many GADA positive patients not requiring insulin treatment for many years (4; 7). Previous research has suggested that genetic variants in the Human Leukocyte Antigen (HLA) region associated with type 1 diabetes are associated with more rapid progression to insulin in patients with clinically defined type 2 diabetes and positive GADA (8).

We have recently developed a Type 1 Diabetes Genetic Risk Score (T1D GRS), which provides an inexpensive £56 in our local clinical laboratory, £16 where DNA has been previously extracted), integrated assessment of a person's genetic susceptibility to type 1 diabetes (9). The score is composed of 30 type 1 diabetes risk variants weighted for effect size, and aids discrimination of type 1 diabetes from type 2 diabetes. The T1D GRS has advantages over HLA typing alone, as it includes more genetic information, is cheaper than conventional HLA typing, and represents a continuous scale of likelihood of type 1 diabetes susceptibility. In young-onset adults (diagnosed between 20-40 years) it can

predict insulin dependence and is independent of and additive to islet-autoantibodies and clinical features (9). It is not known if the T1D GRS will improve the prediction of insulin requirement by GADA in clinically defined type 2 diabetes.

We aimed to determine if the T1D GRS could predict rapid progression to insulin (within 5 years of diagnosis) over and above GADA testing in patients with a clinical diagnosis of type 2 diabetes treated without insulin at diagnosis.

Methods

We examined the relationship between GADA, T1D GRS and progression to insulin therapy using survival analysis in 8,608 participants with clinical type 2 diabetes initially treated without insulin therapy.

Study population

Included participants had a clinical diagnosis of type 2 diabetes after the age of 35 years, and were treated without insulin for the first 6 months from diagnosis and were of white European origin.

To achieve a sufficient number of GADA positive participants, participants were identified in the following cohorts: Genetics of Diabetes Audit and Research Tayside Study (GoDARTS) (10), Hoorn Diabetes Care System (DCS) (11), Diabetes Alliance for Research in England (DARE) (12), Predicting Response to Incretin Based Agents in Type 2 Diabetes (PRIBA) (13), and MRC MASTERMIND Progressors (14) and combined into a single dataset. These cohorts were studies of participants with a clinical diagnosis of type 2 diabetes recruited from primary and secondary care, and are population based with the exception of PRIBA and MRC MASTERMIND Progressor which account for

<10% of participants. Summaries of the cohort recruitment and data collection methods are shown in Supplementary Table 1, a flow diagram of sample selection is shown in Supplementary Figure 1.

Participants known to have had GADA testing performed either in clinical practice or prior to diagnosis (through review of electronic laboratory records) were excluded due to the risk of the result influencing the clinician's treatment decision.

In the GoDarts cohort, participants diagnosed with diabetes before 1st January 1994 were excluded; due to insufficient prescribing information we were unable to define time to insulin prior to this date. In the DARE cohort, only the participants recruited in the Exeter Centre with saved serum were included.

Assessment of diabetes progression (time to insulin)

For GoDarts and DCS cohorts, time to insulin was defined from electronic prescription records. For Exeter Cohorts (DARE, PRIBA and MRC MASTERMIND Progressors), insulin treatment, date of commencing insulin and date of diagnosis were self-reported at a single visit.

Laboratory Measurement

The Academic Department of Blood Sciences at the Royal Devon and Exeter Hospital measured GADA for all five cohorts at a median diabetes duration of 6.1 years, using the same assay from biobanked samples stored at -80C. GADA was performed using the RSR Limited ELISA assay (RSR Ltd, Cardiff, UK) on the Dynex DS2 ELISA Robot (Dynex Technologics, Worthing, UK). The cut-off for positivity was ≥11 units/ml, based on the 97.5th centile of 1,500 controls without diabetes (15). The lowest reportable value (lowest calibrant)

was 5.0 units/ml. The laboratory participates in the International Autoantibody Standardization Programme.

The HbA_{1c} value at latest follow up (closest available result, median 10.6 years diabetes duration) was obtained from electronic healthcare records or measured on a research sample by the Academic Department of Blood Sciences at the Royal Devon and Exeter Hospital.

Assessment of T1D GRS

The development of the T1D GRS has been described previously (9). In brief, T1D GRS consists of 30 common type 1 diabetes genetic variants (single nucleotide polymorphisms (SNPs)) from HLA and non-HLA loci; each variant is weighted by their effect size on type 1 diabetes risk from previously published literature, with weights for DR3/DR4-DQ8 assigned based on imputed haplotypes. The combined score represents an individual's genetic susceptibility to type 1 diabetes. Variants used to derive the score are shown in Supplementary Table 2. For ease of clinical interpretation the score is presented in this article as the centile position of the distribution in the Wellcome Trust Case Control Consortium type 1 diabetes population (16).

In the Exeter cohorts, genotyping was performed using the KASP genotyping assay by LGC Genomics (Hoddesdon, UK) as previously described (9). Genotyping in the GoDarts cohort was performed using custom genotyping arrays (including Immunochip, Cardio-Metabochip (Metabochip) and Human Exome array) from Illumina as previously described (17). Genotyping in the DCS cohort was performed with Illumina's HumanCoreExome Array and imputed using IMPUTE2 (18) into the 1000 Genomes March 2012 reference panel. All SNPs had an INFO > 0.8.

T1D GRS calculation was not performed if genotyping results were missing for either of the two alleles with the greatest weighting (DR3/DR4-DQ8 or HLA_DRB1_15) or if more than two of any other SNPs were missing.

Statistical analysis

We assessed the relationship between time to insulin treatment and each of GADA and T1D GRS using survival analysis. For this analysis, T1D GRS was categorized based on centiles of a type 1 diabetes population (Wellcome Trust Case Control Consortium (16)): <5th centile (< 0.234 (low)), 5th-50th centile (>= 0.234 & <= 0.280 (medium)), >50th centile (> 0.280 (high)) as previously reported (9; 19). GADA was dichotomized into negative or positive based on the cut-off for positivity. Participants were then classified into six risk groups from these categories 1) GADA negative, low T1D GRS 2) GADA negative, medium T1D GRS 3) GADA negative, high T1D GRS 4) GADA positive, low T1D GRS 5) GADA positive, medium T1D GRS 6) GADA positive, high T1D GRS.

Time to insulin data was censored at five years (or the latest available time point not on insulin, if earlier). Survival distributions for time to insulin, stratified by risk groups, were estimated using the Kaplan-Meier product limit estimator (20). The proportional hazard assumption was checked visually and failed. Differences in time to insulin between risk groups were therefore compared using the Wilcoxon (Breslow) test. Positive predicted values were obtained from the product limit estimator which makes allowances for censored observations.

To assess whether clinical characteristics were different across risk groups we performed Wilcoxon test for trend (21) on the continuous variables and Pearson chi-squared test for categorical variables.

To assess whether GADA, T1D GRS (as a continuous covariate), age of diagnosis and BMI (closest available to diagnosis, median 3 years diabetes duration) are independent predictors of rapid progression to insulin we performed multivariate Cox proportional hazards regression analysis (22). When T1D GRS was used as a continuous covariate, the proportional hazard assumption was satisfied. T1D GRS and GADA were added in as separate variables and as an interaction term. The log-linearity assumption was checked by examining Martingale-based residual plots and was considered valid. Study of origin was included as a strata variable to control for effects of cohort differences.

As a 10 SNP T1D GRS combining the 10 alleles with the greatest weightings ordered by published odds ratios (Supplementary Table 3) has also been proposed for clinical practice, we repeated survival analysis using T1D GRS defined by this 10 SNP score using the same centile cut-offs for categorization (9). We also estimated survival distributions for risk groups based on imputed HLA DR3/DR4 genotypes, individually and grouped by number of copies of at risk alleles.

Median follow-up time was calculated using the reverse Kaplan-Meier method (23). All analysis was performed in Stata/SE 15.1 (StataCorp, College Station, TX).

Results

We identified 8,608 participants with a clinical diagnosis of type 2 diabetes meeting all of our inclusion criteria, Table 1 shows the characteristics for these participants. 79.9% (n = 6,879) had been followed for at least five years; median follow up time, calculated as the median time to censoring (insulin treatment or

latest follow up), was 10.5 [95% CI 10.3, 10.6] years. 7.8% (n = 533) of those participants with over five years follow up had progressed to insulin <= 5 years. 3.3% (n = 280) of participants were GADA positive (measured at a median 6.1 years diabetes duration). The distribution of participants by low, medium and high TID GRS category was 53.2% (n = 4,580), 40.7% (n = 3,504) and 6.1% (n = 524) respectively.

Table 1: Participant characteristics. Median (IQR) or % (n = 8 608). *Closest to diagnosis (median 3 years diabetes duration). †Percentage of participants observed for at least five years. ‡At latest follow up. §Centile of participants with type 1 diabetes from the Wellcome Trust Case Control Consortium.

Characteristic	Value
Sex (% Male)	56.4%
Age at diagnosis (years)	60 (52, 68)
BMI (kg/m^2)*	30.4 (27.2, 34.7)
Duration of diabetes (years) at latest follow up	10.6 (6.0, 14.3)
Duration of diabetes (years) at GADA	6.1 (3.3, 10.0)
Insulin treated within 5 years (%)†	7.8%
HbA_{1c} (%)‡	7.0 (6.4, 8.0)
HbA_{1c} (mmol/mol)‡	53 (46, 64)
GADA positive (%)	3.3%
T1D GRS centile§	4.2 (0.6, 16.1)

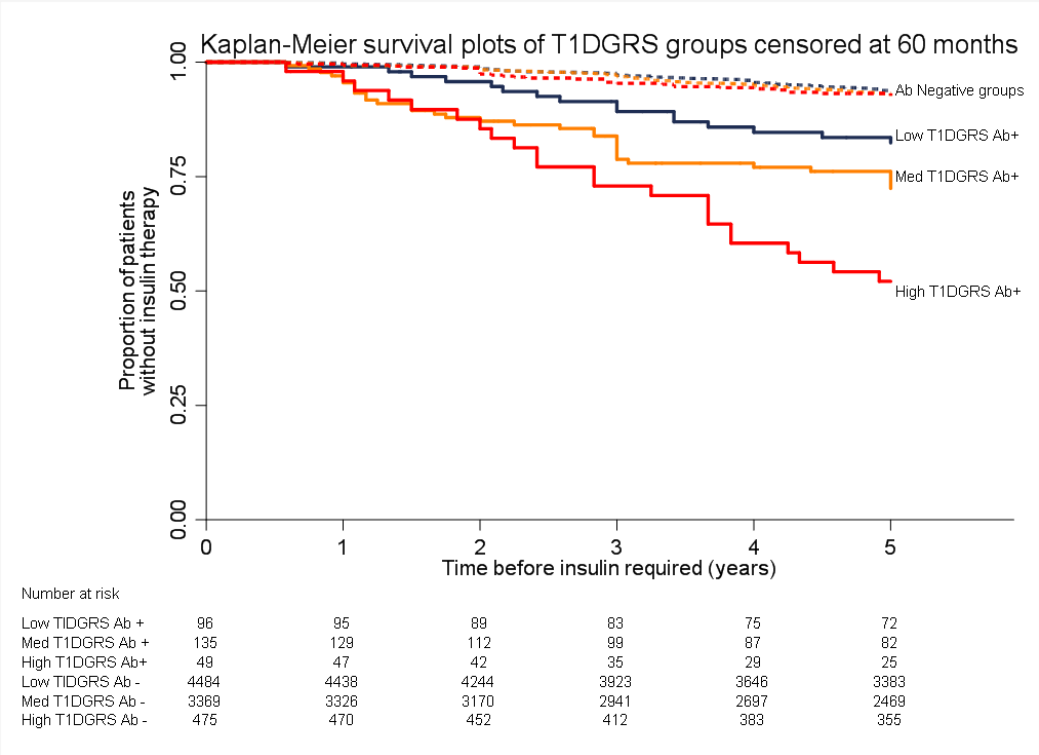
Characteristics of the participants stratified by the individual cohorts are shown in Supplementary Table 4. Statistically significant differences in GADA prevalence, diabetes duration and HbA_{1c} between cohorts were evident and survival distributions differed between studies. These cohort differences were adjusted for by including study of origin as a strata variable in the Cox proportional hazards regression analysis. Year of diagnosis for the participants ranged over a fairly long period (from 1967 to 2015) over which management and treatment practices are likely to have changed. Supplementary Figure 2 shows a reasonable distribution between 1994 and 2015 but with a long tail back to 1967.

High T1D GRS is associated with markedly higher rates of rapid insulin requirement in participants with positive GADA, but is not associated in those who are GADA negative

T1D GRS was strongly predictive of rapid insulin requirement in participants with positive GADA (Figure 1). In GADA positive participants, those with higher T1D GRS progressed to insulin more quickly (p=0.001): probability of requiring insulin at five years post diagnosis (positive predictive value) [95% CI]: 47.9% [35.0%, 62.78%] (high T1D GRS) vs 27.6% [20.5%, 36.5%] (medium T1D GRS) vs 17.6% [11.2%, 27.2%] (low T1D GRS).

T1D GRS was not associated with rapid insulin requirement in GADA negative participants. For the GADA negative participants, the probability of requiring insulin at five years post diagnosis was similar across all risk groups (p=0.4): 7.4% [5.3%, 10.3%] (high T1D GRS) vs 7.3% [6.5%, 8.3%] (medium T1D GRS) vs 6.7% [5.9%, 7.5%] (low T1D GRS).

Figure 1: Kaplan-Meier plot of probability of requiring insulin therapy during 5-year follow-up by risk group of T1D GRS. Solid lines represent GADA positive groups, dashed lines represent GADA negative groups. Blue = low T1D GRS (<5th centile of a type 1 diabetes population (< 0.234)), orange = medium T1D GRS (5th-50th centile of a type 1 diabetes population (>= 0.234 & <= 0.280)), red =high T1D GRS (>50th centile of a type 1 diabetes population (> 0.280)).



Differences in T1D GRS were associated with higher HbA_{1c} in GADA positive participants but no differences in other clinical features

The characteristics of the GADA positive and negative participants split by T1D GRS category are shown in Table 2. In GADA positive participants, HbA_{1c} increased ($p = 0.04$) and BMI decreased ($p = 0.01$) with higher T1D GRS category. In GADA negative participants, clinical characteristics were similar across all categories of T1D GRS.

When comparing the characteristics of GADA positive and negative participants (Table 2), GADA positive participants had a higher T1D GRS (median 0.251 vs 0.231, $p < 0.001$) and a lower BMI (median 28.73 vs 30.48, $p < 0.001$) but similar age of diagnosis (median 59 vs 60 years, $p = 0.052$).

Table 2: Participant characteristics by risk group. Median (IQR) or %. p values given for continuous variables are Wilcoxon-type test for trend, for categorical variables Pearson chi-squared. *Closest to diagnosis (median 3 years diabetes duration). †At latest follow up. ‡Centile of participants with type 1 diabetes from the Wellcome Trust Case Control Consortium.

	<5th T1D GRS centile for type 1 diabetes‡ (low)	5th–50th T1D GRS centile for type 1 diabetes ‡ (medium)	>50th T1D GRS centile for type 1 diabetes ‡ (high)	p-value
GADA negative				
n (% of GADA negative)	4,484 (54%)	3,369 (40%)	475 (6%)	
Sex (% Male)	56.2%	57.0%	56.2%	>0.1
Age at diagnosis (years)	60 (52, 68)	60 (52, 68)	60 (51, 68)	>0.1
BMI (kg/m^2)*	30.5 (27.3, 34.9)	30.4 (27.1, 34.6)	30.7 (27.4, 34.3)	>0.1
Duration of diabetes (years) at latest follow up	10.6 (6.1, 14.4)	10.5 (5.8, 14.1)	10.6 (6.0, 14.4)	>0.1
Duration of diabetes (years) at GADA	6.3 (3.3, 10.0)	6.0 (3.3, 10.0)	6.3 (3.6, 10.2)	>0.1
HbA_{1c} (%)†	7.0 (6.4, 8.0)	7.0 (6.4, 8.0)	6.9 (6.4, 8.0)	>0.1
HbA_{1c} (mmol/mol) †	53 (46, 64)	53 (46, 64)	52 (46, 64)	>0.1
Insulin treated within 5 years (%) (where observed ≥ five years)	6.7%	7.4%	8.1%	>0.1
GADA (units/mL)	4.9 (4.9, 5.0)	4.9 (4.9, 5.0)	4.9 (4.9, 5.0)	>0.1
GADA positive				
n (% of GADA positive)	96 (34%)	135 (48%)	49 (18%)	
Sex (% Male)	51.0%	57.0%	49.0%	>0.1
Age at diagnosis (years)	61 (50, 69)	59 (51, 67)	54 (49, 63)	0.06
BMI (kg/m^2)*	29.6 (26.7, 34.1)	28.7 (25.6, 32.5)	27.7 (25.4, 30.4)	0.01
Duration of diabetes (years) at latest follow up	11.1 (9.0., 13.8)	10.4 (6.7, 14.9)	11.8 (9.1, 15.0)	>0.1
Duration of diabetes (years) at GADA	5.2 (3.1, 9.5)	5.6 (3.0, 10.1)	8.9 (4.9, 11.1)	0.01
HbA_{1c} (%)†	7.3 (6.6, 9.1)	7.8 (6.7, 9.0)	8.1 (7.1, 9.1)	0.04
HbA_{1c} (mmol/mol) †	56 (49, 76)	62 (50, 75)	66 (55, 77)	0.04
Insulin treated within 5 years (%) (where observed ≥ five years)	18.4%	27.8%	40.5%	0.03
GADA (units/mL)	77.6 (24.3, 1191.9)	111.4 (28.8, 1354.9)	175.9 (38.6, 1218.2)	>0.1

T1D GRS and GADA are predictors of rapid insulin requirement and are independent of age and BMI

Table 3 shows the Cox proportional hazards regression model for time to insulin (censored at 5 years) controlled for effects of cohort differences. As expected, the presence of GADA was a significant predictor of time to insulin (Hazard Ratio (HR) 3.43 [2.50, 4.71], p <0.001). T1D GRS was independently associated with time to insulin, but only in the presence of GADA (HR per 1 standard deviation (SD) increase in T1D GRS 1.48 [1.15, 1.90], p = 0.002). These associations were independent of age at diagnosis and BMI.

Table 3: Hazard ratios from Cox proportional regression model (adjusted for cohort) for time to insulin censored at 5 years (30 SNP T1D GRS).* Closest to diagnosis

Variable	Hazard Ratio [95% CI]	p value
GADA negative	1	
GADA positive	3.43 [2.50, 4.71]	<0.001
GADA negative:T1D GRS (per 1 SD increase in T1D GRS)	1.02 [0.94, 1.12]	>0.1
GADA positive:T1D GRS (per 1 SD increase in T1D GRS)	1.48 [1.15, 1.90]	0.002
Age at diagnosis (per 1 year)	0.97 [0.96, 0.97]	<0.001
BMI (per kg/m^2 unit)*	1.00 [0.98, 1.01]	>0.1

A 10 SNP T1D GRS, and HLA type alone are predictive of future insulin requirement

The association between the 10 SNP T1D GRS and rapid insulin requirement was consistent with our findings using the full 30 SNP T1D GRS. The 10 SNP T1D GRS was associated with rapid insulin requirement in the GADA positive risk groups (p < 0.001) but was not associated in the GADA negative groups (p=0.4) (Supplementary Figure 3). In Cox proportional hazards regression model (Supplementary Table 5), the 10 SNP T1D GRS was independently associated with future insulin treatment in GADA positive participants (HR per 1

SD increase in T1D GRS 1.34 [1.05, 1.71], p = 0.02). Kaplan-Meier plots for HLA DR3/DR4 genotype risk groups, individually and grouped by number of at risk alleles, are shown in Supplementary Figures 4 and 5.

Conclusions

In this large study of participants with a clinical diagnosis of type 2 diabetes, we have found that type 1 genetic susceptibility alters the clinical implications of a positive GADA when predicting rapid time to insulin. GADA positive participants with high T1D GRS were more likely to require insulin within 5 years of diagnosis, with 48% progressing to insulin in this time in contrast to only 18% in participants with low T1D GRS. The T1D GRS was independent of and additive to participant's age of diagnosis and BMI. However, T1D GRS was not associated with rapid insulin requirement in participants who were GADA negative.

To our knowledge this is the first study to assess the association between an integrated assessment of type 1 genetic risk and GADA in patients with type 2 diabetes or LADA. A key strength of this study is use of large, predominantly population-based, cohorts of participants diagnosed with type 2 diabetes and to date, is the largest cohort with measured GADA in a western population. This means our results are likely to reflect true associations in patients seen in clinical practice. An additional key strength is the use of a single laboratory and assay for measuring GADA across cohorts, with a very robustly defined threshold for positive GADA based on a large predominantly adult control population. We have demonstrated that our results are independent of and additive to participants' clinical features.

A limitation of our study is that time to insulin has been self-reported in the Exeter cohorts at a single visit, in contrast to other cohorts where electronic healthcare records were available. Insulin commencement was also based on clinical decision making rather than a trial protocol. Both these aspects may introduce imprecision but since both clinicians and participants were unaware of results, systematic bias would be unlikely. An additional limitation of cross-sectional study design is that GADA was measured at a median 6.1 years diabetes duration, which could result in a lower prevalence than if measurement was undertaken at diagnosis. However, in adult populations the difference is likely to be small, with GADA positivity being stable over the first 6 years in UKPDS study participants (adult onset type 2 diabetes) (24) and a modest reduction in prevalence (72% to 63%) observed after 8 years in adult onset type 1 diabetes (25). The results of this study can only be applied to white European populations and we do not have measurement of other islet-autoantibodies in this cohort - the interaction between genetic risk and other islet-autoantibodies would be an area of interest for future research (26).

Our findings are consistent with previous research in a population of participants diagnosed with diabetes between the ages of 20 to 40 years, where the same T1D GRS was predictive of insulin dependent diabetes (9), and other work which has shown this risk score to be additive to islet-autoantibodies in predicting future type 1 diabetes (27). It is also consistent with previous research showing patients defined as LADA who have HLA type associated with type 1 diabetes susceptibility, have more rapid progression to insulin (8), and with research showing a combination of positive islet cell autoantibodies and high risk HLA is associated with low C-peptide in a cohort diagnosed as type 2 diabetes in contrast to either of these features alone (28). While the

relationship between integrated genetic risk of type 1 diabetes and progression of type 2 diabetes or LADA has not been previously assessed, it has previously been shown that a type 2 diabetes genetic risk score covering 61 established type 2 diabetes risk variants is not associated with time to insulin (17) and that a 69 SNP type 2 diabetes genetic risk score has very limited utility in discriminating patients with type 1 from type 2 diabetes (9).

The prevalence of positive GADA in our cohorts was lower than in much of the previous literature, with previous multicentre studies reporting widely varying prevalence of positive GADA in type 2 diabetes populations ranging from 4% to 14% (29; 30). In addition to diabetes duration, differences in the prevalence of GADA positivity between our and other studies may be explained by our use of an assay with higher specificity than used in many other studies (29-33), our lack of an upper age limit (with lower GADA prevalence seen at older ages (4; 33; 34)), and our use of predominantly population cohorts not selected from secondary care where treatment with insulin is more frequent. We have used a robustly defined high specificity (97.5%) threshold to define positive GADA in line with current clinical laboratory practice, using a large control population. Detectable GADA are commonly found in healthy adult non-diabetic populations and therefore a threshold based on a control population is recommended to robustly define GADA positivity (31-33). An additional potential reason for low autoantibody prevalence is that we have excluded a small number of cohort participants who had GADA tested in clinical practice, which may have influenced treatment choice. However only 47 participants were excluded of whom only 13 were GADA positive, so the effect on overall prevalence is small.

Our findings have clear implications for clinical practice. The T1D GRS represents a novel clinical test that can be used to enhance the prognostic

value of GADA testing. For predicting future insulin requirement in patients with apparent type 2 diabetes who are GADA positive, T1D GRS may be clinically useful and can be used as an additional test in the screening process. However, in patients with type 2 diabetes who are GADA negative, there is no benefit gained from genetic testing. This is unsurprising as the prevalence of underlying autoimmunity in patients with a clinical phenotype of type 2 diabetes who are GADA negative is likely to be extremely low, therefore most GADA negative participants with high T1D GRS will have non-autoimmune diabetes. The use of this two-step testing approach may facilitate a precision medicine approach to patients with apparent type 2 diabetes; patients who are likely to progress rapidly are identified for targeted management which may include increased monitoring, early therapy intensification and/or interventions aimed at slowing progression (35; 36).

The costs of analysing the T1D GRS are relatively modest and may fall further as genetic testing is rapidly becoming less expensive (37). This may allow introduction of the T1D GRS into clinical practice. While the test cost could potentially be reduced further by using 10 SNPs or imputing HLA type alone, the majority of test costs are attributable to DNA extraction, sample handling and test interpretation, with cost for genotyping additional SNPs as low as 8 pence per SNP. Savings would therefore be modest and, while this study does not have sufficient statistical power to directly compare different risk scores in islet-autoantibody-positive participants, this may come at a cost of reduced test accuracy. The use of a risk score approach has an additional advantage over using HLA alone, as it provides genetic information expressed as a simple to use continuous variable.

While using the T1D GRS in a two stage approach may have clinical utility, approaches that go beyond single tests and thresholds to integrate multiple clinical features and biomarkers are likely to have the greatest use for clinical practice. The T1D GRS is additive to other predictive features such as age of diagnosis and BMI, and dichotomizing the test to use thresholds will lose predictive value. While the negative predictive value of a low T1D GRS in participants with GADA is high (<5th centile 92%), positive predictive values are modest, with the majority of high T1D GRS participants not requiring insulin by 5 years. Therefore approaches that combine different predictive features on a continuous basis, using prediction models (clinical calculators), may have the greatest utility in accurately predicting future insulin requirement in this group, and are an important area for future research (38). Additional areas for future research include the association between T1D GRS and progression where multiple islet-autoantibodies have been tested, and assessment in a prospective setting where islet-autoantibodies have been measured at diabetes diagnosis.

In conclusion, a Type 1 Diabetes Genetic Risk Score alters the clinical implications of a positive GADA test in patients with clinical type 2 diabetes, and is independent of and additive to clinical features. This therefore represents a novel test for identifying patients with rapid progression in this population.

Acknowledgments

The authors thank participants who took part in these studies and the research teams who undertook cohort recruitment. We thank Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital for assistance with conducting the study.

GoDarts: We are grateful to all the participants in this study, the general practitioners, the Scottish School of Primary Care for their help in recruiting the participants, and to the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The study complies with the Declaration of Helsinki. We acknowledge the support of the Health Informatics Centre, University of Dundee for managing and supplying the anonymized data and NHS Tayside, the original data owner.

References

1. U.K. Prospective Diabetes Study Group: U.K. Prospective Diabetes Study 16: Overview of 6 Years' therapy of Type II Diabetes: A Progressive Disease. Diabetes 1995;44:1249
2. Fonseca VA: Defining and Characterizing the Progression of Type 2 Diabetes. Diabetes care 2009;32:S151
3. Groop LC, Bottazzo GF, Doniach D: Islet Cell Antibodies Identify Latent Type I Diabetes in Patients Aged 35–75 Years at Diagnosis. Diabetes 1986;35:237
4. Turner R, Stratton I, Horton V, Manley S, Zimmet P, Mackay IR, Shattock M, Bottazzo GF, Holman R: UKPDS 25: autoantibodies to islet-cell cytoplasm and glutamic acid decarboxylase for prediction of insulin requirement in type 2 diabetes. The Lancet 350:1288-1293
5. Tuomi T, Groop LC, Zimmet PZ, Rowley MJ, Knowles W, Mackay IR: Antibodies to Glutamic Acid Decarboxylase Reveal Latent Autoimmune Diabetes Mellitus in Adults With a Non—Insulin-Dependent Onset of Disease. Diabetes 1993;42:359
6. Pozzilli P, Di Mario U: Autoimmune Diabetes Not Requiring Insulin at Diagnosis (Latent Autoimmune Diabetes of the Adult). Diabetes care 2001;24:1460
7. Liu L, Li X, Xiang Y, Huang G, Lin J, Yang L, Zhao Y, Yang Z, Hou C, Li Y, Liu J, Zhu D, Leslie RD, Wang X, Zhou Z: Latent autoimmune diabetes in adults with low-titer GAD antibodies: similar disease progression with type 2 diabetes: a nationwide, multicenter prospective study (LADA China Study 3). Diabetes care 2015;38:16-21
8. Maioli M, Pes GM, Delitala G, Puddu L, Falorni A, Tolu F, Lampis R, Orru V, Secchi G, Cicalo AM, Floris R, Madau GF, Pilosu RM, Whalen M, Cucca F: Number of autoantibodies and HLA genotype, more than high titers of glutamic acid decarboxylase autoantibodies, predict insulin dependence in latent autoimmune diabetes of adults. European journal of endocrinology 2010;163:541-549
9. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN: A Type 1 diabetes genetic risk score can aid discrimination between Type 1 and Type 2 diabetes in young adults. Diabetes care 2016;39:337-344
10. Hebert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, Donnelly LA, Tavendale R, Leese G, Colhoun HM, Dow E, Morris AD, Doney AS, Lang CC, Pearson ER, Smith BH, Palmer CNA: Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). International journal of epidemiology 2017;
11. van der Heijden AA, Rauh SP, Dekker JM, Beulens JW, Elders P, t Hart LM, Rutters F, van Leeuwen N, Nijpels G: The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. BMJ open 2017;7:e015599
12. Diabetes Alliance for Research in England (DARE) [article online], Available from http://www.diabetesgenes.org/content/diabetes-alliance-research-england-dare-previously-known-exeter-research-alliance-extra-stud. Accessed 23 November 2017
13. Jones AG, McDonald TJ, Shields BM, Hill AV, Hyde CJ, Knight BA, Hattersley AT, for the PSG: Markers of beta cell failure predict poor glycemic response to GLP-1 receptor agonist therapy in type 2 diabetes. Diabetes care 2016;39:250-257

14. RetroMASTER - Retrospective Cohort MRC ABPI STratification and Extreme Response Mechanism in Diabetes [article online], Available from https://www.clinicaltrials.gov/ct2/show/NCT02109978.
15. McDonald TJ, Colclough K, Brown R, Shields B, Shepherd M, Bingley P, Williams A, Hattersley AT, Ellard S: Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabetic medicine : a journal of the British Diabetic Association 2011;28:1028-1033
16. The Wellcome Trust Case Control C: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-678
17. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CNA, Pearson ER: Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. Diabetes care 2014;37:718
18. Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 2009;5:e1000529
19. Patel KA, Oram RA, Flanagan SE, De Franco E, Colclough K, shepherd M, Ellard S, Weedon MN, Hattersley AT: Type 1 Diabetes Genetic Risk Score: a novel tool to discriminate monogenic and type 1 diabetes. Diabetes 2016;65:2094-2099
20. Kaplan EL, Meier P: Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 1958;53:457-481
21. Cuzick J: A wilcoxon-type test for trend. Statistics in medicine 1985;4:87-90
22. Cox D: Regression models and life tables. Journal of the Royal Statistical Society, Series B (Methodological) 1972;34:187 - 220
23. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials 1996;17:343-346
24. Desai M, Cull CA, Horton VA, Christie MR, Bonifacio E, Lampasona V, Bingley PJ, Levy JC, Mackay IR, Zimmet P, Holman RR, Clark A: GAD autoantibodies and epitope reactivities persist after diagnosis in latent autoimmune diabetes in adults but do not predict disease progression: UKPDS 77. Diabetologia 2007;50:2052-2060
25. Schölin A, Björklund L, Borg H, Arnqvist H, Björk E, Blohmé G, Bolinder J, Eriksson JW, Gudbjörnsdottir S, Nyström L, Östman J, Karlsson AF, Sundkvist G: Islet antibodies and remaining β-cell function 8 years after diagnosis of diabetes in young adults: a prospective follow-up of the nationwide Diabetes Incidence Study in Sweden. Journal of Internal Medicine 2004;255:384-391
26. Bottazzo GF, Bosi E, Cull CA, Bonifacio E, Locatelli M, Zimmet P, Mackay IR, Holman RR: IA-2 antibody prevalence and risk assessment of early insulin requirement in subjects presenting with type 2 diabetes (UKPDS 71). Diabetologia 2005;48:703-708
27. Redondo MJ, Geyer S, Steck AS, Sharp S, Wentworth J, Weedon M, Antinozzi P, Pugliese A, Oram R: A type 1 diabetes genetic risk score predicts progression of islet autoimmunity and development of type 1 diabetes in individuals at risk (Abstract) 53rd EASD Annual Meeting of the European Association for the Study of Diabetes: 51. Diabetologia 2017;60:1-608
28. Groop L, Miettinen A, Groop P-H, Meri S, Koskimies S, Bottazzo GF: Organ-Specific Autoimmunity and HLA-DR Antigens as Markers for β-Cell Destruction in Patients With Type II Diabetes. Diabetes 1988;37:99-103
29. Buzzetti R, Zampetti S, Maddaloni E: Adult-onset autoimmune diabetes: current knowledge and implications for management. Nature Reviews Endocrinology 2017;13:674

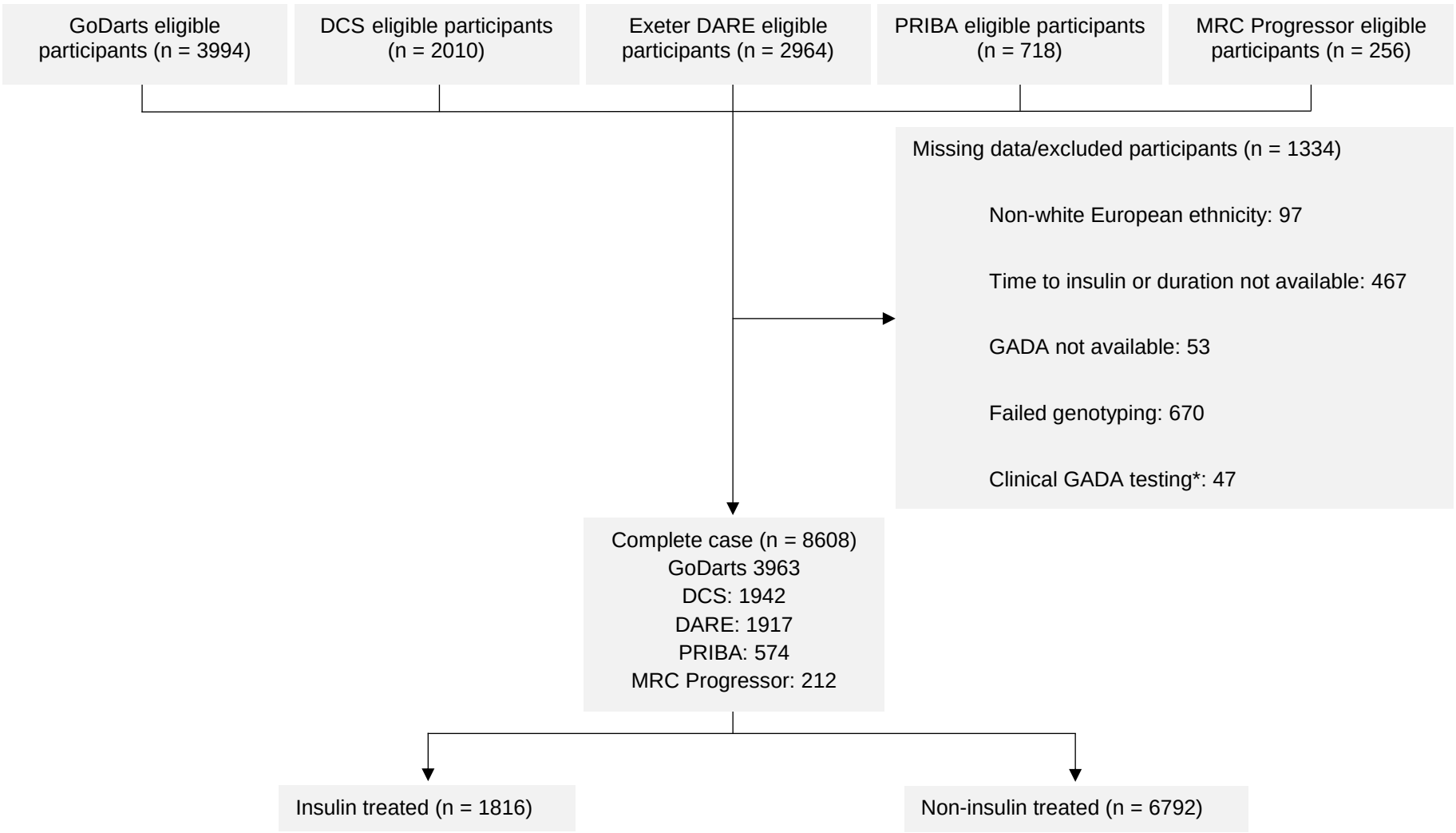
30. Laugesen E, Østergaard JA, Leslie RDG, the Danish Diabetes Academy W, Workshop S: Latent autoimmune diabetes of the adult: current knowledge and uncertainty. Diabetic Medicine 2015;32:843-852
31. Bonifacio E, Yu L, Williams AK, Eisenbarth GS, Bingley PJ, Marcovina SM, Adler K, Ziegler AG, Mueller PW, Schatz DA, Krischer JP, Steffes MW, Akolkar B: Harmonization of Glutamic Acid Decarboxylase and Islet Antigen-2 Autoantibody Assays for National Institute of Diabetes and Digestive and Kidney Diseases Consortia. The Journal of clinical endocrinology and metabolism 2010;95:3360-3367
32. Bingley PJ: Clinical applications of diabetes antibody testing. The Journal of clinical endocrinology and metabolism 2010;95:25-33
33. Tuomi T, Carlsson A, Li H, Isomaa B, Miettinen A, Nilsson A, Nissén M, Ehrnström BO, Forsén B, Snickars B, Lahti K, Forsblom C, Saloranta C, Taskinen MR, Groop LC: Clinical and genetic characteristics of type 2 diabetes with and without GAD antibodies. Diabetes 1999;48:150
34. Tuomi T, Santoro N, Caprio S, Cai M, Weng J, Groop L: The many faces of diabetes: a disease with increasing heterogeneity. The Lancet 2014;383:1084-1094
35. Florez JC: Precision Medicine in Diabetes: Is It Time? Diabetes care 2016;39:1085-1088
36. Leslie RD, Palmer J, Schloot NC, Lernmark A: Diabetes at the crossroads: relevance of disease classification to pathophysiology and treatment. Diabetologia 2016;59:13-20
37. Christensen KD, Dukhovny D, Siebert U, Green RC: Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. Journal of Personalized Medicine 2015;5:470-486
38. Hattersley AT, Patel KA: Precision diabetes: learning from monogenic diabetes. Diabetologia 2017;60:769-777

Supplementary Material

Supplementary Table 1: Cohort recruitment and data collection methods summary

	GoDarts	DCS	DARE	PRIBA	MRC Progressor
Included participants	3963	1942	1917	574	212
Data collection period	From 1998	From 1998	2007 to 2017	2011 to 2013	2013 to 2015
Study design	Longitudinal	Longitudinal	Cross-sectional	Longitudinal	Cross-sectional
Setting	Primary and secondary care in Tayside, Scotland	Primary and secondary care in West-Friesland, Netherlands	Primary and secondary care in eight diabetes research regions, England and retinal screening clinics.	Primary and secondary care in South West England	Primary and secondary care in Exeter, Dundee and Oxford, England
Inclusion criteria	Clinical diagnosis of type 2 diabetes.	Clinical diagnosis of type 2 diabetes.	Clinical diagnosis of diabetes (any type).	Clinical diagnosis of type 2 diabetes. Clinician determined requirement for DPP-IV inhibitor or GLP-1 analogue (HbA_{1C} >7.5%)	Clinical diagnosis of type 2 diabetes non-insulin treated within 6 months of diagnosis. Participants were selected on the basis of rapid or slow progression to insulin therapy (<7, >7 years). Age 18-90 inclusive.
Data collection	Clinical measurements and blood collected at initial visit. Follow up clinical data constantly updated using electronic medical record linkage.	Clinical measurements collected at initial visit, and repeated annually. Blood collected at one of the annual visits. Additional health data collected using electronic medical record linkage.	Clinical measurements and blood sample collected at visit. Ongoing biochemical data collected from pathology laboratories.	Clinical measurements and blood taken at initial visit. Follow up clinical measurements and blood collected at three and six months.	Clinical measures and fasting blood sample taken at visit.

Supplementary Figure 1: Participant flow diagram * identified through search of electronic laboratory records.



Supplementary Table 2: Type 1 diabetes SNPs included in the genetic risk score with weights. Effect allele is the risk increasing allele on the positive strand.

SNP	Gene	Odds Ratio	Weight	Effect Allele
rs2187668, rs7454108	DR3/DR4	48.18	3.87	
	DR3/DR3	21.12	3.05	
	DR4/DR4	21.98	3.09	
	DR4/X	7.03	1.95	
	DR3/X	4.53	1.51	
rs1264813	HLA_A_24	1.54	0.43	T
rs2395029	HLA_B_5701	2.5	0.92	T
rs3129889	HLA_DRB1_15	14.88	2.70	A
rs2476601	PTPN22	1.96	0.67	A
rs689	INS	1.75	0.56	T
rs12722495	IL2RA	1.58	0.46	T
rs2292239	ERBB3	1.35	0.30	T
rs10509540	C10orf59	1.33	0.29	T
rs4948088	COBL	1.3	0.26	C
rs7202877		1.28	0.25	G
rs12708716	CLEC16A	1.23	0.21	A
rs3087243	CTLA4	1.22	0.20	G
rs1893217	PTPN2	1.2	0.18	G
rs11594656	IL2RA	1.19	0.17	T
rs3024505	IL10	1.19	0.17	G
rs9388489	C6orf173	1.17	0.16	G
rs1465788		1.16	0.15	C
rs1990760	IFIH1	1.16	0.15	T
rs3825932	CTSH	1.16	0.15	C
rs425105		1.16	0.15	T
rs763361	CD226	1.16	0.15	T
rs4788084	IL27	1.16	0.15	C
rs17574546		1.14	0.13	C
rs11755527	BACH2	1.13	0.12	G
rs3788013	UBASH3A	1.13	0.12	A
rs2069762	IL2	1.12	0.11	A
rs2281808		1.11	0.10	C
rs5753037		1.1	0.10	T

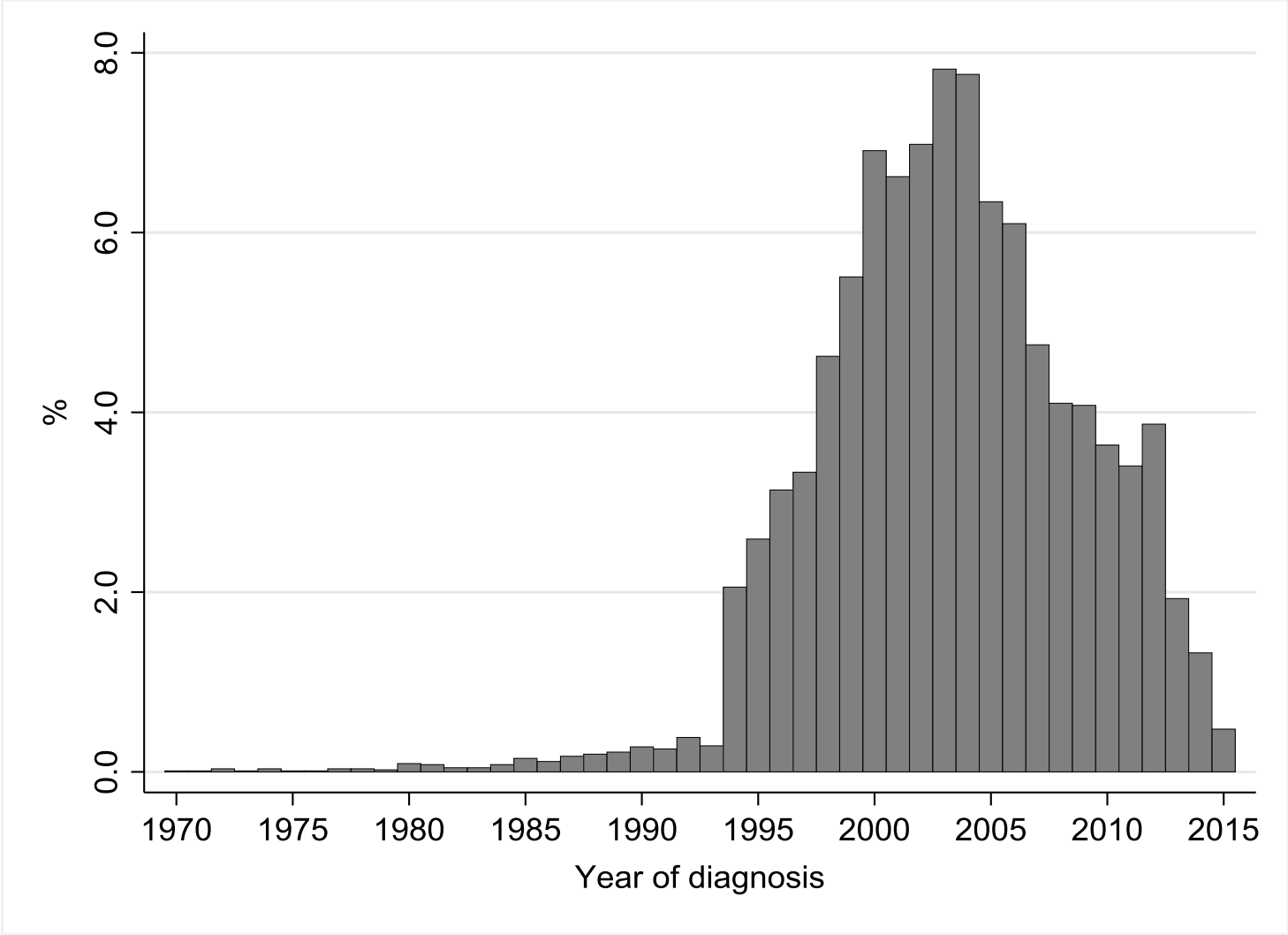
Supplementary Table 3: Type 1 diabetes SNPs included in the 10 SNP T1D GRS

SNP	Gene	Odds Ratio	Weight	Effect Allele
rs2187668, rs7454108	DR3/DR4	48.18	3.87	
	DR3/DR3	21.12	3.05	
	DR4/DR4	21.98	3.09	
	DR4/X	7.03	1.95	
	DR3/X	4.53	1.51	
rs1264813	HLA_A_24	1.54	0.43	T
rs2395029	HLA_B_5701	2.5	0.92	T
rs3129889	HLA_DRB1_15	14.88	2.70	A
rs2476601	PTPN22	1.96	0.67	A
rs689	INS	1.75	0.56	T
rs12722495	IL2RA	1.58	0.46	T
rs2292239	ERBB3	1.35	0.30	T
rs10509540	C10orf59	1.33	0.29	T

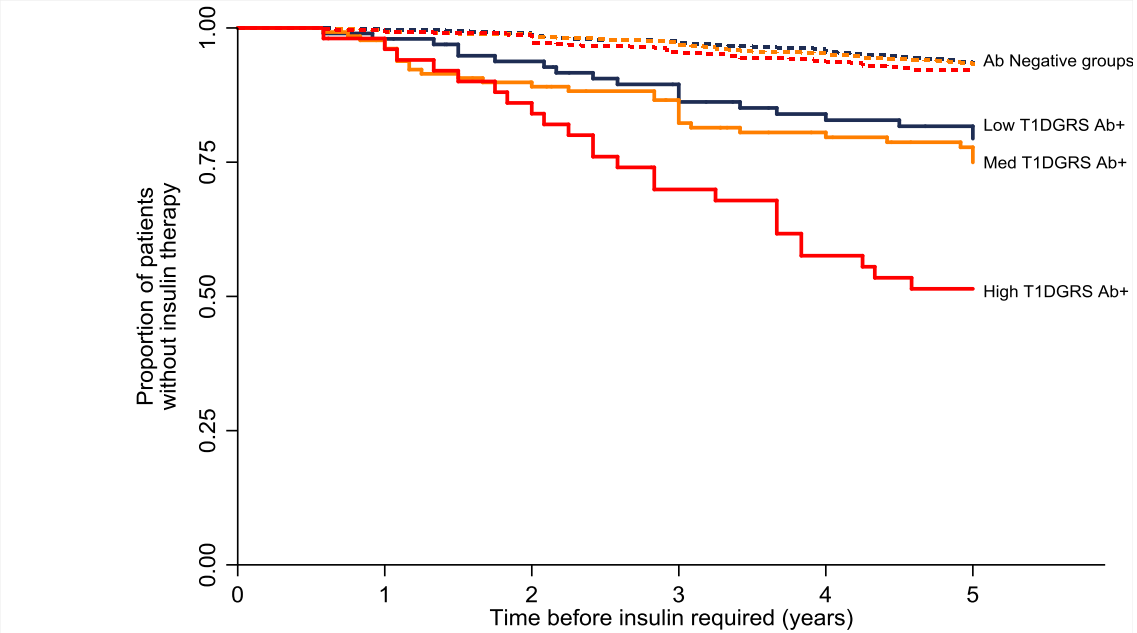
Supplementary Table 4: Participant characteristics stratified by cohort. Median (IQR) or %
Kruskal-Wallis used for comparison testing continuous variables, chi-square for categorical variables
Exeter cohorts are shown combined due to low numbers in PRIBA and MRC Progressor
* Closest to diagnosis
† At latest follow up
‡ Percentage of participants observed for at least five years
§ Centile of participants with type 1 diabetes from the Wellcome trust case control consortium.

	DCS Hoorn (n = 1 942)	*GoDarts (n = 3 963)*	*Exeter studies (n= 2 702)*	*p-value*
Sex (% Male)	54.6%	54.8%	60.2%	<0.001
Age at diagnosis (years)	60 (53, 67)	61 (54, 68)	59 (50, 67)	<0.001
BMI (kg/m^2)*	29.5 (26.8, 33.2)	30.4 (27.2, 34.6)	31.1 (27.5, 35.7)	<0.001
Duration of diabetes (years) †	7.1 (4.3, 11.0)	12.8 (10.3, 15.7)	7.0 (3.0, 12.3)	<0.001
Duration of diabetes (years) at GADA	8.2 (5.3, 12.2)	5.1 (2.7, 8.0)	7.0 (3.0, 12.0)	<0.001
Insulin treated within 5 years (%)‡	5.8	7.4	10.2	<0.001
HbA_{1c} (%)†	6.5 (6.1, 7.1)	7.2 (6.5, 8.2)	7.3 (6.6, 8.4)	<0.001
HbA_{1c} (mmol/mol) †	48 (43, 54)	55 (48, 66)	56 (49, 68)	<0.001
GADA Positive (%)	2.2%	3.9%	3.1%	<0.001
GADA (units/mL)	2.6 (2.0, 3.7)	5.0 (5.0, 5.0)	4.9 (4.9, 4.9)	<0.001
T1D GRS centile§	4.7 (0.9, 16.1)	3.9 (0.5, 15.9)	4.2 (0.7, 16.3)	<0.001

Supplementary Figure 2: Histogram of year of diabetes diagnosis.



Supplementary Figure 3: Kaplan-Meier plot of probability of requiring insulin therapy by risk group using 10 SNP T1D GRS. Solid lines represent GADA positive groups, dashed lines represent GADA negative groups. Blue = low T1D GRS centile, orange = medium T1D GRS centile, red = high T1D GRS centile.



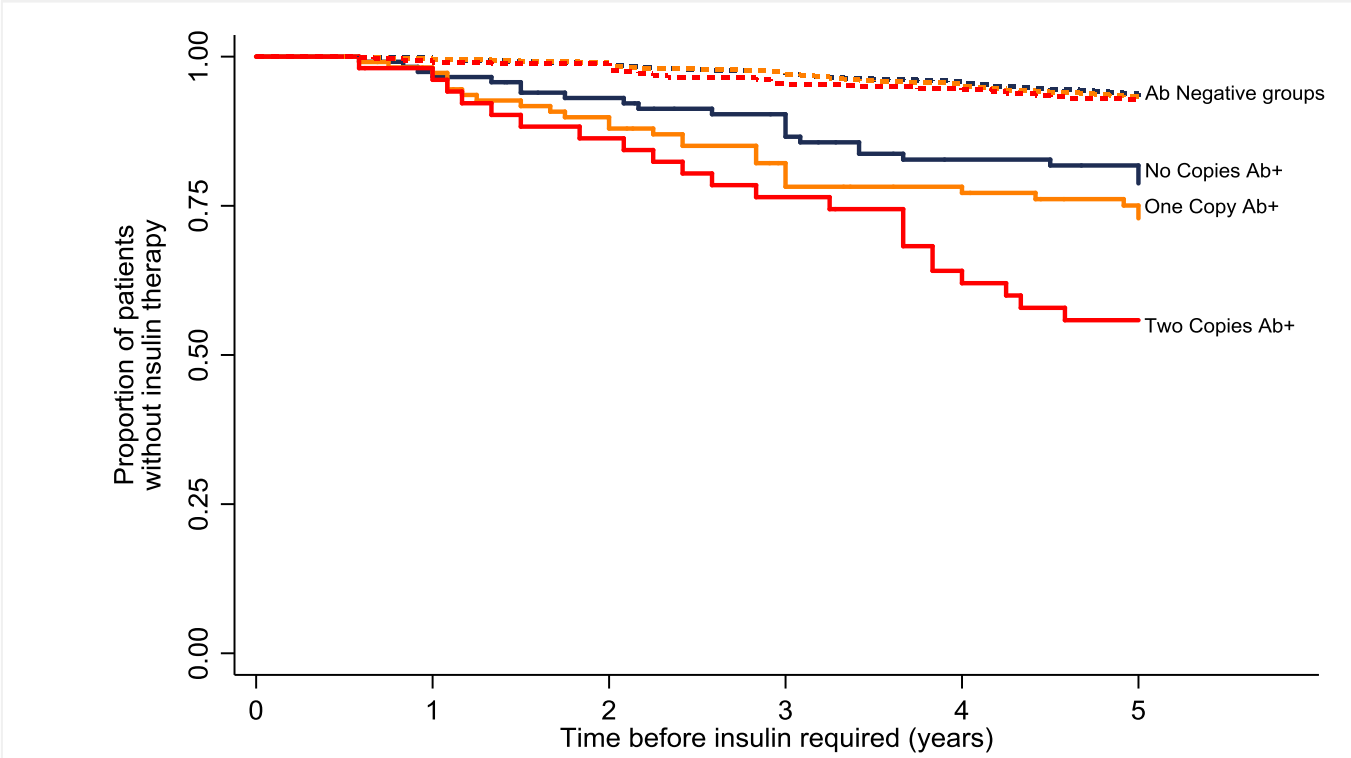
Number at risk

Low TIDGRS Ab +	98	96	89	82	74	71
Med T1DGRS Ab +	131	126	111	101	89	83
High T1DGRS Ab+	51	49	43	34	28	25
Low TIDGRS Ab -	4291	4247	4066	3765	3504	3238
Med T1DGRS Ab -	3545	3501	3335	3084	2826	2601
High T1DGRS Ab -	492	486	465	427	396	368

Supplementary Table 5: Hazard ratios from Cox proportional regression model for time to insulin censored at 5 years (10 SNP T1D GRS) * Closest to diagnosis

Variable	Hazard Ratio [95% CI]	p value
GADA Negative	1	
GADA Positive	3.70 [2.74, 4.99]	<0.001
GADA Negative:10 SNP T1D GRS (per 1 SD change in T1D GRS)	1.04 [0.96, 1.14]	>0.1
GADA Positive:10 SNP T1D GRS (per 1 SD change in T1D GRS)	1.34 [1.05, 1.71]	0.02
Age at diagnosis (per 1 year)	0.97 [0.96, 0.98]	<0.001
BMI (per kg/m^2 unit)	1.00 [0.98, 1.01]	>0.1

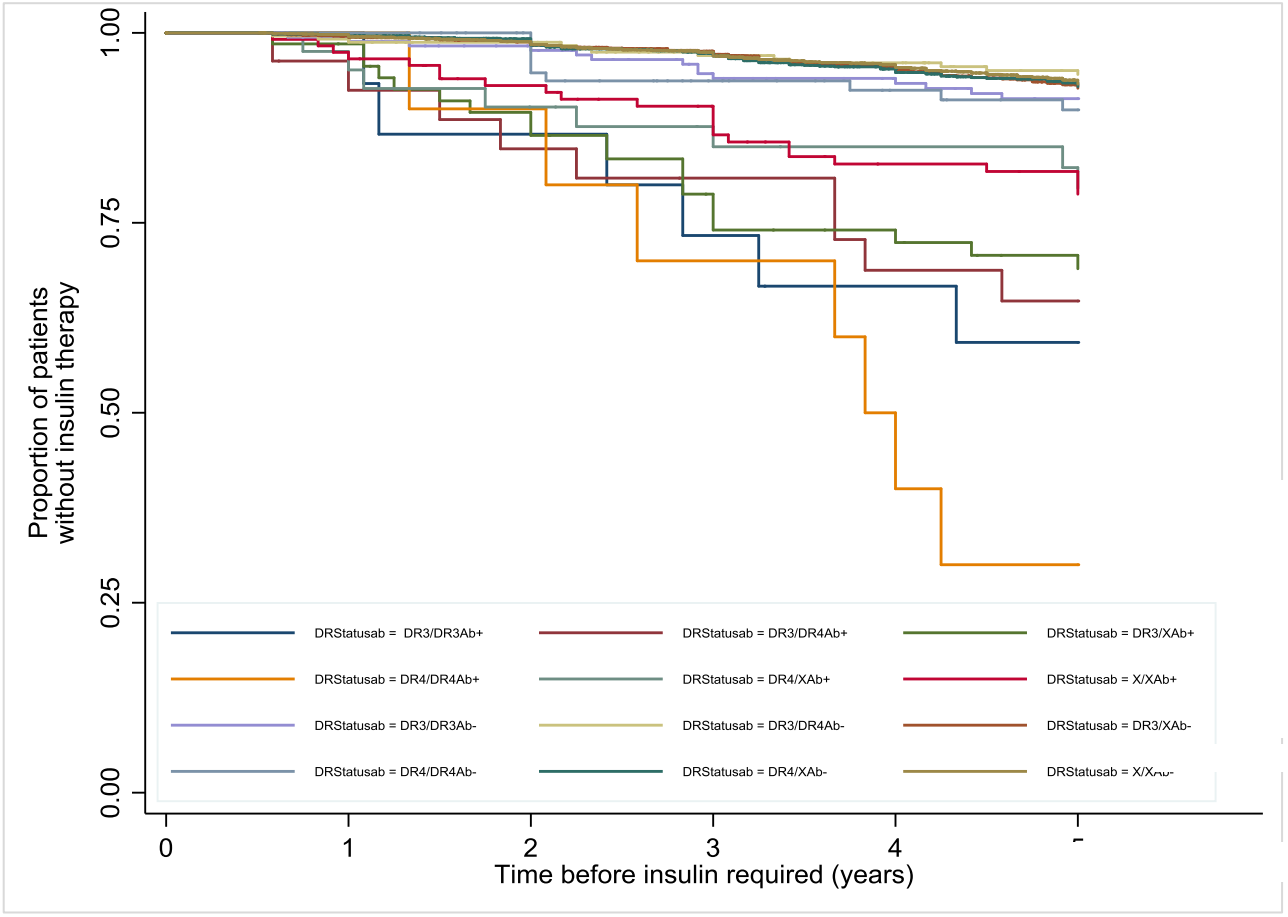
Supplementary Figure 4: Kaplan-Meier plot of probability of requiring insulin therapy by risk group using HLA DR3/DR4 alleles. Solid lines represent GADA positive groups, dashed lines represent GADA negative groups. Blue = No DR3/DR4 copies, orange = one DR3/DR4 copy, red = two DR3/DR4 copies.



Number at risk

NoCopies Ab+	117	114	104	96	84	82
OneCopy Ab +	111	107	95	83	76	70
Two Copies Ab+	52	50	44	38	31	27
NoCopies Ab -	4617	4565	4355	4038	3740	3456
OneCopies Ab -	3186	3152	3016	2791	2574	2365
Two Copies Ab -	524	516	494	446	411	385

Supplementary Figure 5: Kaplan-Meier plot of probability of requiring insulin therapy by risk group using HLA DR3/DR4 alleles.



Chapter 5.

Predicting early insulin requirement in adults diagnosed with type 2 diabetes: development and external validation of a multivariable survival model

Anita L Lynam, John M Dennis, Femke Rutters, Louise A Donnelly, Andrew T Hattersley, Richard A Oram, Colin NA Palmer, Amber A van der Heijden, Fiona Carr, Petra J M Elders, Mike N Weedon, Roderick C Slieker, Leen M 't Hart, Ewan R Pearson, Beverley M Shields, Angus G Jones

Expected to be submitted to Diabetes Care, October 2019

Acknowledgments of co-authors and contributions to paper

I was instrumental in formulation of the study concept, research methods and design. I performed the review of existing literature. I cleaned the data. I designed, guided and performed the statistical analysis, and interpreted the results. I drafted the manuscript.

Beverley Shields and Angus Jones conceived the idea. Andrew Hattersley and Angus Jones researched the Exeter data. Richard Oram and Mike Weedon researched the genetic data, Femke Rutters, Roderick Slieker, Petra Elders, Amber van der Heijden and Leen M 't Hart researched the DCS data, and Louise Donnelly, Fiona Carr, Colin Palmer and Ewan Pearson researched the GoDarts data. Beverley Shields, John Dennis, Andrew Hattersley and Angus Jones assisted with analysing the data and clinical interpretation. John Dennis accessed the ADOPT data and performed the ADOPT validation analysis. Beverley Shields and Angus Jones reviewed and gave feedback on the draft manuscript. All authors critically revised the manuscript and approved the final version.

Rachel Nice of the Blood Sciences Department, Royal Devon and Exeter Hospital conducted additional autoantibody analysis on the GADA islet-autoantibody Exeter samples. Paul Dickman and Sarwar Mozumder provided me with advice on applying the model to other datasets and model implementation.

Abstract

Objective

The rate of glycaemic deterioration in patients with clinically diagnosed type 2 diabetes is highly variable. We aimed to develop and validate a multivariable prognostic model to predict rapid glycaemic progression leading to a requirement for insulin therapy within five years in adult patients diagnosed with type 2 diabetes.

Research Design and Methods

We examined the relationship between seven potential prognostic factors; Age at diagnosis, BMI, Sex, HbA_{1c}, HDL, GAD65 autoantibodies (GADA) and a Type 1 Diabetes Genetic Risk Score, and time to insulin therapy using survival analysis in 3,232 participants with clinical type 2 diabetes (onset >=35 years, treated without insulin from diagnosis). Separate models were developed without knowledge of GADA status, and in GADA positive and negative participants. External validation was performed in observational (Diabetes Care System (DCS) n = 1,241) and (for glycaemic progression on monotherapy) trial (ADOPT, n = 3,487) datasets.

Results

Area under the receiving operator curve (ROC AUC) for insulin requirement at 5 years ranged from 0.74 (95% CI [0.71, 0.77]) (model without GADA) to 0.80 [0.71, 0.88] (GADA positive model). Results were consistent in external validation (model without GADA ROC AUC 0.80 [0.75, 0.85], ADOPT 0.70 [0.67, 0.73]). 70% of participants had <10% probability of insulin requirement at 5 years in a model without biomarker measurement.

Conclusions

Prediction models integrating clinical features with biomarkers may assist clinicians in identifying patients with high risk of progression and those who may benefit most from GADA testing.

The rate of glycaemic deterioration leading to a requirement for insulin therapy in patients with clinically diagnosed type 2 diabetes is highly variable (1); in many patients, glycaemia can be successfully managed with lifestyle changes or oral agents for many years whilst others require insulin therapy soon after diagnosis (2; 3). This heterogeneity may reflect differences in underlying pathophysiology (4-9).

Being able to identify which patients will rapidly progress (or conversely remain stable over many years) may have utility in clinical practice and research. In clinical practice this could facilitate prioritised monitoring and treatment escalation for those likely to progress rapidly, and may allow targeting of therapies with specific effects on glycaemic deterioration (10), however this approach would need to be low cost and therefore based on routinely measured features or inexpensive biomarkers. In research, those patients likely to rapidly progress could be targeted to maximise cost effectiveness of trials of interventions aimed at slowing diabetes progression.

A number of clinical and genetic factors have been found to be associated with the rate of glycaemic deterioration leading to a requirement for insulin therapy in patients with clinically diagnosed type 2 diabetes (1; 4; 5; 11-18). Whilst the definition of failure varies between studies (initiation of pharmacologic treatment, requirement for second or third line treatment, or requirement for insulin), the association between younger age at diagnosis and rapid progression has been a consistent finding and is strongly associated with disease progression (1). Additional clinical features reported to be associated with future progression of type 2 diabetes include, HbA_{1c} or fasting glucose, HDL, Triacylglyceride, alanine transaminase, sex, beta-cell function (Homeostatic Model Assessment (HOMA)), LDL, serum creatinine, smoking

and ethnicity (1; 4; 5; 11-18). However, in a recent large study of progression to insulin therapy, only age at diagnosis, HDL, Triacylglyceride and BMI were independent predictors (11).

The presence of GAD65 autoantibodies (GADA) or (less commonly) other islet-autoantibodies is strongly associated with rapid progression to insulin, however many autoantibody-positive patients do not have early insulin requirement (13; 19; 20). A Type 1 Diabetes Genetic Risk Score (T1D GRS) has been shown to be associated with faster progression to insulin but only in participants who are GADA positive (20). No association was found with type 2 diabetes generic risk (11; 21).

Rather than relying on single features in isolation, the most effective approach to predicting type 2 diabetes progression is likely to be through combining different features, as is now common for outcome prediction in many areas of clinical practice. There are however, no prognostic models that combine clinical features and biomarkers to predict progression in individuals with a clinical diagnosis of type 2 diabetes.

We aimed to develop and validate a prognostic model to predict early insulin requirement in adult patients newly diagnosed with type 2 diabetes.

Methods

We used data from existing prospective studies to develop and validate a multivariable prognostic model to predict progression to insulin therapy, from diagnosis, in adult patients with clinical type 2 diabetes.

Study population

Participants were eligible for the study (model development or validation) if they had a clinical diagnosis of type 2 diabetes after the age of 35 years, and were treated without insulin from diagnosis.

Participants known to have had GADA testing performed either in clinical practice or prior to diagnosis (through a review of electronic laboratory records) were excluded due to the risk of the result influencing the clinician's treatment decision to commence insulin (n=107).

Development cohort

For model development, participants were identified in the Genetics of Diabetes Audit and Research Tayside Study (GoDARTS) (22). GoDarts is a population cohort comprising of longitudinal clinical data (measured at recruitment and from electronic medical record linkage) of participants with a clinical diagnosis of type 2 diabetes recruited from primary and secondary care in Tayside, Scotland, UK. Participants diagnosed with diabetes before 1st January 1994 were excluded from our study; due to insufficient prescribing information we were unable to define time to insulin prior to this date.

External validation cohort

For external validation, participants meeting our study inclusion criteria were identified in the Hoorn Diabetes Care System (DCS) (23). DCS is a longitudinal study of participants diagnosed with type 2 diabetes recruited from primary and secondary care in West-Friesland, Netherlands, clinical data is collected at recruitment and at annual visits.

Assessment of the performance of the model in predicting shorter term glycaemic progression (monotherapy failure) was undertaken in 3,487 participants in the Diabetes Outcome Progression Trial (ADOPT) (10; 24; 25), ADOPT is an intention to treat randomised drug efficacy trial in participants aged 30-75 years with recently diagnosed (< 3 years) type 2 diabetes. Participants were eligible for ADOPT if they had been previously managed with diet/exercise only and had fasting plasma glucose ranging from 126 to 240 mg/dl (7–13 mmol/l) at screening and from 126 to 180 mg/dl (7–10 mmol/l) at randomisation (10).

All participants included in this study were of white-European origin.

Model outcome measure: Time to insulin therapy

The primary outcome was time to insulin therapy, defined as the number of months between diabetes diagnosis and commencement of continuous insulin therapy obtained from electronic prescription records.

In ADOPT, the primary outcome was time to monotherapy failure which was defined according to the study primary outcome as a confirmed level of fasting plasma glucose of more than 180 mg/dl (10.0 mmol/l) (10). Time to insulin could not be assessed in this cohort.

Prognostic factors

Study prognostic factors were selected based on reported independent associations with glycaemic progression of diabetes in previous literature (1; 11) and availability in study cohorts. We examined seven potential prognostic factors; Age at diagnosis, BMI, Sex, HbA_{1c}, HDL, GADA and a T1D GRS. While Triglycerides have previously been reported to be independently associated

with more rapid progression in the GoDARTS cohort (11) it was not included in our study due to limited data availability in the GoDarts participants (25% Triglycerides missing), high collinearity with HDL, and the need for patients to fast prior to measurement which limits clinical utility. To maximise potential utility, and account for potential different interactions between clinical features and progression in autoimmune and non-autoimmune diabetes, models were developed for routinely available predictors without knowledge of GADA status, and separate models then developed for both GADA positive and negative participants, with and without examination of T1D GRS respectively.

BMI

Height and weight measurements were collected at recruitment visit and used to calculate BMI.

Laboratory Measurement

HbA_{1c} and HDL

HbA_{1c} values were included in the model development if collected within +/- 6 months of reported diagnosis date. The closest available value was used for analysis with values prior to diagnosis excluded if in the non-diabetic range (<6.5% (48 mmol/mol)). HDL values were included where the sample was collected within 12 months of diagnosis (before or after diagnosis), with the closest available value used for study analysis.

In the GoDarts cohort, HbA_{1c} and HDL were collected from electronic medical record linkage as previously described (22) (or recruitment visit where this was conducted < 6 months from diagnosis).

In the DCS cohort, HbA_{1c} and HDL were measured at a recruitment visit and were repeated annually as previously described (23). Measurement of HbA_{1c} and HDL was conducted using fasting blood. HbA_{1c} was based on the turbidimetric inhibition immunoassay for haemolysed whole EDTA blood (Cobas c501, Roche Diagnostics, Mannheim, Germany) and is expressed in mmol/mmol according to the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) as well as percentage according to the Diabetes Control and Complications Trial (DCCT)/ National Glycohemoglobin Standardization (23). HDL was determined enzymatically (Cobas c501, Roche Diagnostics) (23).

Measurement of HbA_{1c} and HDL in ADOPT has been described previously (25).

Type 1 Diabetes Genetic Risk Score (T1D GRS)

The development of the T1D GRS has been described previously (26). In brief, T1D GRS consists of 30 common type 1 diabetes genetic variants (single nucleotide polymorphisms (SNPs)) from HLA and non-HLA loci; each variant is weighted by their effect size on type 1 diabetes risk from previously published literature, with weights for DR3/DR4-DQ8 assigned based on imputed haplotypes. The combined score represents an individual's genetic susceptibility to type 1 diabetes. Variants used to derive the score are shown in Supplementary Table 1. For ease of clinical interpretation the score is presented in this article as the centile position of the distribution in the Wellcome Trust Case Control Consortium type 1 diabetes population (27).

Genotyping in the GoDarts cohort was performed using custom genotyping arrays (including Immunochip, Cardio-Metabochip (Metabochip) and Human Exome array) from Illumina as previously described (11). Genotyping in the

DCS cohort was performed with Illumina's HumanCoreExome Array and imputed using IMPUTE2 (28) into the 1000 Genomes March 2012 reference panel. All SNPs had an INFO > 0.8.

T1D GRS calculation was not performed if genotyping results were missing for either of the two alleles with the greatest weighting (DR3/DR4-DQ8 or HLA_DRB1_15) or if more than two of any other SNPs were missing.

GADA

GADA was measured for GoDarts and DCS by the Academic Department of Blood Sciences at the Royal Devon and Exeter Hospital using the RSR Limited ELISA assay (RSR Ltd, Cardiff, UK) on the Dynex DS2 ELISA Robot (Dynex Technologics, Worthing, UK). Sample collection for GADA measurement was a median diabetes duration of 4.5 years (GoDarts) and 7.2 years (DCS). The cut-off for positivity was ≥11 units/ml, based on the 97.5th centile of 1,559 controls without diabetes (29). The lowest reportable value (lowest calibrant) was 5.0 units/ml. The laboratory participates in the International Autoantibody Standardisation Programme.

Missing data

This study is a complete case analysis. To assess the appropriateness of this approach to deal with missing data, we first looked at the missing data patterns to describe the missing data (Supplementary Table 2). The percentage of participants in the development data meeting our inclusion criteria with missing data for the study prognostic factors was 18%.

The missing data mechanism was then investigated using the predictors of missingness method. This method compares the characteristics of the

participants with missing data to those of participants without missing data and can be used to assess the plausibility of the data being missing completely at random (MCAR). This involves using logistic models each with a binary outcome of missing data relating to the variable of interest (1 = missing data, 0 = not missing), with the remaining prognostic factors treated as predictor variables. Significant predictor variables suggest that the data is not MCAR. None of the predictor variables was significant suggesting that the data was MCAR and therefore a complete case analysis was considered appropriate.

Statistical analysis

Model development

We applied a Royston-Parmar flexible parametric survival model (RP) (30-32) programmed in Stata (stpm2) (33). RP models have advantages over Cox models; they can be used when the proportional hazards assumption is not met, can predict survival/failures probabilities over time for individual participants, and have smoothed survival and cumulative hazards functions. We followed the approaches of Royston and Lambert for developing and reporting our RP model (34). Median follow-up time was calculated using the reverse Kaplan-Meier method (35). All analysis was performed in Stata/SE 15.1 (StataCorp, College Station, TX).

Continuous prognostic factors

Each continuous predictor was first modelled (univariate, Cox model) as linear, log-transformed and transformed using orthogonalised restricted cubic splines (3 knots) (36). Nonlinearity between each factor and the outcome was then assessed both visually using plots and statistically using Bayes information

criteria (BIC). Each continuous predictor was mean centred in the modelling to produce a meaningful baseline survival function. T1D GRS was normalised.

Time-dependent covariates

Proportional hazards for each prognostic factor (non-transformed) was checked visually using plots of scaled Schoenfeld residuals against time.

Scale and baseline complexity

Scale and baseline complexity were selected by inspecting the Akaike information criterion (AIC) and BIC statistics of a simple multivariable preliminary model consisting of all prognostic factors (categorised continuous (non-transformed) predictors) with varying scale (log cumulative hazard (hazard), log cumulative odds (odds), standard normal deviate (probit) (normal) and value of theta using the Aranda-Ordaz family of link functions (theta)) and degrees of freedom (1 to 5) (34).

Selection of prognostic factors

We first built a RP model including all prognostic factors without any transformations. We then investigated the inclusion of transformed and time dependent effects to improve the model fit assessed using AIC and BIC. Interactions between continuous variables were not assessed in the model. The goodness of fit of the continuous covariates included in the final model was assessed visually using plots of smoothed martingale residuals.

Evaluation of model performance: Internal validation

Bootstrapping with replacement (1,000 repetitions) was used to estimate optimism adjusted explained variation on the natural scale of the model (R^2_D) (37; 38) and area under the receiver operating characteristic curve (ROC AUC)

at five years. ROC AUC was calculated using the failure probability at five years and a five-year censored outcome (participants with < 5 year duration and non-insulin treated were excluded).

The available range of discrimination was assessed visually by plotting failure probabilities against time at specified centiles of the distribution of the prognostic index.

Evaluation of model performance: External validation

The model was fitted to both external datasets. The quality of the model predictions were evaluated using ROC AUC at five years calculated using same method as previously described. Calibration was assessed visually using a calibration plots at five years.

GADA models

A separate model applicable to participants with a clinical diagnosis of type 2 diabetes who are known to be GADA positive and a model for those known to be GADA negative were developed and validated using the same methods as previously described. T1D GRS was used as a potential predictor in the GADA positive model only.

Results

For the development cohort, we identified 3,232 participants with a clinical diagnosis of type 2 diabetes meeting all of our inclusion criteria. A flow diagram describing the flow of participants through the study is shown in Supplementary Figure 1. Table 1 shows the characteristics for these participants. 97% (n = 3,147) had been followed for at least five years; median follow up time, calculated as the median time to censoring (insulin treatment or latest follow

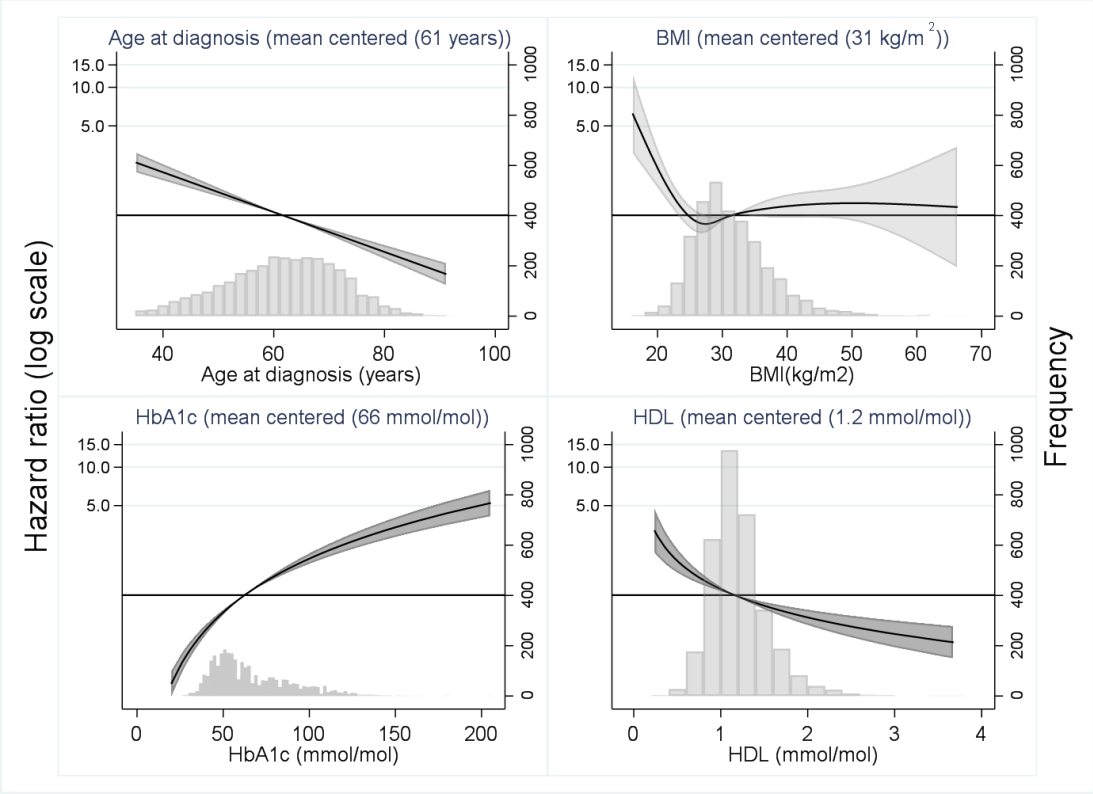
up), was 11.8 (95% CI [11.6, 11.9]) years. 8.8% (n =278) of those participants with over five years follow up had progressed to insulin within 5 years.

Table 1: Participant characteristics for No GADA model cohorts.
Median (IQR) or %
*At first visit
†Percentage of patients observed for at least five years
‡ measured < 6 months post diagnosis
§ Closest to diagnosis (within 12 months pre or post diagnosis)
|| Not followed post failure

	GoDarts development (n = 3,232)	*DCS Validation (n = 1,241)*	*ADOPT validation (n = 3,487)*
Sex (% Male)	54.6%	54.0%	59.1%
Age at diagnosis (years)	62 (54, 69)	61 (54, 67)	58 (51, 65)
BMI (kg/m^2)*	30.4 (27.2, 34.7)	29.5 (26.8, 33.2)	31.0 (27.8, 35.3)
Duration of diabetes (years) at latest follow up	12.3 (9.9, 14.9)	6.3 (4.0, 10.2)	Not Available[||]
Insulin treated within 5 years (%)[†]	8.8%	8.0%	Not Applicable
Monotherapy failure by 4 years	Not Applicable	Not Applicable	15.1%
HbA_{1c} (%)[‡]	7.6 (6.6, 9.5)	6.6 (6.1, 7.5)	7.2 (6.7, 7.8)
HbA_{1c} (mmol/mol) [‡]	60 (49, 80)	48.6 (43.2, 58.5)	55.2 (49.7, 61.7)
HDL (mmol/L) [§]	1.2 (1.0, 1.4)	1.15 (0.99, 1.36)	1.2 (1.0, 1.4)

In univariate analysis, increased risk of progressing to insulin therapy was associated with younger age at diagnosis, lower HDL, higher HbA_{1c} and being female; there was a U-shaped association with BMI, with risk lowest in those with a BMI of 30 ((Figure 1, Supplementary Figure 2 and Supplementary Table 3).

Figure 1: Best fit univariate association between continuous variables and progression to insulin therapy assessed using relative to mean centred hazard obtained from Cox models with 95% confidence intervals. Horizontal line at hazard ratio = 1. Assumes relative effect is the same throughout the follow-up period (n = 3,232).



Supplementary Table 4 shows the univariate R^2_D and C-statistic for each of the best fit clinical features and biomarkers.

Covariate effects varied over time (time dependent)

The effect of HDL, BMI and HbA_{1c} varied over time, invalidating the assumption that each covariate is independent of time. Supplementary Figure 3 shows the effect of HDL and HbA_{1C} reducing over the first five years and BMI increasing over the same period. For HbA_{1c} (considered in isolation) this means that, for

example, at diagnosis, the risk of rapid insulin requirement (hazard ratio) is much higher in patients with a high HbA_{1c} compared to those patients with a lower HbA_{1c} but after five years, the difference in risk between patients with higher and lower HbA_{1c} measured at diagnosis is much less. These time dependent effects can be modelled using spline functions with the knot location specified at five years but have the disadvantage of creating a more complex model with an increased number of degrees of freedom (model parameters). When the model fit (AIC and BIC) was assessed with the inclusion of HDL, BMI and HbA_{1c} as time dependent variables, only the inclusion of HbA_{1c} significantly improved the model fit. Since HDL and BMI were weaker predictors and did not improve the model fit when used as time dependent effects, we decided to include only HbA_{1c} as a time dependent effect. A limitation with the use of spline functions for dealing with time dependent covariates is that the individual model parameters for the time dependent splines are almost impossible to interpret (34). Essentially the time dependent splines represent the situation shown in the Supplementary Figure 3.

Combining clinical features and biomarkers in a prognostic model improves model performance

Age at diagnosis (linear), HDL (log-transformed), sex (male), BMI (3-knot spline) and HbA_{1c} (log-transformed) as a time varying covariate were statistically significant predictors of time to insulin therapy and were included in the final model (Supplementary Table 5). HDL, BMI and HbA_{1c} were time-dependent effects (Supplementary Figure 3) but only HbA_{1c} was included as a time varying covariate since using time varying covariates for HDL and BMI did not significantly improve the model fit. The parameter estimates for the full model are shown in Supplementary Table 6.

ROC AUC of the predicted probabilities for discriminating those who are insulin treated by five years was 0.74 [0.71, 0.77]. The explained variation (R^2_D) of the final model was modest (19% [16%, 23%]). Supplementary Table 7 shows the impact of each covariate in the model on R^2_D. HbA_{1c} was the most important covariate in the model accounting for most of the explained variation (on its own R^2_D = 12%), BMI and sex added the least.

Internal validation results suggest robust model performance

The distribution of the model predicted probabilities for requiring insulin by 5 years was skewed and in most cases fairly low (median (range) 5.3% (0.3 - 62.75)) (Supplementary Figure 4).

The model showed reasonable discrimination for requirement of insulin by 5 years, with those in the highest deciles of predicted probability having the highest risk (18% predicted to require insulin for the 90th centile compared with 1.6% in the lowest decile), (Supplementary Figure 5). The model calibration was good (Supplementary Figure 6).

Results of the bootstrap internal validation showed low levels of optimism (Supplementary Table 8).

Similar model performance in external validation cohort

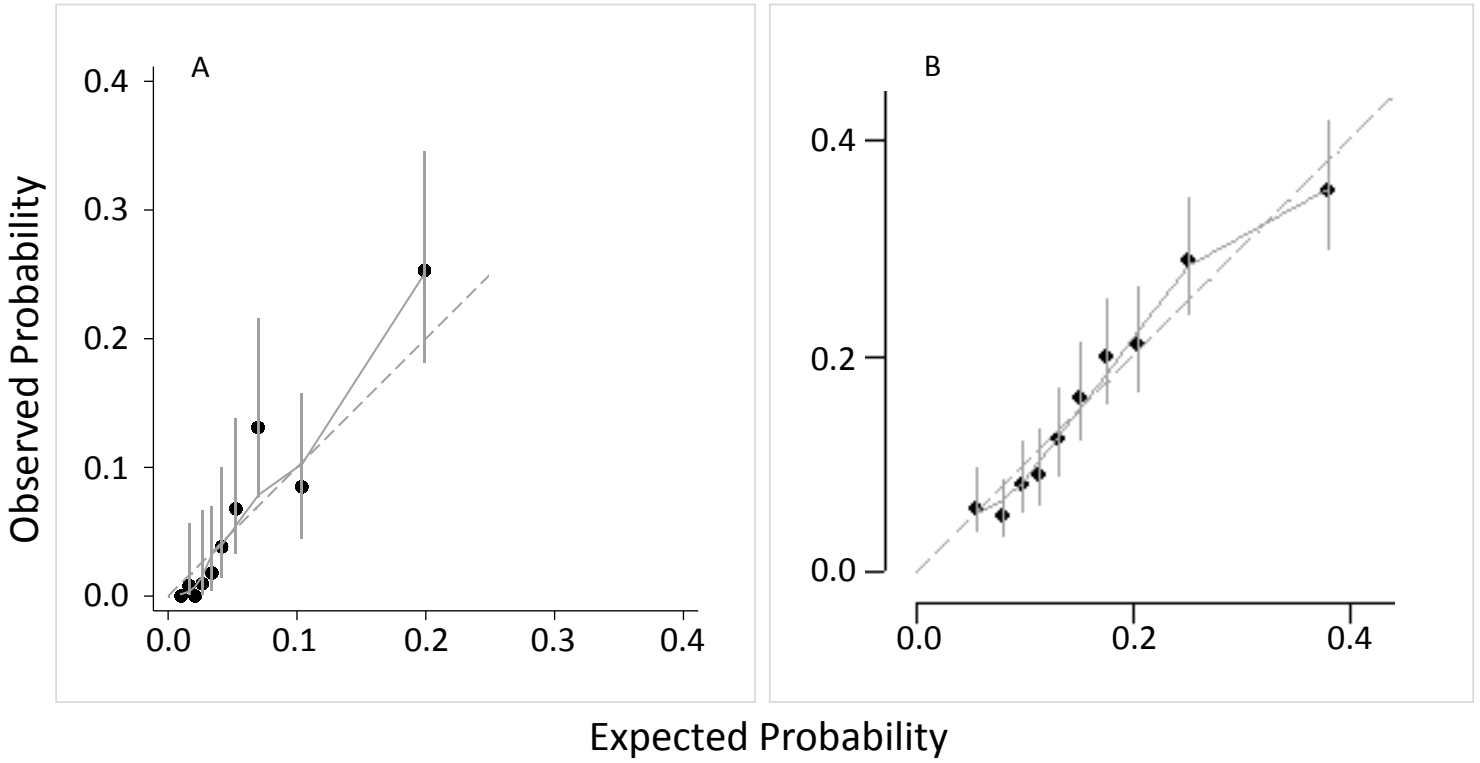
1,241 participants in the DCS study and 3,487 in ADOPT met criteria for external validation (Supplementary Figures 7 & 8). Table 1 shows the characteristics for the DCS and ADOPT participants included in the external validation.

The ROC AUC at 5 years for the DCS external validation cohort was 0.80 [0.75, 0.85]. The model calibrated reasonably well in the DCS cohort (Figure 2 (A)),

although probabilities were slightly underestimated overall (5.7% expected to be insulin treated within 5 years v 6.5% observed), with the model fitting less well at the extremes.

To deal with a shorter follow up duration in ADOPT (four year treatment period (10)), we evaluated the model using the ADOPT external validation cohort at four years. ROC AUC at four years was 0.70 [0.67, 0.73]. Consistent with the different outcome assessed (monotherapy failure as opposed to insulin requirement) the expected probabilities were lower than the observed outcome (expected/observed = 0.27), we therefore recalibrated the model using a baseline hazard (intercept) update (Figure 2 (B)).

Figure 2A: DCS external validation calibration plot of expected versus observed failure probabilities at t = 5 years.
Figure 2B: Re-calibrated ADOPT validation calibration plot of expected versus observed failure probabilities at t = 4 years.
Dashed grey line is reference line where observed = expected probabilities. Black filled circles are risk groups using deciles of expected probabilities, vertical grey solid lines are 95% CIs. Grey solid line is lowess smoother.



A GADA positive model showed good performance in internal validation

We identified 131 participants who were GADA positive in the development data meeting all of our inclusion criteria (Supplementary Table 9). The final model consisted of age at diagnosis (log-transformed and TVC), BMI (log-transformed), HbA_{1c} (log-transformed) and sex (Supplementary Table 10). The parameter estimates for the full model are shown in Supplementary Table 11.

ROC AUC at five years was 0.80 [0.71, 0.88]. R^2_D was reasonable (33% [18%, 46%]). HbA_{1c} was the most important factor in the model (on its own R^2_D = 16%) (Supplementary Table 12). Internal validation suggested low levels of optimism (Supplementary Table 13).

Due to small sample size (n = 28 and n = 138 GADA positive participants meeting inclusion criteria in DCS and ADOPT respectively, of whom only 9 and 18 met the study glycaemic failure definition), external validation could not be performed.

A GADA negative model has similar discrimination performance to the main model

We identified 3,101 participants who were GADA negative in the development data meeting all of our inclusion criteria (Supplementary Table 14). The final model consisted of age at diagnosis (linear), HbA_{1c} (log-transformed and TVC), HDL (log-transformed) and sex (Supplementary Table 15). The parameter estimates for the full model are shown in Supplementary Table 16. ROC AUC at five years was 0.73 [0.69, 0.76]. R^2_D was modest (22% [18%, 26%]) with HbA_{1c} again the strongest predictor (R^2_D 14%). Internal validation showed low levels of optimism (Supplementary Table 17).

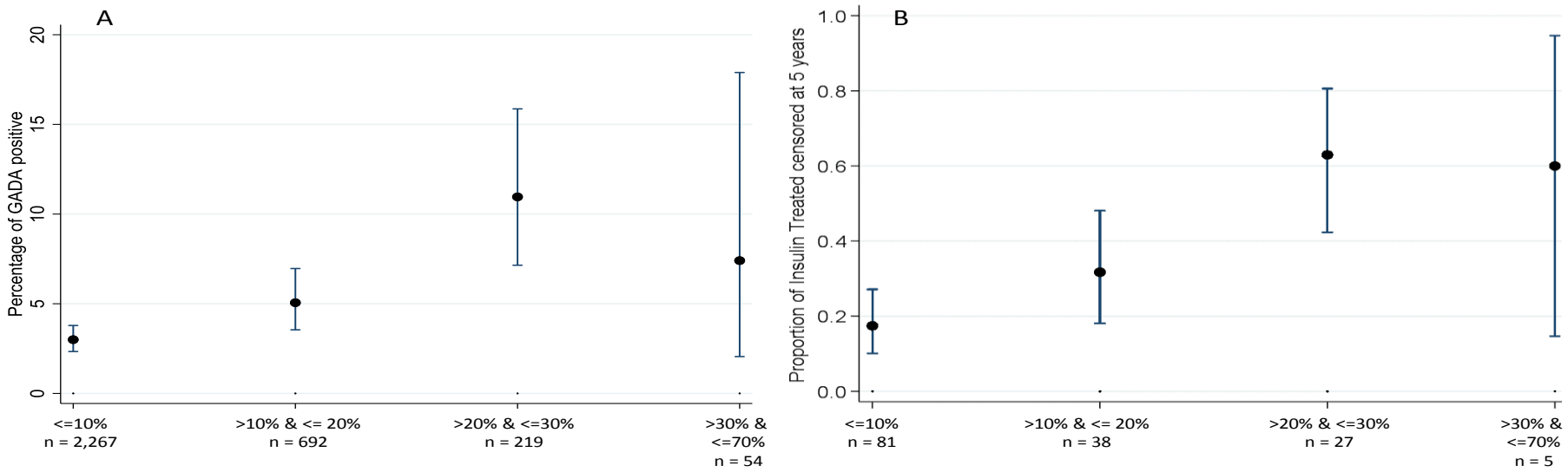
We identified 1,213 participants who were GADA negative in the DCS external validation cohort satisfying our inclusion criteria (Supplementary Table 14). The calibration plot at five years (Supplementary Figure 9) shows that the range of probabilities is again narrow. The model showed reasonable calibration but underestimated in the higher risk groups. There was a slight increase in ROC AUC at 5 years (ROC AUC 0.76 [0.70, 0.82]) when compared to the internal validation results.

We also identified 3,208 participants in the ADOPT data (Supplementary Table 14). ROC AUC at four years was 0.71 [0.68, 0.74]. After adjusting the model to reflect the difference in outcome incidence (baseline hazard (intercept) update), the results of the calibration (Supplementary Figure 10) were very similar to the results of the ADOPT external validation of the main model.

The majority of participants who have low model probabilities may be unlikely to benefit from GADA testing

Figure 3 shows the probability of a positive GADA test (3A) and how much difference a GADA positive result makes to a patients probability of progressing to insulin within five years (3B). Participants with higher model probability have far higher likelihood of testing GADA positive (3A), and have higher rates of rapid insulin requirement if they test positive for GADA (3B). In contrast the vast majority of participants who have low model probability have a low likelihood of a positive GADA test, and low likelihood of rapid progression even if they test positive, for example 70% of participants have model probability ≤10%, this group had only a 3% GADA positive rate in the development dataset and only 17% of the GADA positive participants rapidly progressed to insulin within five years.

Figure 3A: Percentage of GADA positive participants by model probability of five year insulin requirement without GADA testing (no GADA measurement model). Development (GoDarts) cohort n = 3232.
Figure 3B: Proportion of GADA positive participant progressing to insulin within five years (95% CI) in the GoDarts and DCS cohorts (n = 151), by risk of five year insulin requirement without GADA testing (no GADA measurement model).
Error bars denote 95% CI



Probability of five year insulin requirement (no GADA measurement model)

Conclusions

We have developed, evaluated and externally validated prognostic models combining clinical features and biomarkers to provide estimates of a patient's risk of progression to insulin therapy within five years of diagnosis.

All models had a ROC AUC > 0.7 and performed similarly in external cohorts providing confidence in the validity of the models. The performance of our models is similar to those of other existing prognostic models routinely used in clinical practice for example in cardiovascular disease and mortality prediction (39; 40). The initial poor calibration of the models in ADOPT is expected due to the use of a different outcome (moderate glycaemia on monotherapy, in contrast to insulin treatment used for model development), however this was addressed by a simple recalibration of the model. Models were highly predictive of a participant being GADA positive, with those who had low model probability having very low rates of GADA positivity, and GADA positive participants in this group only low probability of 5 year insulin requirement.

To our knowledge, this is the first study to develop a prognostic model for progression to insulin therapy in adult participants with type 2 diabetes. A key strength of this study is our use of a population cohort (GoDarts) for model development, this means that our results are likely to be a true representation of patients seen in primary care. Additional key strengths are the availability of clinical features at diagnosis, our systematic approach to model development (34) and our use of separate cohorts for external validation. We have used both unambiguous definitions of the prognostic factors and reproducible measurements which means our models are usable in clinical practice (41). We built separate GADA models rather than simply using GADA as a covariate due

to the presence of statistically significant interactions between GADA and each of BMI, T1D GRS and HDL, inclusions of these interactions would have resulted in a highly complex model.

Limitations of this study include that insulin commencement was based on clinical decision making rather than a trial protocol (1), we have however addressed this by externally validating the models in ADOPT using a trial glucose threshold. In addition the models have been developed on a white European population, validating these models in other ethnicities is therefore an important area of future work. An additional limitation of our study is that GADA was measured at a median 4.9 years diabetes duration, which could result in a lower prevalence than if measurement was undertaken at diagnosis. However, in adult populations the error is likely to be small, with GADA positivity being stable over the first six years in UKPDS study participants (42). In addition we did not have sufficient external data to externally validate the GADA positive model. Lastly we were not able to use the most recently published Type 1 Diabetes Genetic Risk Score (which has modestly improved performance), due to unavailability of all the relevant SNPs in our cohorts (43).

We have previously shown that the T1D GRS is independently associated with rapid progression to insulin in patients diagnosed with type 2 diabetes who were GADA positive, using a larger cohort of 8,608 participants (20). In this study, T1D GRS was not statistical significant in the GADA positive model when adjusted for the other clinical features and biomarkers. This may reflect a smaller cohort, more advanced modelling to optimize use of other features, and the inclusion of additional predictive features including HbA_{1c}.

Our models have the potential to facilitate the management of patients diagnosed with type 2 diabetes by allowing identification of individuals who have a high probability of rapid glycaemic progression and may benefit from more intensive treatment or monitoring. A potential related role is in assisting targeted GADA testing to the minority of patients who have higher risk of islet antibody positivity, and in whom a positive antibody will be associated with high rates of progression. While we have focused on five year insulin requirement our use of the flexible parametric survival models means that survival probabilities can be calculated for any time point. We envisage that the model will be implemented as a dynamic web-based tool similar to that used in a cancer survival model (44) and may potentially be used at diagnosis alongside recently published prediction models for diabetes classification (45).

In conclusion, prediction models integrating clinical features with biomarkers may assist clinicians in identifying patients with high risk of progression and those who may benefit most from GADA testing.

References

1. Donnelly LA, Zhou K, Doney ASF, Jennison C, Franks PW, Pearson ER: Rates of glycaemic deterioration in a real-world population with type 2 diabetes. Diabetologia 2018;61:607-615
2. U.K. Prospective Diabetes Study Group: U.K. Prospective Diabetes Study 16: Overview of 6 Years' therapy of Type II Diabetes: A Progressive Disease. Diabetes 1995;44:1249
3. Fonseca VA: Defining and Characterizing the Progression of Type 2 Diabetes. Diabetes care 2009;32:S151
4. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad RB, Aly DM, Almgren P, Wessman Y, Shaat N, Spégel P, Mulder H, Lindholm E, Melander O, Hansson O, Malmqvist U, Lernmark Å, Lahti K, Forsén T, Tuomi T, Rosengren AH, Groop L: Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology
5. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT: Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. The Lancet Diabetes & Endocrinology 2019;7:442-451
6. Tuomi T, Carlsson A, Li H, Isomaa B, Miettinen A, Nilsson A, Nissén M, Ehrnström BO, Forsén B, Snickars B, Lahti K, Forsblom C, Saloranta C, Taskinen MR, Groop LC: Clinical and genetic characteristics of type 2 diabetes with and without GAD antibodies. Diabetes 1999;48:150
7. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT: Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine 2015;7:311ra174
8. Udler MS: Type 2 Diabetes: Multiple Genes, Multiple Diseases. Current Diabetes Reports 2019;19:55
9. Pearson ER: Type 2 diabetes: a multifaceted disease. Diabetologia 2019;62:1107-1112
10. Viberti G, Kahn SE, Greene DA, Herman WH, Zinman B, Holman RR, Haffner SM, Levy D, Lachin JM, Berry RA, Heise MA, Jones NP, Freed MI: A Diabetes Outcome Progression Trial (ADOPT). Diabetes care 2002;25:1737
11. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CNA, Pearson ER: Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. Diabetes care 2014;37:718
12. Levy J, Atkinson AB, Bell PM, McCance DR, Hadden DR: Beta-cell deterioration determines the onset and rate of progression of secondary dietary failure in type 2 diabetes mellitus: the 10-year follow-up of the Belfast Diet Study. Diabetic medicine : a journal of the British Diabetic Association 1998;15:290-296
13. Turner R, Stratton I, Horton V, Manley S, Zimmet P, Mackay IR, Shattock M, Bottazzo GF, Holman R: UKPDS 25: autoantibodies to islet-cell cytoplasm and glutamic acid decarboxylase for prediction of insulin requirement in type 2 diabetes. The Lancet 1997;350:1288-1293
14. Matthews DR, Cull CA, Stratton IM, Holman RR, Turner RC: UKPDS 26: sulphonylurea failure in non-insulin-dependent diabetic patients over six years. Diabetic Medicine 1998;15:297-303
15. Ringborg A, Lindgren P, Yin DD, Martinell M, Stålhammar J: Time to insulin treatment and factors associated with insulin prescription in Swedish patients with type 2 diabetes. Diabetes & Metabolism 2010;36:198-203

16. Cook MN, Girman CJ, Stein PP, Alexander CM, Holman RR: Glycemic control continues to deteriorate after sulfonylureas are added to metformin among patients with type 2 diabetes. Diabetes care 2005;28:995-1000
17. Pani LN, Nathan DM, Grant RW: Clinical Predictors of Disease Progression and Medication Initiation in Untreated Patients With Type 2 Diabetes and A1C Less Than 7%. Diabetes care 2008;31:386
18. Waldman B, Jenkins AJ, Davis TME, Taskinen M-R, Scott R, O'Connell RL, Gebski VJ, Ng MKC, Keech AC: HDL-C and HDL-C/ApoA-I Predict Long-Term Progression of Glycemia in Established Type 2 Diabetes. Diabetes care 2014;37:2351
19. Groop LC, Bottazzo GF, Doniach D: Islet Cell Antibodies Identify Latent Type I Diabetes in Patients Aged 35–75 Years at Diagnosis. Diabetes 1986;35:237
20. Grubb AL, McDonald TJ, Rutters F, Donnelly LA, Hattersley AT, Oram RA, Palmer CNA, van der Heijden AA, Carr F, Elders PJM, Weedon MN, Slieker RC, 't Hart LM, Pearson ER, Shields BM, Jones AG: A Type 1 Diabetes Genetic Risk Score Can Identify Patients With GAD65 Autoantibody–Positive Type 2 Diabetes Who Rapidly Progress to Insulin Therapy. Diabetes care 2019;42:208
21. Hornbak M, Allin KH, Jensen ML, Lau CJ, Witte D, Jørgensen ME, Sandbæk A, Lauritzen T, Andersson Å, Pedersen O, Hansen T: A Combined Analysis of 48 Type 2 Diabetes Genetic Risk Variants Shows No Discriminative Value to Predict Time to First Prescription of a Glucose Lowering Drug in Danish Patients with Screen Detected Type 2 Diabetes. PLOS ONE 2014;9:e104837
22. Hebert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, Donnelly LA, Tavendale R, Leese G, Colhoun HM, Dow E, Morris AD, Doney AS, Lang CC, Pearson ER, Smith BH, Palmer CNA: Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). International journal of epidemiology 2017;
23. van der Heijden AA, Rauh SP, Dekker JM, Beulens JW, Elders P, t Hart LM, Rutters F, van Leeuwen N, Nijpels G: The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. BMJ open 2017;7:e015599
24. Kahn SE, Haffner SM, Heise MA, Herman WH, Holman RR, Jones NP, Kravitz BG, Lachin JM, O'Neill MC, Zinman B, Viberti G: Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. N Engl J Med 2006;355:2427-2443
25. Viberti G, Lachin J, Holman R, Zinman B, Haffner S, Kravitz B, Heise MA, Jones NP, O'Neill MC, Freed MI, Kahn SE, Herman WH: A Diabetes Outcome Progression Trial (ADOPT): baseline characteristics of Type 2 diabetic patients in North America and Europe. Diabetic medicine : a journal of the British Diabetic Association 2006;23:1289-1294
26. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN: A Type 1 diabetes genetic risk score can aid discrimination between Type 1 and Type 2 diabetes in young adults. Diabetes care 2016;39:337-344
27. The Wellcome Trust Case Control C: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-678
28. Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 2009;5:e1000529

29. McDonald TJ, Colclough K, Brown R, Shields B, Shepherd M, Bingley P, Williams A, Hattersley AT, Ellard S: Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from Type 1 diabetes. Diabetic medicine : a journal of the British Diabetic Association 2011;28:1028-1033
30. Royston P, Parmar MK: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine 2002;21:2175-2197
31. Lambert PC, Royston P: Further Development of Flexible Parametric Models for Survival Analysis. The Stata Journal 2009;9:265-290
32. Royston P: Flexible Parametric Alternatives to the Cox Model, and more. The Stata Journal 2001;1:1-28
33. Lambert PC: STPM2: Stata module to estimate flexible parametric survival models, Statistical Software Components S457128. Boston College Department of Economics, revised 11 Oct 2018, 2010
34. Royston P, Lambert P: Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. USA, Stata Press, 2011
35. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials 1996;17:343-346
36. Lambert PC: RCSGEN, Stata module to generate restriced cubic splines and their derivatives, Statistical Software Components S456986, Boston College Department of Economics, revised 15 Sep 2015. 2008
37. Royston P, Sauerbrei W: A new measure of prognostic separation in survival data. Statistics in medicine 2004;23:723-748
38. Royston P: Explained variation for survival models. Stata Journal 2006;6:83-96
39. Liao Y, McGee DL, Cooper RS, Sutkowski MBE: How generalizable are coronary risk prediction models? Comparison of Framingham and two national cohorts. American Heart Journal 1999;137:837-845
40. Knaus WA, Draper EA, Wagner DP, Zimmerman JE: APACHE II: a severity of disease classification system. Critical care medicine 1985;13:818-829
41. Moons KG, Altman DG, Vergouwe Y, Royston P: Prognosis and prognostic research: application and impact of prognostic models in clinical practice. Bmj 2009;338:b606
42. Desai M, Cull CA, Horton VA, Christie MR, Bonifacio E, Lampasona V, Bingley PJ, Levy JC, Mackay IR, Zimmet P, Holman RR, Clark A: GAD autoantibodies and epitope reactivities persist after diagnosis in latent autoimmune diabetes in adults but do not predict disease progression: UKPDS 77. Diabetologia 2007;50:2052-2060
43. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, Schneider DA, Locke JM, Tyrrell J, Weedon MN, Hagopian WA, Oram RA: Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. Diabetes care 2019;42:200-207
44. Mozumder SI, Dickman PW, Rutherford MJ, Lambert PC: InterPreT cancer survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and cancer epidemiologists. Cancer epidemiology 2018;56:46-52
45. Lynam AL, McDonald T, Hill A, Dennis JM, Oram R, Pearson E, Weedon M, Hattersley A, Owen K, Shields B, Jones A: Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18 to 50. BMJ Open (in press) 2019;

Supplementary material

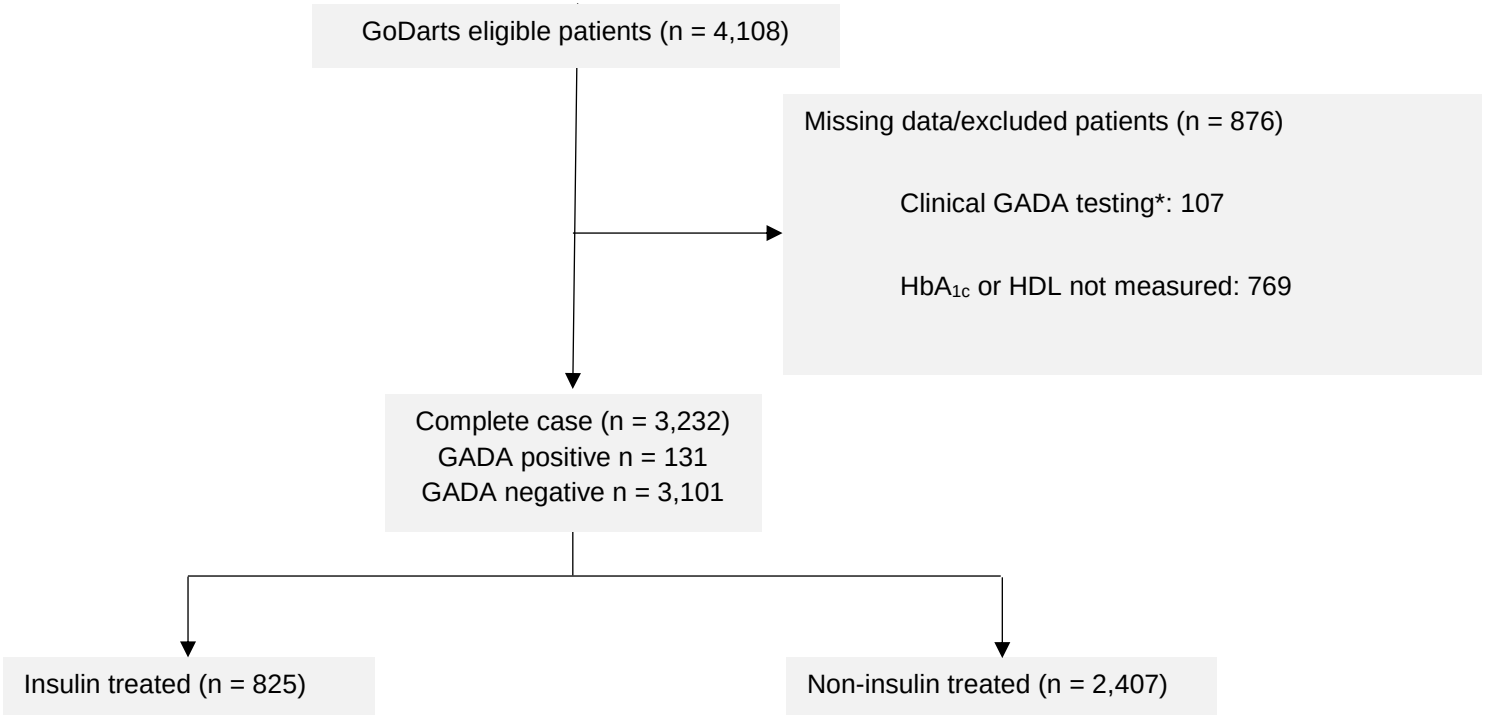
Supplementary Table 1: Type 1 diabetes SNPs included in the genetic risk score with weights. Effect allele is the risk increasing allele on the positive strand.

SNP	Gene	Odds Ratio	Weight	Effect Allele
rs2187668, rs7454108	DR3/DR4	48.18	3.87	
	DR3/DR3	21.12	3.05	
	DR4/DR4	21.98	3.09	
	DR4/X	7.03	1.95	
	DR3/X	4.53	1.51	
rs1264813	HLA_A_24	1.54	0.43	T
rs2395029	HLA_B_5701	2.5	0.92	T
rs3129889	HLA_DRB1_15	14.88	2.70	A
rs2476601	PTPN22	1.96	0.67	A
rs689	INS	1.75	0.56	T
rs12722495	IL2RA	1.58	0.46	T
rs2292239	ERBB3	1.35	0.30	T
rs10509540	C10orf59	1.33	0.29	T
rs4948088	COBL	1.3	0.26	C
rs7202877		1.28	0.25	G
rs12708716	CLEC16A	1.23	0.21	A
rs3087243	CTLA4	1.22	0.20	G
rs1893217	PTPN2	1.2	0.18	G
rs11594656	IL2RA	1.19	0.17	T
rs3024505	IL10	1.19	0.17	G
rs9388489	C6orf173	1.17	0.16	G
rs1465788		1.16	0.15	C
rs1990760	IFIH1	1.16	0.15	T
rs3825932	CTSH	1.16	0.15	C
rs425105		1.16	0.15	T
rs763361	CD226	1.16	0.15	T
rs4788084	IL27	1.16	0.15	C
rs17574546		1.14	0.13	C
rs11755527	BACH2	1.13	0.12	G
rs3788013	UBASH3A	1.13	0.12	A
rs2069762	IL2	1.12	0.11	A
rs2281808		1.11	0.10	C
rs5753037		1.1	0.10	T

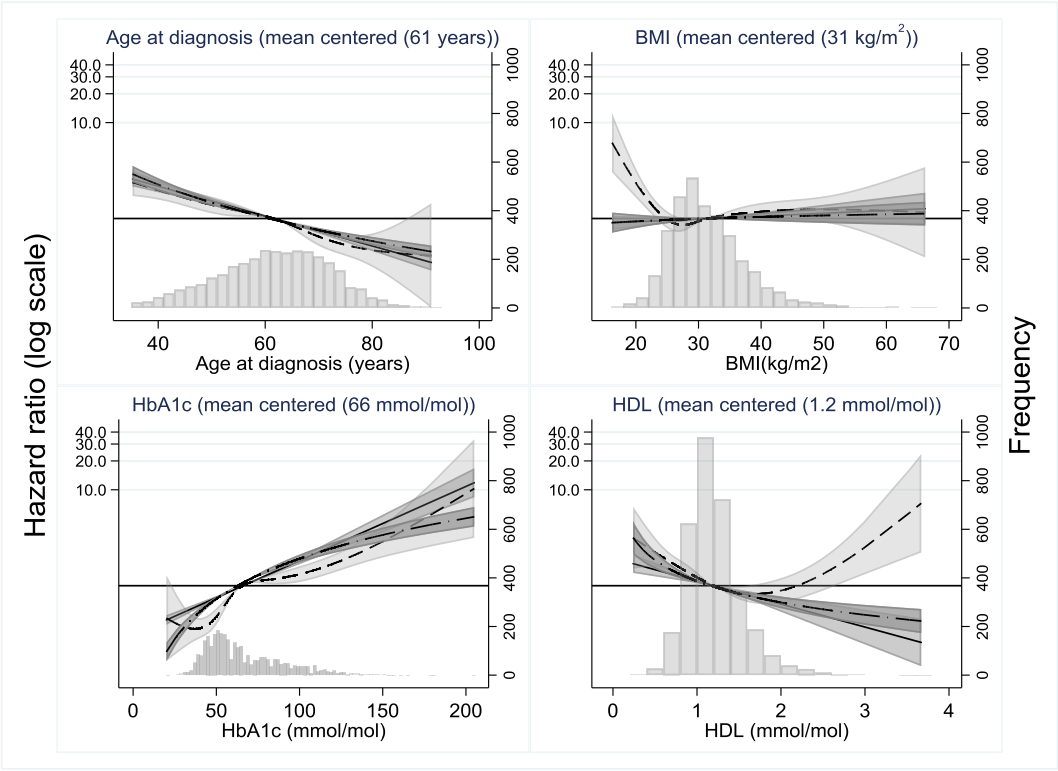
Supplementary Table 2: Missing data pattern (n = 4,001). 1 means complete. Table only includes predictor variables that had missing data.

Percent missing	HDL	HbA_{1c}
82%	1	1
6%	0	0
6%	1	0
5%	0	1

Supplementary Figure 1: Participant flow diagram (GoDarts development cohort) ***** identified through search of electronic laboratory records.



Supplementary Figure 2: Univariate association between continuous variables in Model 1 and insulin treated outcome assessed using relative to mean centred hazard obtained from Cox models with 95% confidence intervals. Black solid lines are linear model, black dashed lines are 3 knot models, dash-dot lines are log transformed. Horizontal line at hazard ratio = 1. Assumes relative effect is the same throughout the follow-up period (n = 3232).



Supplementary Table 3: Selection of functional forms for continuous prognostic factors. Obtained from Cox models. N=2407 used in calculating BIC.

	Linear		Log-transformed		3-knot spline	
	AIC	BIC	AIC	BIC	AIC	BIC
Age at Diagnosis	12387.72	12393.51	12390.97	12396.75	12391.94	12415.09
BMI	12495.06	12500.84	12495.84	12501.62	12477.11	12500.26
HbA_{1c}	12324.11	12329.89	12310.86	12316.64	12302.48	12325.63
HDL	12473.46	12479.25	12464.39	12470.17	12455.04	12478.19

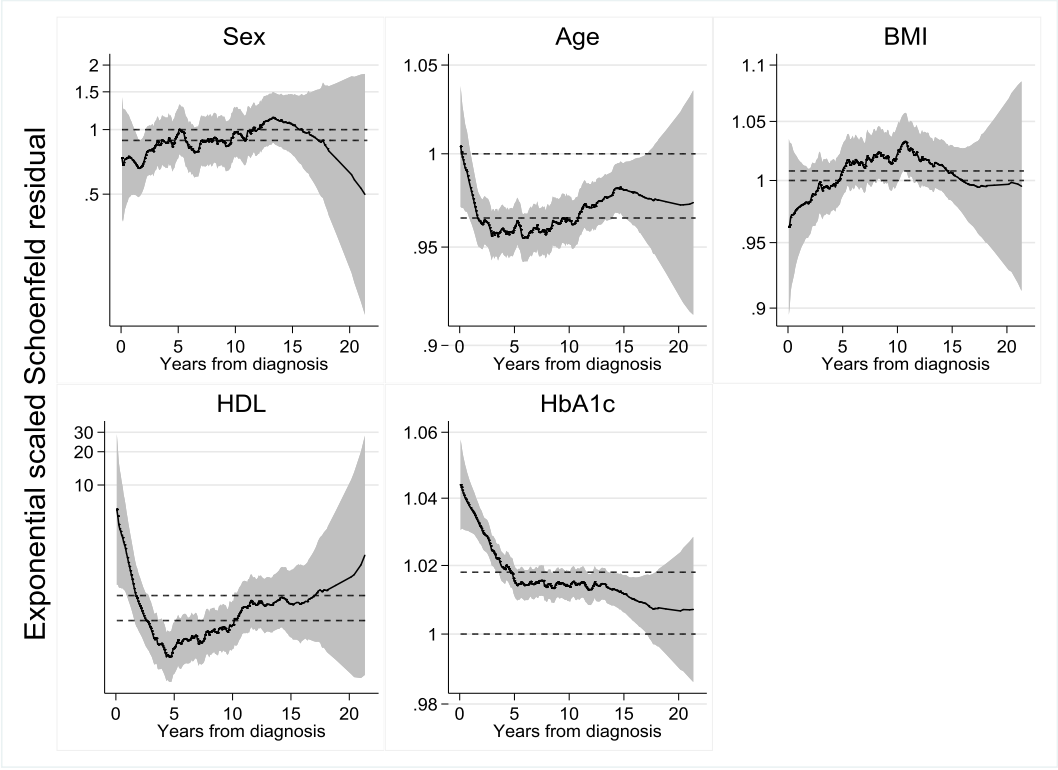
Supplementary Table 4: Performance of univariate associations with progression to insulin therapy using simple Cox model.

Prognostic factor	Adjusted R^2_D [95% CI]	C-statistic [95% CI]
Age at diagnosis (linear)	0.08 [0.05, 0.11]	0.61 [0.59, 0.63]
BMI (3-knot spline)	0.01 [0.00, 0.02]	0.54 [0.51, 0.56]
HbA_{1c} (log-transformed)	0.12 [0.09, 0.16]	0.66 [0.64, 0.67]
HDL (log-transformed)	0.03 [0.01, 0.04]	0.57 [0.54, 0.59]
Sex (Male)	0.003 [0.00, 0.01]	0.52 [0.50, 0.54]

Supplementary Table 5: Hazard Ratios of the four covariates in the RP model (normal scale with 3 d.f.) for time to insulin.
* Centered variables. † Log-transformed. ‡Derived spline variables for BMI. Full model including derived spline variables for the baseline normal cumulative hazard, derived spline variables for the time-dependent effect of HbA_{1c} and intercept is shown in separate table.

Variable	Hazard Ratio [95% CI]	P value
HbA_{1c} (mmol/mol)*†	2.66 [2.31, 3.06]	<0.001
Age at diagnosis (years)*	0.98 [0.97, 0.98]	<0.001
Sex (male)	0.87 [0.78, 0.95]	0.004
HDL (mmol/mol)* †	0.74 [0.62, 0.89]	0.001
BMI_1*‡	0.95 [0.91, 1.00]	0.056
BMI_2*‡	0.93 [0.89, 0.97]	0.001
BMI_3*‡	1.09 [1.04, 1.14]	<0.001
BMI_4*‡	0.98 [0.94, 1.02]	0.349

Supplementary Figure 3: Univariate analysis smoothed Schoenfeld residuals against time plots. Dashed reference lines at 1 (null effect) and estimated value of HR. Trends in the running line smoother indicate non-proportional hazards.



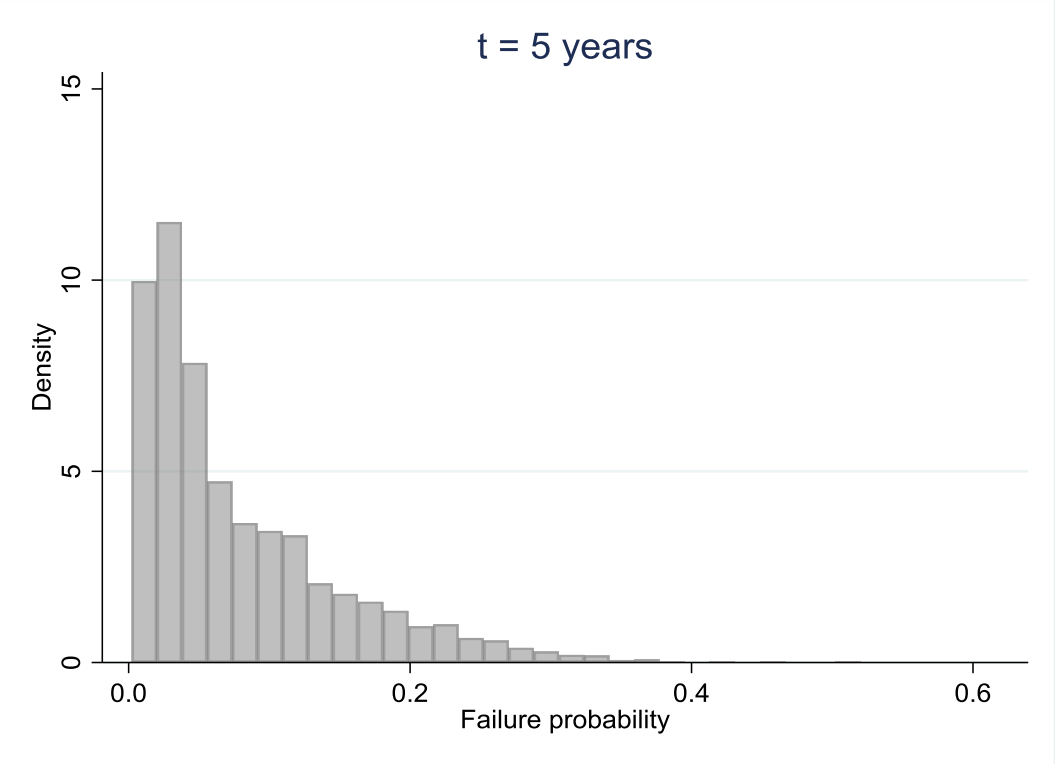
Supplementary Table 6: Model coefficients for the RP model (normal scale with 3 d.f.) for model replication purpose.
* Centered variables. † Log-transformed. ‡Derived spline variables for BMI. §Derived spline variables for the baseline normal cumulative hazard. ‖Derived spline variables for the time-dependent effect of Hba_{1c}. HbA_{1c} is centered on 66 mmol/mol (8.2%), BMI on 31 kg/m2, Age is centered on 61 years and HDL on 1.2 mml/mol. BMI knots at knots 16.20, 27.20, 30.40, 34.70, 66.20.

Variable	Beta coefficient [95% CI]	P value
HbA_{1c} (mmol/mol)*†	0.9770372 [0.8365849, 1.11749]	<0.001
Age at diagnosis (years)*	-0.0218331 [-0.0267874, -0.0168788]	<0.001
Sex (male)	-0.1438583 [-0. 2413587, -0.0463578]	0.004
HDL (mmol/mol)* †	-0.3005002 [-0.4800388, -0.1209615]	0.001
BMI_1*‡	-0.046877 [-0.0950246, 0.0012707]	0.056
BMI_2*‡	-0.0745811 [-0.1189969, -0.0301652]	0.001
BMI_3*‡	0.0829741 [0.0393008, 0.1266475]	<0.001
BMI_4*‡	-0.0214095 [-0.0662199, 0.023401]	0.349
rcs1§	0.46694 [0.4379549, 0.4959251]	< 0.001
rcs2§	-0.1295396 [-0.1515924, -0.1074868]	< 0.001
rcs3§	-0.0278038 [-0.045251, -0.0103567]	0.002
$rcsHbA_{1c}1$‖	-0.0820199 [-0.154833, -0.0068254]	0.032
$rcsHbA_{1c}2$‖	0.0623191 [0.0097391, 0.1148992]	0.020
Intercept	-0.9035062 [-1.003537, -0.8034755]	

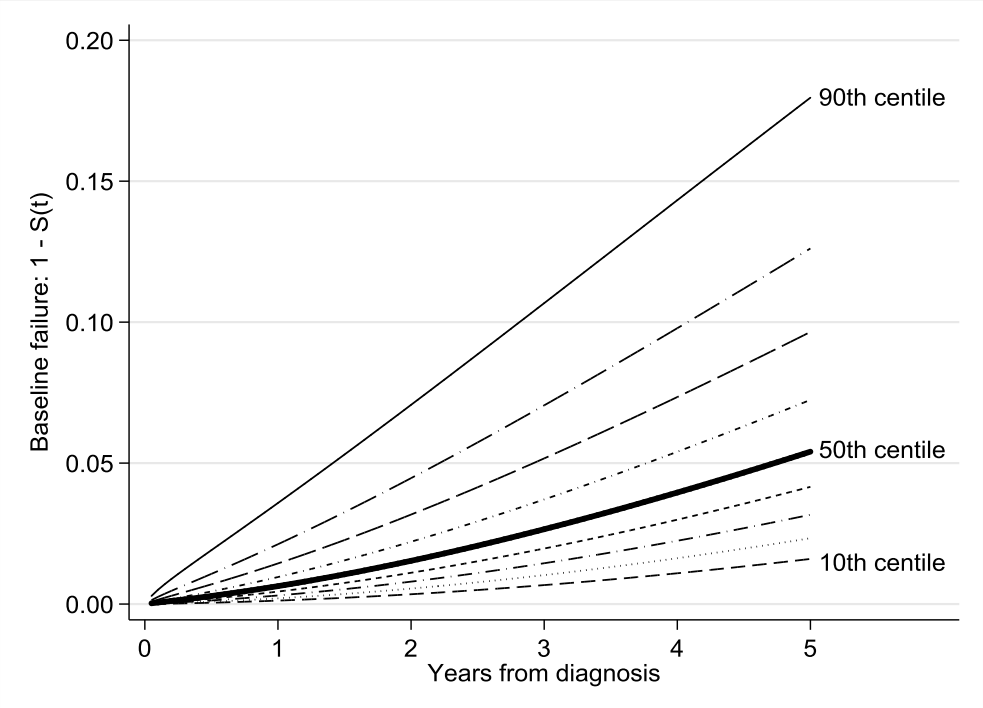
Supplementary Table 7: Effects of removing variables on explained variation (R^2_D). The effect of removing one covariate singularly is shown in the second and third columns. The effect of removing covariates cumulatively in descending order of importance is shown in the fourth and fifth columns. The effect of removing covariates cumulatively in ascending order of importance is shown in the sixth and seventh columns.

Variable removed	Single	Cumulative (greatest first)	Cumulative (least first)	
	R^2_D	R^2_D	R^2_D	order
Full model	0.19	0.19	0.19	
HbA_{1c}	0.09	0.09		
Age at diagnosis	0.15	0.04	0.12	4
HDL	0.19	0.00	0.18	3
BMI	0.19	0.00	0.18	2
Sex (male)	0.19	-	0.19	1

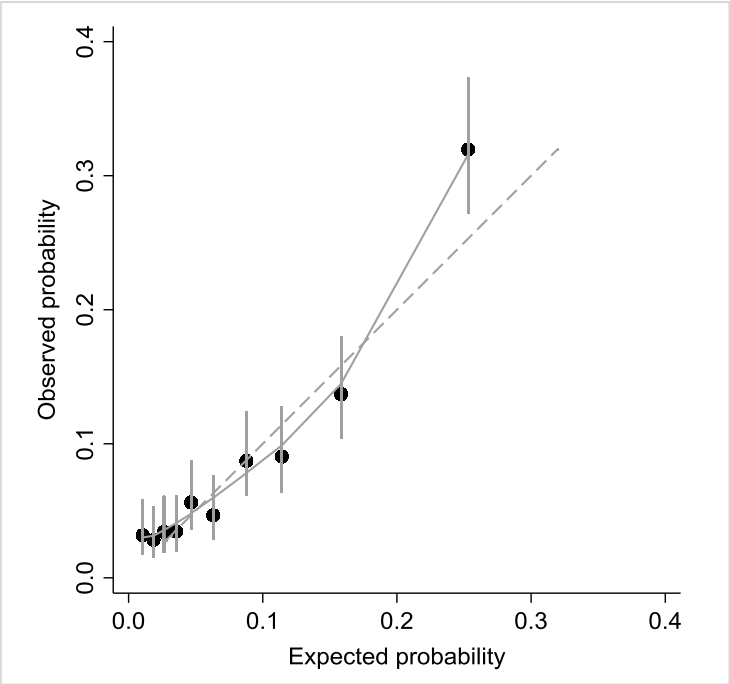
Supplementary Figure 4: Distribution of predicted failure probabilities of all patients in development data at 5 years



Supplementary Figure 5: Failure probabilities (first five years from diagnosis) at the 10th, 20th …90th deciles of the prognostic index (linear predictor). 10th centile (low risk) is the lowermost dashed line, 90th centile (high risk) is the uppermost solid line. The bold solid line represents 50th centile.



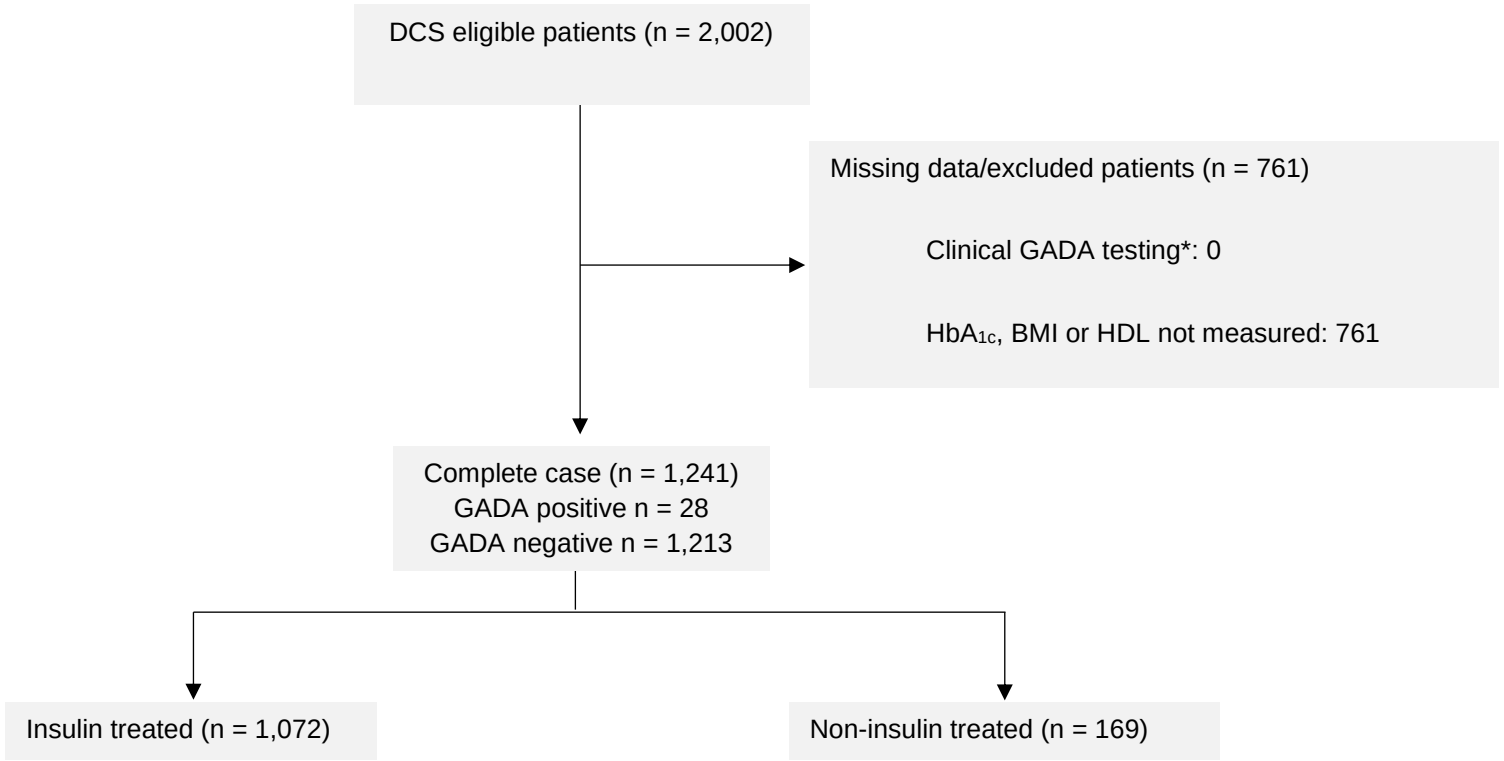
Supplementary Figure 6: GoDarts internal validation calibration plot plot of expected versus observed failure probabilities at t = 5 years.
Dashed grey line is reference line where observed = expected probabilities. Black filled circles are risk groups using deciles of expected probabilities, vertical grey solid lines are 95% CIs. Grey solid line is lowess smoother.



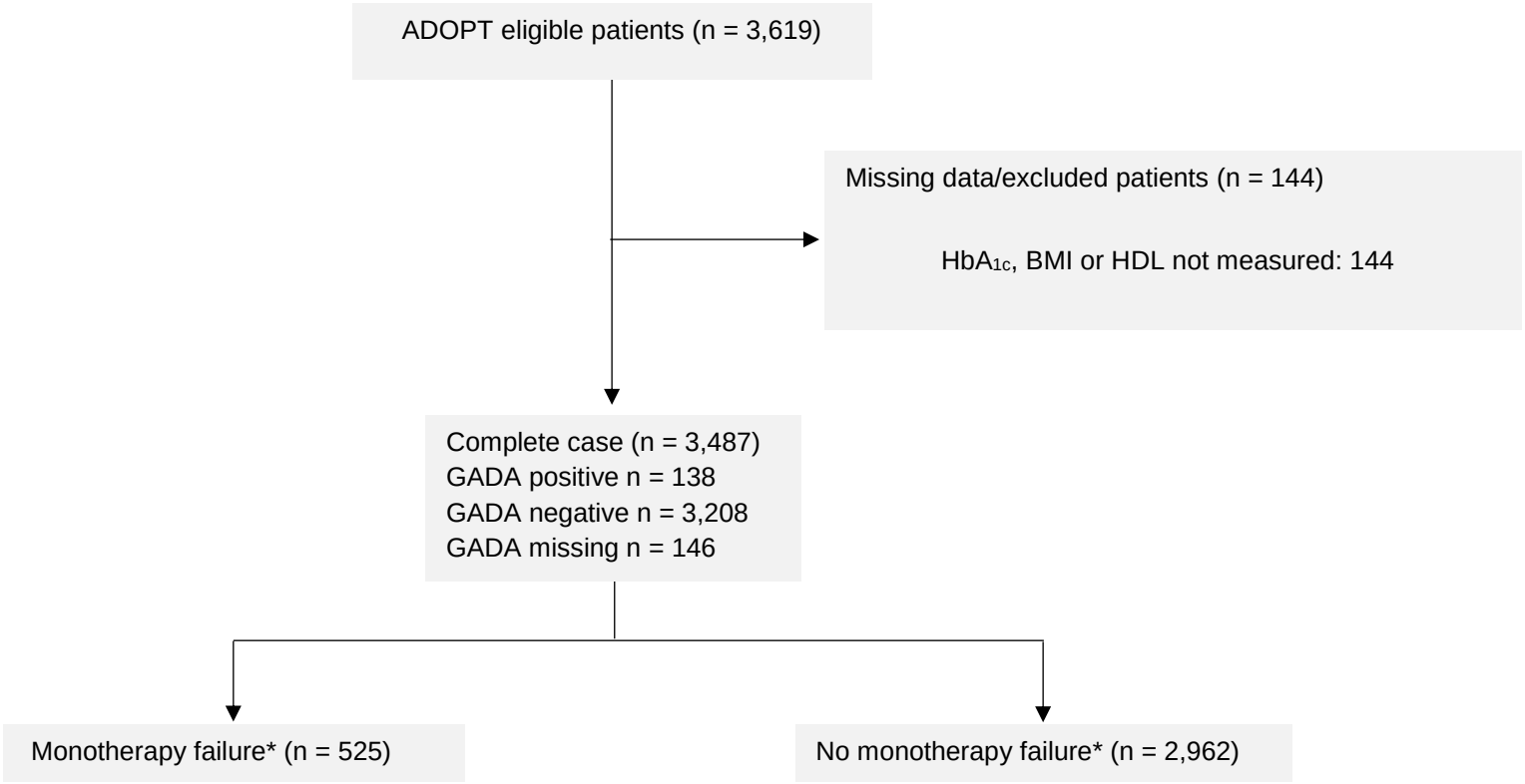
Supplementary Table 8 Model performance results for the internal validation

Performance parameter	Internal validation (1,000 bootstrap)		
	Apparent (SD)	Test (SD)	Optimism
Explained variation (R^2_D) (1)	0.199 (0.021)	0.191 (0.009)	0.008
ROC AUC (5 yr)	0.741 (0.016)	0.736 (0.002)	0.005

Supplementary Figure 7: Participant flow diagram (DCS external validation cohort) ***** identified through search of electronic laboratory records.



Supplementary Figure 8: Patient flow diagram (ADOPT external validation cohort). * Failures in five years.



Supplementary Table 9: Patient characteristics for GADA positive development cohorts. Median (IQR) or %
* At first visit
†Percentage of patients observed for at least five years
‡ measured < 6 months post diagnosis
§ Closest to diagnosis (within 12 months pre or post diagnosis)
|| Centile of participants with type 1 diabetes from the Wellcome Trust Case Control Consortium.

	GoDarts Development (n = 131)
Sex (% Male)	48.9%
Age at diagnosis (years)	62 (56, 70)
BMI (kg/m^2)*	28.7 (25.7, 32.7)
Duration of diabetes (years) at latest follow up	12.5 (10.2, 14.9)
Failure within 5 years (%)†	30.2%
HbA_{1c} (%)‡	8.6 (6.9, 10.8)
HbA_{1c} (mmol/mol) ‡	71.0 (52.0, 95.0)
HDL (mmol/L) §	1.2 (1.0, 1.5)
Duration of diabetes (years) at GADA	4.9 (2.1, 6.9)
T1D GRS||	6.8 (1.1, 37.5)

Supplementary Table 10: Hazard Ratios of the four covariates in the GADA positive model (hazard scale with 1 d.f.) for time to insulin. * Centered variables. † Log-transformed. Full model including derived spline variables for the baseline normal cumulative hazard, derived spline variables for the time-dependent effect of HbA_{1c} and intercept for model replication purpose is shown in separate table.

Variable	**Hazard Ratio [95% CI]**	**P value**
HbA_{1c} (mmol/mol)*†	3.55 [1.68, 7.47]	0.001
Age at diagnosis (years)* †	0.07 [0.02, 0.32]	0.001
Sex (male)	0.58 [0.34, 0.96]	0.036
BMI* †	0.06 [0.02, 0.22]	<0.001

Supplementary Table 11: Model coefficients for the GADA positive model (hazard scale with 1 d.f.) for model replication purpose.
* Centered variables. † Log-transformed. §Derived spline variables for the baseline hazard cumulative hazard. ‖Derived spline variables for the time-dependent effect of age at diagnosis. HbA_{1c} is centered on 75 mmol/mol, Age at diagnosis is centered on 61 years, BMI is centered on 29 kg/m^2.

Variable	Beta coefficient [95% CI]	P value
HbA_{1c} (mmol/mol)*†	1.266847 [0.5230154, 2.010679]	0.001
Age at diagnosis (years)* †	-2.605545 [-4.086671, -1.124419]	0.001
Sex (male)	-0.5520431 [-1.066946, -0.0371403]	0.036
BMI (kg/m^2) *†	-2.859483 [-4.181024, -1.507469]	< 0.001
rcs1§	1.028493 [0.8105269, 1.246459]	< 0.001
rcsAge1‖	-1.521259 [-2.796028, -0.2464909]	0.019
rcsAge2‖	0.153941 [-0.5701638, 0.8780459]	0.677
Intercept	-0.9048484 [-1.257163, -0.5525342]	

Supplementary Table 12: Effects of removing variables on explained variation (R^2_D) on the GADA positive model. The effect of removing one covariate singularly is shown in the second and third columns. The effect of removing covariates cumulatively in descending order of importance is shown in the fourth and fifth columns. The effect of removing covariates cumulatively in ascending order of importance is shown in the sixth and seventh columns.

Variable removed	Single	Cumulative (greatest first)	Cumulative (least first)	
	R^2_D	R^2_D	R^2_D	order
Full model	0.326	0.326	0.326	
HbA_{1c}	0.249	0.249		4
BMI	0.203	0.059	0.161	3
Age at diagnosis	0.229	0.006	0.232	2
Sex (male)	0.273	-	0.273	1

Supplementary Table 13: Model performance results for the internal validation (GADA positive model)

Performance parameter	Internal validation (1,000 bootstrap)		
	Apparent (SD)	Test (SD)	Optimism
Explained variation (R^2_D) (1)	0.355 (0.099)	0.328 (0.035)	0.027
ROC AUC (5 yr)	0.809 (0.040)	0.792 (0.013)	0.017

Supplementary Table 14: Patient characteristics for GADA negative cohorts. Median (IQR) or %
* At first visit
†Percentage of patients observed for at least five years
‡ measured < 6 months post diagnosis
§ Closest to diagnosis (within 12 months pre or post diagnosis)
|| Not followed post failure.

	GoDarts Development (n = 3,101)	DCS Validation (n = 1,213)	ADOPT Validation (n = 3,208)
Sex (% Male)	54.8%	54.3%	59.2%
Age at diagnosis (years)	62 (54, 69)	61 (54, 67)	58 (51, 65)
BMI (kg/m^2)*	30.5 (27.2, 34.8)	29.5 (26.8, 33.2)	31.0 (27.8, 35.3)
Duration of diabetes (years) at latest follow up	12.3 (9.9, 15.0)	6.3 (4.0, 10.2)	Not Available||
Insulin treated within 5 years (%)†	7.9%	7.1%	14.9%
HbA_{1c} (%)‡	7.6 (6.6, 9.4)	6.6 (6.1, 7.5)	7.2 (6.7, 7.8)
HbA_{1c} (mmol/mol) ‡	60.0 (49.0, 79.0)	48.6 (43.2, 58.5)	55.2 (49.7, 61.7)
HDL (mmol/L) §	1.15 (0.99, 1.36)	1.15 (0.99, 1.36)	1.2 (1.0, 1.4)
Duration of diabetes (years) at GADA	4.5 (2.4, 7.0)	7.2 (4.7, 11.4)	1.0 (0.0, 1.0)

Supplementary Table 15: Hazard Ratios of the four covariates in the GADA negative model (odds scale with 2 d.f.) for time to insulin. * Centered variables. † Log-transformed. Full model including derived spline variables for the baseline odds cumulative hazard and intercept for model replication purpose is shown in separate table.

Variable	Hazard Ratio [95% CI]	P value
HbA_{1c} (mmol/mol)*†	5.52 [4.26, 7.15]	<0.001
Age at diagnosis (years)*	0.96 [0.95, 0.97]	<0.001
Sex (male)	0.77 [0.65, 0.92]	0.004
HDL (mmol/L) * †§	0.53 [0.38, 0.73]	<0.001

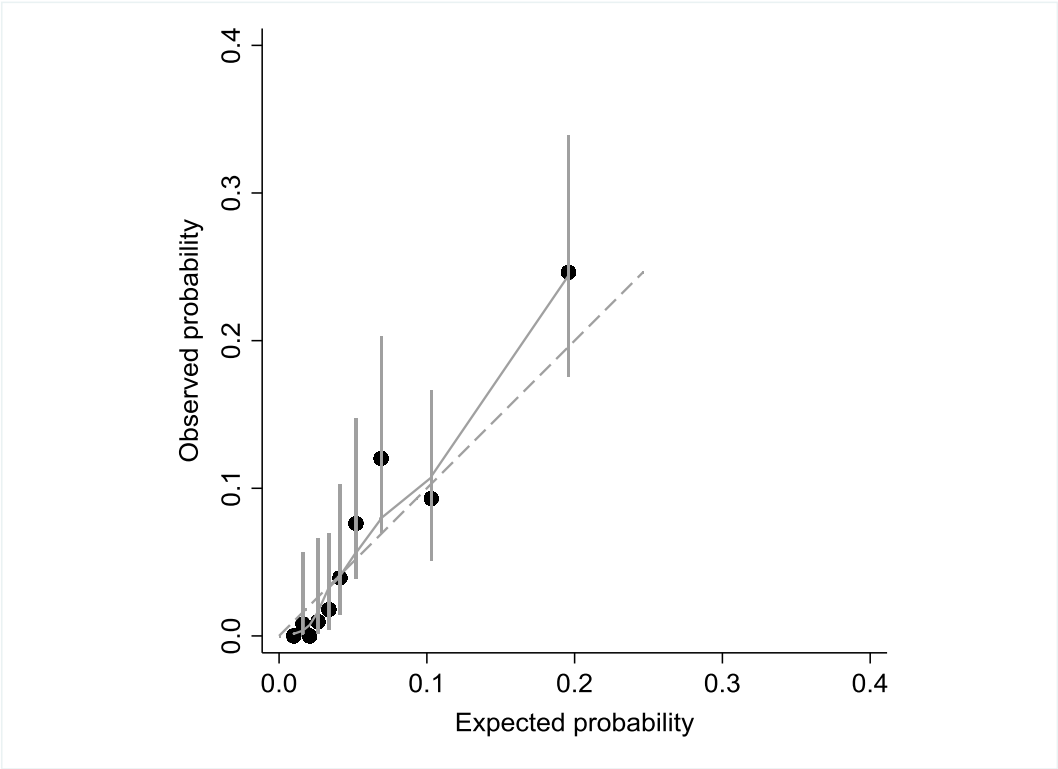
Supplementary Table 16: Model coefficients for the RP GADA negative model (odds scale with 2 d.f.) for model replication purpose.
* Centered variables. † Log-transformed. ‡Closest to diagnosis. §Derived spline variables for the baseline normal cumulative hazard. ‖Derived spline variables for the time-dependent effect of HbA_{1c}. HbA_{1c} is centered on 66 mmol/mol (8.2%), Age centered on 61 years and HDL on 1.2 mml/mol.

Variable	**Beta coefficient [95% CI]**	**P value**
HbA_{1c} (mmol/mol)*†	1.707518 [1.448497, 1.966538]	<0.001
Age at diagnosis (years)*	-0.0395409 [-0.048312, -0.0307698]	<0.001
Sex (male)	-0.2552507 [-0.4289788, -0.0815225]	0.004
HDL (mmol/L) *†‡	-0.6422273 [-0.9718084, -0.3126463]	< 0.001
rcs1§	0.9315873 [0.8557963, 1.007378]	< 0.001
rcs2§	-0.1583448 [-0.2177518, -0.0989377]	< 0.001
$rcsHbA_{1c}$‖	-0.3280567 [-0.5007058, -0.1554076]	< 0.001
$rcsHbA_{1c}$‖	0.0955222 [-0.0272555, 0.2182999]	0.127
Intercept	-1.580503 [-1.711754, -1.449251]	

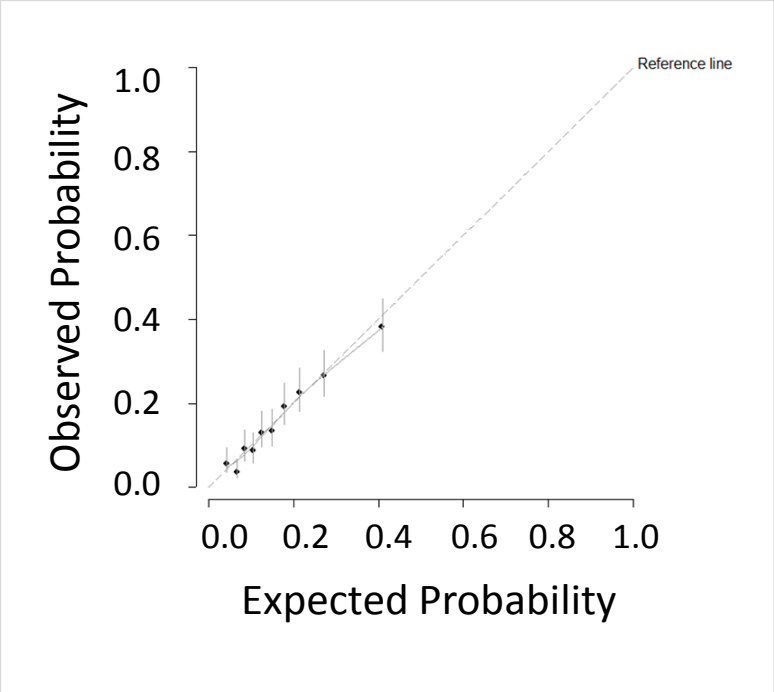
Supplementary Table 17: Model performance results for the internal validation (GADA negative model)

Performance parameter	**Internal validation (1,000 bootstrap)**		
	Apparent (SD)	**Test (SD)**	**Optimism**
Explained variation (R^2_D) (1)	0.219 (0.025)	0.215 (0.015)	0.004
ROC AUC (5 yr)	0.725 (0.017)	0.724 (0.001)	0.001

Supplementary Figure 9: DCS external validation (GADA negative model) calibration plot of expected versus observed failure probabilities at t = 5 years. Dashed grey line is reference line where observed = expected probabilities. Black filled cricles are risk groups using deciles of expected probabilities, vertical grey solid lines are 95% CIs. Grey solid line is lowess smoother.



Supplementary Figure 10: ADOPT external validation (GADA negative model) calibration plot of expected versus observed failure probabilities at t = 4 years. Dashed grey line is reference line where observed = expected probabilities. Black filled cricles are risk groups using deciles of expected probabilities, vertical grey solid lines are 95% CIs. Grey solid line is lowess smoother.



Supplementary references

1. Royston P, Sauerbrei W: A new measure of prognostic separation in survival data. Statistics in medicine 2004;23:723-748

Chapter 6.

Discussion

The work presented in this thesis has investigated the development of clinical prediction models to assist with the classification and care of patients diagnosed with diabetes in clinical practice.

The first two studies of this thesis investigate the development and validation of a diagnostic model for identifying type 1 diabetes requiring rapid insulin therapy in young adults. We first developed a multivariable diagnostic model combining five clinical features and biomarkers (age of diagnosis, BMI, GADA and IA-2 islet-autoantibodies, T1D GRS) using logistic regression. Performance was assessed using both internal and external validation; the results indicated that the model had high discrimination and calibration ability. In the next study, we used the same dataset to assess if machine learning would have superior performance over logistic regression in this setting. We built comparative models using five commonly used supervised machine learning algorithms (Gradient Boosting Machine, Support Vector Machine, K-Nearest Neighbours, Neural Network and Random Forest) and compared their performance to that of logistic regression. In this setting, there was no performance gain in using machine learning.

The remaining studies investigated glycaemic deterioration in patients diagnosed with type 2 diabetes. We first discovered that a genetic risk score (T1D GRS) can be used to identify patients with rapid glycaemic deterioration requiring insulin treatment over and above GADA testing. We went on to develop and externally validate a multivariable prognostic model built using Royston-Parmar flexible parametric survival analysis (RP) to identify patients with a high risk of rapid progression. Performance of the model was modest with baseline HbA_{1c} explaining most of the variation.

The remainder of this chapter gives an overview of the main findings of this thesis and discusses the work's conclusions, implications, limitations and potential areas for further research.

Discussion of chapter 2: Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18 to 50

Misclassification of diabetes subtype is common particularly in young adult patients where, due to increasing rates of obesity, discriminating between type 1 and young-onset type 2 diabetes can be challenging. Current guidance on diabetes classification at diagnosis focuses on aetiopathological definitions rather than the patient's treatment requirements, with no clear criteria for use in clinical practice. There is no single diagnostic test that can robustly classify diabetes at diagnosis and no clinical prediction models are available to assist clinical decision making.

In this study we developed and validated a diagnostic model to classify type 1 diabetes at diagnosis using a robust definition based on the requirement for rapid insulin therapy.

Conclusions

A diagnostic model combining clinical features and biomarkers has a higher accuracy for identifying type 1 diabetes with rapid insulin requirement than using single features in isolation.

Implication of findings

This study delivers a diagnostic model that has the potential to be used in clinical practice to assist clinicians to accurately identify patients with type 1 diabetes requiring rapid insulin therapy and to help reduce diabetes misclassification.

The development of multiple models with different combinations of the five predictors means that the model still retains utility in situations where autoantibody and/or genetic testing is either not indicated or not available; the model can be still be used in at least one form. For example, genetic but not autoantibody data is available in many biobanks but in clinical care, genetic testing is not yet routinely performed. This development approach allows a staged approach to classification of diabetes; the clinical features-only model can be used to identify patients with diagnostic uncertainty who may benefit most from additional testing without incurring any financial cost.

In addition to aiding clinical decision-making, the model could facilitate a triage-based approach to diabetes subtype diagnosis; probabilities derived on clinical features alone could be used as criteria for requesting autoantibody or genetic testing. It could also be used as a tool for evidence-based classification in diabetes research where it could be incorporated into the participant selection process.

The model is presented in a website (beta version available at https://www.diabetesgenes.org/t1dt2d-prediction-model/) (Figure 1) which provides predictions based on user input predictor values (R code provided in Appendix 1).

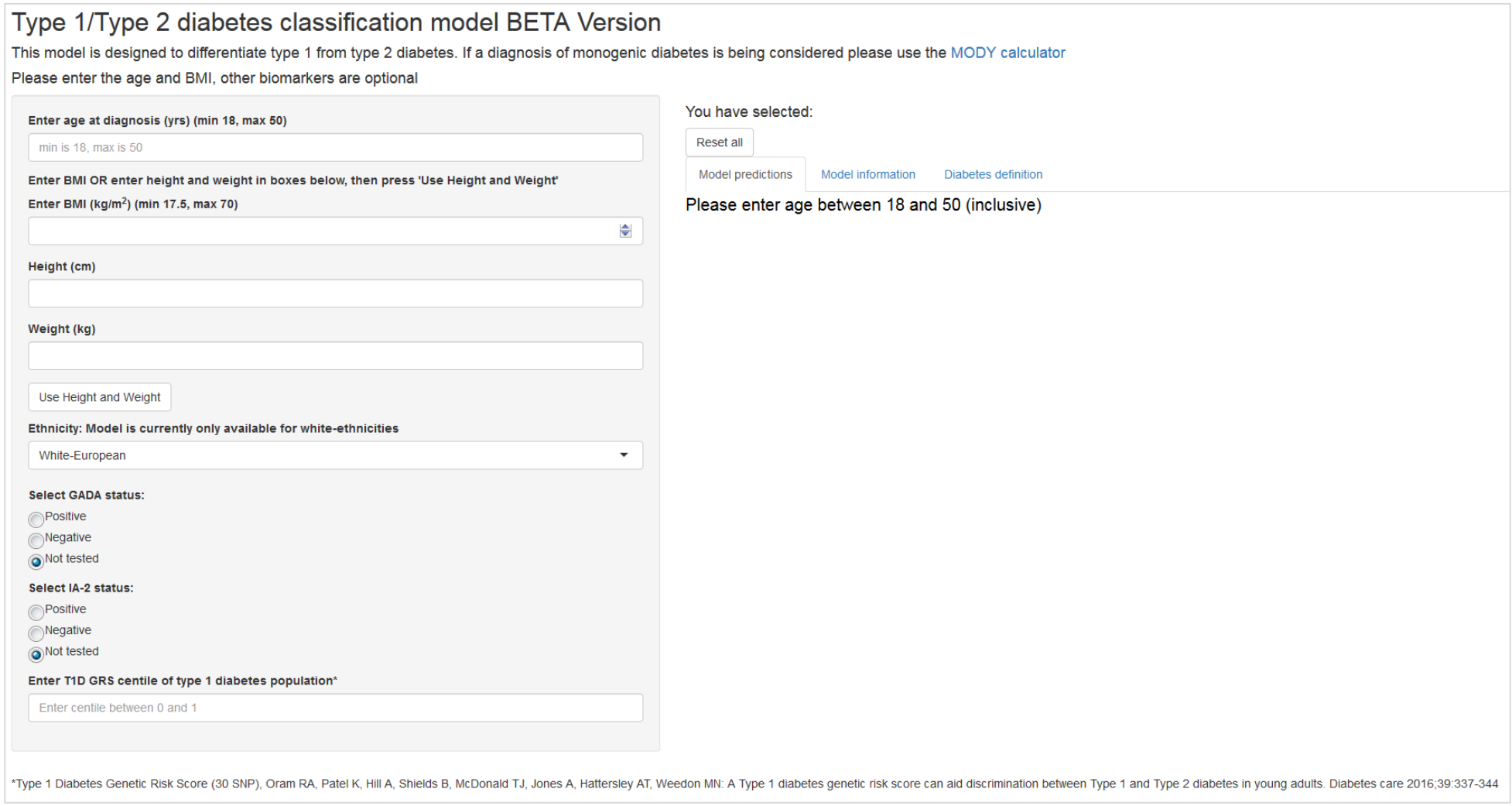


Figure 1: Classification model web calculator

Subsequent work

There have not to our knowledge been any other published studies proposing a diagnostic model for classifying type 1 and type 2 diabetes at diagnosis. A recent study examining the frequency of type 1 diabetes has however highlighted the need for improved diabetes classification in older adults (1).

A clustering algorithm comprising of five diabetes subgroups has recently been published; one cluster being defined by the presence of GADA positivity only regardless of other features (severe autoimmune diabetes (SAID) cluster), and the other four based on GADA negativity and differences in age at diagnosis, BMI, HbA_{1c} and HOMA 2 (type 2-like clusters) (2) but did not show that the clusters could be used to inform treatment decisions (3). Whilst this cluster model has limitations, it does identify a future direction for diabetes prediction models where the focus shifts from classification of diabetes to predicting other aspects of the disease such as complications and treatment responses.

A new improved T1D GRS has been published subsequent to our study (T1D GRS2) (4). The new T1D GRS2 includes 67 SNP's compared to 30 in T1D GRS and has greater performance. An area of future work would be to update the model with the new T1D GRS2.

We followed up this study in Chapter 3 to compare the performance of machine learning algorithms to that of logistic regression.

Limitations

The limitations of this work are predominantly related to the use of existing cross-sectional data to build the models: ideally we would have carried out a new study allowing us to use predictors measured at diagnosis and follow-up

data to assess development of severe insulin deficiency and insulin requirement.

The use of existing data meant that we were limited to modelling only the features that were available in the datasets. The cross-sectional nature of the data meant that predictors that present at diagnosis such as presentation glycaemia, ketosis, or weight loss were not available in the datasets. It also meant that predictor variables were for most participants measured some years post diagnosis; since BMI and islet-autoantibodies change over time in adult onset diabetes (5), the model predictions are likely to be under-estimated. Data such as date of diagnosis and time to insulin were self-reported by the patient rather than obtained from patient medical records which may have introduced error when assigning the outcome.

There are two limitations connected to the use of C-peptide as our gold standard outcome. Firstly, an issue with stored samples for participants in the DARE cohort prior to February 2010 (when immediate freezing of aliquoted samples was introduced) restricted the availability of C-peptide data in DARE participants before this date: sample degradation and poor sample collection can cause falsely low values (6). Secondly, we did not consider renal impairment in the participants, which may cause falsely high values (7).

Both GADA and IA-2 titres (concentrations) were dichotomised for use as predictors in the model. We did not use them as continuous predictors because the rounding of titres lower or above the levels of assay detection causes peaks at either end in the continuous distribution and because titre values are not normally available in clinical practice; results are normally reported simply as either positive or negative.

The datasets used to develop the model predominantly consisted of white-European participants, which meant that we did not have sufficient data to include ethnicity as a predictor in the model. The model does not therefore reflect differences in prevalence in certain ethnic subgroups (8) and restricts the use of the model to a white-European population. Another limitation related to the use of the model is that is unsuitable for extrapolation beyond the ages of 18 and 50 years.

Finally, we were unable to externally validate all combinations of the model as IA-2 islet-autoantibodies and T1D GRS were not available in the external dataset.

Future research

The main direction of future research is to implement the model into clinical practice. This will involve additional validation and expanding the use of the model into other ethnicities and age groups.

Important areas of future research are updating the existing predictor coefficients using data measured at diagnosis and assessing the performance of the model with the inclusion of additional features available only at diagnosis. This data will be available in the Getting the Right Classification and Treatment From Diagnosis in Adults With Diabetes (StartRight) study (9). StartRight is a new prospective observational study of newly diagnosed adult participants designed to assess the relationship of clinical features and biomarkers to diabetes subtype.

There is an opportunity to perform prospective external validations using the StartRight study (9) and United Kingdom prospective diabetes study (UKPDS) (10). UK Biobank is a valuable new source of data for future validation or

updating of the model. UK Biobank is a long-term national project to build a detailed resource for health researchers consisting of data and stored samples on more than 500,000 UK volunteers aged 40-69 years when recruited (11). It now contains GP medical records and there are plans to include C-peptide measurement.

An assessment of the clinical usefulness of the model would be an interesting addition to this study and would be useful for implementing the model into clinical use. Follow up research could include a health economic and implementation study (12) to evaluate the impact of the model in clinical use. For example, a study to evaluate differences in classification or outcomes could involve comparing decisions made using the model predictions versus clinician based decisions.

Discussion of chapter 3: Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the classification of type 1 and type 2 diabetes in young adults

The previous chapter suggested that a diagnostic model to classify type 1 diabetes at diagnosis built using logistic regression achieved good performance. In this study we compared the discrimination and calibration performance of five machine learning algorithms to logistic regression using our diabetes classification model from the previous chapter as an example.

Conclusions

Optimised machine algorithms performed no better than logistic regression to classify type 1 and type 2 diabetes in young adults.

Implication of findings

A recent systematic review study found no evidence of superior performance of machine learning over logistic regression and concluded that improvements in methodology and reporting of comparison studies are needed (13). In our study we demonstrated the application of a methodological approach and provided our code to allow our approach to be replicated in future comparison studies (Appendix 2).

This study demonstrates the utility of comparing machine learning to traditional regression modelling when developing and selecting clinical prediction models, and re-iterates the need to validate models on external data. In our diabetes setting, we provided confidence that machine learning would not have yielded better performance than that achieved using logistic regression.

Subsequent work

There are no studies that have published a performance comparison of a diabetes classification model.

Limitations

An essential aspect of this model comparison study was the external validation, which we would ideally have carried out using with all five predictors. It is a severe limitation that the lack of IA-2 and T1D GRS in the external dataset meant that the only external validation that we were able to perform involved using just three of the predictor variables: Age at diagnosis, BMI and GADA. This is a very small number of predictors for machine learning algorithms even in medicine (between 5 and 20 predictors is more relevant). Machine learning is generally associated with processing large numbers of predictors but in medicine, with the possible exception of image data, a few meaningful predictors is more common. It is possible that a comparison of a model comprising of more variables and the use of a larger sample size might have given enough power to the machine learning for it to outperform logistic regression.

Furthermore, logistic regression may have been slightly disadvantaged in our comparison by not considering non-linearity.

Future research

The methodological approach that we have applied to our study could be used as a framework for independent researchers to externally validate other studies that have performed similar comparisons. In particular, it would be interesting to

use our framework to examine other comparison studies identified as having a risk of bias in the recent systematic review (13).

Future research to address the main limitations of this study could include adding more predictors and an assessment of a more flexible logistic regression model with splines or fractional polynomials. A logistic regression model with interactions would also be of interest especially if the number of predictors were increased. The effect of the use of SMOTE for class imbalance on risk estimates is an important issue that could also be addressed in future research.

Discussion of chapter 4: A Type 1 Diabetes Genetic Risk Score can identify patients with GAD65 autoantibody positive type 2 diabetes that rapidly progress to insulin therapy

The rate of glycaemic progression in patients with clinically diagnosed type 2 diabetes is highly variable. There may be clinical utility in identifying patients who are most likely to rapidly progress to requiring insulin therapy, enabling clinicians to prioritise high risk patients for more frequent monitoring and treatment escalation.

GADA has been associated with rapid glycaemic deterioration, but the predictive value of this test is limited in patients with type 2 diabetes. Previous research has suggested that type 1 diabetes genetic variants in the HLA region are associated with rapid progression to insulin therapy in patients with clinically diagnosed type 2 diabetes who are positive GADA (14).

We used survival analysis to investigate if a diagnostic test for type 1 diabetes genetic variants (T1D GRS) could identify rapid progression to insulin therapy in adult patients with a clinical diagnosis of type 2 diabetes over and above GADA testing.

Conclusions

We found that participants who were GADA positive and had a high T1D GRS progressed to insulin therapy more rapidly than the other GADA positive participants. There was no difference in the rate of progression by T1D GRS in participants who were GADA negative.

Our finding that GADA is associated with time to insulin in patients clinically diagnosed with type 2 diabetes is consistent with previous studies (15, 16);

what we add to previous knowledge is our finding that T1D GRS is independently associated with time to insulin in this population, but only in the presence of GADA.

Implication of findings

T1D GRS alters the clinical implications of a positive GADA test when predicting time to insulin requirement in patients with a clinical diagnosis of type 2 diabetes. There is no prognostic value in genetic testing for patients who are GADA negative; genetic testing should be indicated in clinical practice only for patients who are GADA positive to more accurately assess their risk of requiring rapid insulin therapy. The use of this two-step testing approach may facilitate a precision medicine approach to treating patients diagnosed with type 2 diabetes. Our findings could also be applied to participant selection for future type 2 diabetes clinical trials investigating immune intervention or other interventions to slow progression.

Whilst original reports of the T1D GRS focused on aiding discrimination between type 1 and type 2 diabetes (17) it has also been used in subsequent studies for other applications such as discriminating monogenic and type 1 diabetes (18) and predicting progression of islet-autoimmunity (19). The use of T1D GRS in our study adds another novel practical application of its use to the literature: to assist identification of patients diagnosed with type 2 diabetes who will require early insulin therapy over and above GADA testing.

Our results support the findings of a recent study that suggested the presence of GADA in patients clinically diagnosed with type 2 diabetes is indicative of two heterogeneous populations with very different phenotypes (20). The first subtype is an autoimmune late onset type 1 diabetes: patients have a genetic

susceptibility to type 1 diabetes and require rapid insulin therapy. The second is a non-autoimmune diabetes: the GADA result is a false positive which will be common with islet-autoantibody testing in low prior prevalence populations (type 2 diabetes not requiring initial insulin).

The presence of genetic susceptibility to type 1 diabetes may increase the likelihood that a patient who is GADA positive has true underlying type 1 diabetes, rather than being a false positive result.

Subsequent work

Although this study has been cited by studies relating to the use of genetics in diabetes (21, 22), there have not to our knowledge been any other published studies investigating the use of GADA and T1D GRS to identify patients diagnosed with type 2 diabetes requiring early insulin therapy.

We followed up this study by using GADA and T1D GRS in the development of a multivariable prognostic model to predict rapid insulin requirement in adult patients diagnosed with type 2 diabetes requiring early insulin therapy (chapter 5).

Limitations

The main limitation of this study is our reliance on using initiation of insulin therapy based on a clinical decision rather than trial protocol to define the endpoint. This is problematic because there is likely to be variation in the decision to initiate insulin therapy between both clinician and patient which may have introduced inertia bias. Clinicians were unaware of the patients T1D GRS and immunology test results at the date of treatment decision so systematic bias is unlikely. Actual clinical requirement for insulin was not known and we

were unable to distinguish between relative or absolute requirement for insulin as C-peptide was not routinely measured in this cohort.

In the Cox proportional hazards regression analysis, we discovered a statistically significant association between year of diagnosis and time to insulin consistent with previous studies (23) reflecting changes in prescribing patterns over time, specifically an increasing delay of insulin therapy initiation (24). Our results were not adjusted for year of diagnosis but this finding may be an important consideration in future work.

There were several limitations relating to the use of Exeter-based cohorts. Firstly, the use of self-reported time to insulin in the Exeter based cohorts may have introduced imprecision. Our use of complete-case analysis meant that we excluded a large number of DARE participants because GADA was missing; GADA testing was only performed in the study for participants who were younger at diagnosis and stored serum was not available for all participants meaning that we could not perform additional GADA testing for all participants where GADA was missing. Most of the 3,542 participants excluded from our analysis were from the DARE cohort, there were statistically significant differences in the clinical features of these excluded participants but they were not considered clinically relevant.

To achieve sufficient numbers, we had to combine several cohorts from different studies - ideally we would have used a single cohort. There were two limitations related to the use of a combined dataset; statistically significant differences in GADA prevalence, diabetes duration and HbA_{1c} between cohorts were evident and survival distributions differed between studies. We dealt with

these limitations by including study of origin as a strata variable in the Cox proportional hazards regression analysis.

Another limitation is the measurement of GADA post diagnosis. This limitation has been discussed earlier in this discussion in relation to chapter 2 and is not repeated here. In addition to GADA, IA-2 has previously been associated with time to insulin in patients diagnosed with type 2 diabetes (25) but we were unable to assess the interaction between IA-2 (or any other islet-autoantibody) and T1D GRS in our study as this data was not available.

In our Cox proportional hazards regression analysis, the relationship between the continuous covariates and progression to insulin was assumed to be linear. This may be an invalid assumption and should be considered in any future research.

The implications of our findings are not generalisable to patients who are of non-white European ethnicity or are younger than 35 years at diagnosis.

Future research

There is an opportunity for this study to be repeated in a new prospective study in which GADA is measured at diabetes diagnosis and initiation of insulin is based on a trial protocol. Our study could be extended in this new prospective setting to investigate the use of other islet-autoantibodies such as IA-2 and ZnT8. The findings from a study using IA-2 would be of particular interest for LADA diagnosis which is currently based on GADA only.

A follow up study to assess whether prior likelihood of autoimmune diabetes alters the association between clinical features and biomarkers, and progression to insulin therapy would be of interest. This could be achieved by

performing survival analysis for different age and BMI subgroups. It would also be of clinical interest to examine if the interaction between GADA and T1D GRS was consistent at different ages. Our primary outcome was short term progression to insulin therapy (5 years), a secondary outcome that could be investigated in future research is long term progression to insulin in those participants who did not require insulin therapy by five years. Future research could be based on the new improved T1D GRS2 (4) which was published subsequent to our study and incorporating different centiles for GADA positivity.

An important area for future research would be to apply our findings to the development of a prognostic model to predict rapid insulin requirement in individual patients diagnosed with type 2 diabetes. The use of a prediction model combining multiple predictors including GADA and T1D GRS is likely to have the greatest utility.

Discussion of chapter 5: Predicting early insulin requirement in adults diagnosed with type 2 diabetes: development and external validation of a multivariable survival model.

Chapter 4 identified genetic susceptibility to type 1 diabetes, measured using T1D GRS, alters the implications of a positive GADA result in patients diagnosed with type 2 diabetes and could be used to identify patients at high risk of rapid progression to insulin therapy.

A previous Diabetes UK-funded Diabetes Remission Clinical Trial (DiRECT) study found other clinical features and biomarkers in addition to GADA and T1D GRS that were independently associated with progression to insulin in patients diagnosed with type 2 diabetes (23). We combine our findings from chapter 4 and findings from the DiRECT study to develop and validate a multivariable prognostic model to predict rapid insulin requirement from diagnosis in individual adult participants diagnosed with type 2 diabetes.

Conclusions

The rate of glycaemic deterioration from first diagnosis defined by requirement for insulin therapy is generally slow and fairly constant in the majority of patients diagnosed with type 2 diabetes. Prognostic models integrating clinical features and biomarkers have the potential to identify those patients at high risk of rapid progression to insulin. High HbA_{1c} measured at diagnosis is the strongest predictor of rapid progression.

Implication of findings

Identifying patients at high risk of rapid progression to insulin has utility in both clinical practice and research. In clinical practice, individual predictions of a

patient's progression to insulin can be used to optimise their treatment and set monitoring priorities. In research, patients likely to rapidly progress could be targeted to maximise the cost effectiveness of clinical trials of interventions aimed at slowing diabetes progression.

Use of the model as a triage-based tool for identifying of patients who would benefit most from GADA testing or those patients who are more likely to be GADA positive has significant clinical interest. Firstly, there is a financial benefit in testing only a minority of patients who will benefit most from additional testing. Currently one of the reasons why GADA testing is not indicated routinely for patients with type 2 diabetes is that it would be too expensive to test everyone given the huge incidence of this disease. Secondly, altering the number of GADA tests performed by only testing those patients who are more likely to be GADA positive (increasing the prior likelihood) will have the benefit of increasing the positive predicted value of the test.

Subsequent work

There have not to our knowledge been any other published studies proposing a prognostic model for identifying early insulin requirement in adults diagnosed with type 2 diabetes.

Limitations

The measuring of GADA post diagnosis in the GoDarts and DCS cohorts may have resulted in some false negatives since GADA levels decrease over time (26). This would have no impact on our main model but may have caused estimation bias in our GADA models due to the classification of participants by their GADA status (negative or positive).

The development of separate GADA models has an advantage in terms of clinical utility but resulted in a small numbers of participants available to build the GADA positive model (n = 131) and wide confidence intervals in the model estimates. In addition, we were unable to perform external validation of the GADA positive model for the same reason. We were unable to use DARE Exeter-based cohort to increase our sample size as HbA_{1c} measured at diagnosis was not available for the majority of participants.

In contrast to our findings in chapter 4, T1D GRS was not statistical significant in the GADA model when adjusted for the other clinical features and biomarkers, this may have been a power issue in this small dataset or it could be that the combined features are capturing a prior likelihood effect (identifying the GADA false positives) much better than T1D GRS in our previous study. Much larger studies would be needed to increase the number of GADA positive participants to allow the T1D GRS to be re-assessed and external validation of the model to be performed. It would be useful to investigate the effect of the new T1D GRS2 (4).

The use of separate GADA models dealt with the presence of statistically significant interactions between GADA and each of BMI, T1D GRS and HDL. Future work could investigate interactions between the continuous predictor variables and assess the impact of including any required interaction terms on model performance, the practicality of implementing a potentially complex model would also need to be assessed.

The predictor variables included in the models are based on features and biomarkers which are routinely measured or inexpensive to measure in the U.K. The models may have a different utility outside the U.K. where clinical practice

and availability of the tests are likely to vary. The findings of this study can only be applied to participants of white- European ethnicity.

The main limitation of our study is that insulin initiation was based on clinical decision making rather than a trial protocol. There is uncontrollable extraneous variability in both the time of diagnosis and the length of time before insulin initiation and therefore neither the start nor the endpoint of the survival period is fixed in relation to any underlying progress of the disease. Many patients with type 2 diabetes are diagnosed whilst in the early stages of the disease by routine testing; otherwise, because hyperglycaemia develops gradually, a patient may go undiagnosed for many years before experiencing classic diabetes symptoms.

In clinical practice, clinical inertia may affect insulin initiation decisions (27, 28); decisions to initiate insulin therapy may also be influenced by factors other than high HbA_{1c} (29). In patient-centred care approaches, there will inevitably be between patient variations in the decision to start insulin with many patients having a strong preference to avoid insulin initiation (30, 31). There may be also be between clinician/practice variability in the prescribing patterns (32) and/or compliance to the HbA_{1c} level guidelines at which insulin is initiated (31, 33). We did not have HbA_{1c} measured at time of insulin initiation so we were unable to check if insulin was initiated according to guidelines.

We assessed the performance of the model in ADOPT clinical trial data (34); this dataset had the advantage of an outcome event defined using a trial protocol but the disadvantage of being based on a different diabetes outcome (monotherapy failure). Ideally we would have developed and validated the

model using a prospective trial where biomarkers could be measured at time of diagnosis and insulin initiation.

Drug therapy prior to insulin initiation was not available in the datasets so we were unable to adjust for the adequacy of glycaemic control in different therapies (32, 35-39). Number of visits and inadequate monitoring (40) may have affected time to insulin initiation but we did not have the data to check this.

At a population level, glycaemic control is improving over time (41) with the time to insulin initiation increasing (24). This may be explained by the introduction of newer oral agents over time (24). All of these will have an impact on the predictions when using time to insulin initiation as the outcome. Our finding that earlier year of diagnosis was associated with higher rates of glycaemic deterioration was consistent with previous studies (16, 23). Year of diagnosis was not adjusted for in our model since its practical implementation would have been difficult. Our failure to adjust for calendar year of diagnosis in the model may have resulted in over-estimated predictions. This limitation should be considered when implementing the model for clinical use, possibly by applying periodic adjustments.

We decided to use RP to develop our model as it is a preferred approach in situations where individual predictions are required and there is a need to incorporate the time-dependent effects (42, 43). We encountered limitations in the usability of the model arising from the inclusion of time-dependent effects; the beta-coefficients are difficult to interpret and publishing the model is more complex.

Data was not available for other islet-autoantibodies in our development dataset (GoDarts) so we were restricted to using GADA. It would be interesting to

incorporate IA-2 and/or ZnT8 into the models as this is likely to increase performance but their inclusion would likely involve re-designing the model.

674 patients in our development cohort died before they had progressed to insulin therapy; we did not investigate the impact or adjust for competing risks in our model which means that the model coefficients may be over-stated. A competing risks analysis would be an interesting area for future research.

Future research

The Innovative Medicines Initiative Diabetes Research on Patient Stratification (DiRECT) study (44) may be a potentially valuable resource for future research aimed at either updating or validating the model although its usefulness is limited by a relatively short follow up time. DiRECT is a large study collecting biomarkers associated with glycaemic deterioration in participants recently diagnosed with type 2 diabetes. Participants are recruited close to diagnosis and followed-up for between 18 and 36 months, clinical features and biomarkers are collected at baseline and repeated at two follow up visits. A further area of research would be to use DiRECT or similar studies to assess whether the addition of further genetic, metabolomic or proteomic data improves the model performance over and above the simple clinical features and clinical biomarkers currently used.

A study comparing the performance and clinical utility of our model to identify patients with rapid progression to insulin to that of the type 2 diabetes subgroups derived by Ahlqvist et al in the recently published cluster analysis (2) would be interesting future research.

Another area of future research would involve using the risk predictions from our model together with health economic modelling methodologies to

investigate whether targeted interventions can be a cost-effective approach for managing type 2 diabetes outcomes. Related future research could include process evaluation of the model using clinician focus groups or questionnaires and performing decision curve analysis to describe the clinical effects of the model.

There is the option in future research to address this research question as a binary regression problem – the event outcome being insulin initiation by five years. It would be interesting to compare both the performance of the two approaches and their respective clinical uptake.

Final remarks

This thesis demonstrates that routinely measured clinical features and biomarkers can be used in multivariable prediction models to aid clinical decisions regarding the classification and care of adult patients diagnosed with diabetes.

The importance and challenges of both correctly classifying patients with diabetes according to their treatment requirements at diagnosis and identifying patients with type 2 diabetes who are likely to rapidly progress to insulin have been highlighted in this thesis: clinical prediction models to identify patients likely to develop diabetes (45) or to identify MODY (46) have been published but there are currently no published clinical prediction models that address these two challenges.

For classification of type 1 and type 2 diabetes, the studies in this thesis identify five diagnostic predictors that can be used at diagnosis, in varying combinations, to accurately predict a patient's individual risk of type 1 diabetes requiring early insulin therapy. The studies in this thesis also identify clinical features and biomarkers that can be combined to predict risk of rapid glycaemic deterioration, from diagnosis, in patients with type 2 diabetes.

This thesis is concerned with the application of statistics to develop valid clinical prediction models that can be used in clinical practice. Methodologically, the systematic approach applied to the model development and validation undertaken in this thesis has been statistically robust and has followed methodological literature and published guidelines on how to perform prediction research (47). An important focus of this thesis is the clinical utility of these models; only features that are routinely indicated in clinical practice have been

implemented in the models. Future work is now required to make these models ready for implementation into clinical practice.

The models developed in this thesis could be used to implement a personalised approach to managing patients with diabetes in clinical practice: evidence-based predictions obtained from the models can be used to inform treatment decisions in conjunction with clinical expertise. In addition, predictions could be used to implement triage-based protocols for additional islet-autoantibody of genetic testing or referrals for prioritised monitoring.

Discussion references

1. 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. The Lancet Diabetes & Endocrinology. 2018;6(2):122-9.
2. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology.
3. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. The Lancet Diabetes & Endocrinology. 2019;7(6):442-51.
4. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, et al. Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. Diabetes care. 2019;42(2):200-7.
5. Tridgell DM, Spiekerman C, Wang RS, Greenbaum CJ. Interaction of Onset and Duration of Diabetes on the Percent of GAD and IA-2 Antibody–Positive Subjects in the Type 1 Diabetes Genetics Consortium Database. Diabetes care. 2011;34(4):988.
6. McDonald TJ, Perry MH, Peake RWA, Pullan NJ, O'Connor J, Shields BM, et al. EDTA Improves Stability of Whole Blood C-Peptide and Insulin to Over 24 Hours at Room Temperature. PLOS ONE. 2012;7(7):e42084.
7. Jones AG, Hattersley AT. The clinical utility of C-peptide measurement in the care of patients with diabetes. Diabetic Medicine. 2013;30(7):803-17.
8. Golden SH, Brown A, Cauley JA, Chin MH, Gary-Webb TL, Kim C, et al. Health disparities in endocrine disorders: biological, clinical, and nonclinical factors--an Endocrine Society scientific statement. The Journal of clinical endocrinology and metabolism. 2012;97(9):E1579-639.
9. clinicalTrials.gov. Getting the Right Classification and Treatment From Diagnosis in Adults With Diabetes (StartRight) 2018 [Available from: https://clinicaltrials.gov/ct2/show/NCT03737799.
10. King P, Peacock I, Donnelly R. The UK prospective diabetes study (UKPDS): clinical and therapeutic implications for type 2 diabetes. Br J Clin Pharmacol. 1999;48(5):643-8.
11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 2015;12(3):e1001779.
12. Nilsen P. Making sense of implementation theories, models and frameworks. Implementation Science. 2015;10(1):53.
13. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019.
14. Maioli M, Pes GM, Delitala G, Puddu L, Falorni A, Tolu F, et al. Number of autoantibodies and HLA genotype, more than high titers of glutamic acid

decarboxylase autoantibodies, predict insulin dependence in latent autoimmune diabetes of adults. European journal of endocrinology. 2010;163(4):541-9.
15. Turner R, Stratton I, Horton V, Manley S, Zimmet P, Mackay IR, et al. UKPDS 25: autoantibodies to islet-cell cytoplasm and glutamic acid decarboxylase for prediction of insulin requirement in type 2 diabetes. The Lancet. 1997;350(9087):1288-93.
16. Donnelly LA, Zhou K, Doney ASF, Jennison C, Franks PW, Pearson ER. Rates of glycaemic deterioration in a real-world population with type 2 diabetes. Diabetologia. 2018;61(3):607-15.
17. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, et al. A Type 1 diabetes genetic risk score can aid discrimination between Type 1 and Type 2 diabetes in young adults. Diabetes care. 2016;39(3):337-44.
18. Patel KA, Oram RA, Flanagan SE, De Franco E, Colclough K, shepherd M, et al. Type 1 Diabetes Genetic Risk Score: a novel tool to discriminate monogenic and type 1 diabetes. Diabetes. 2016;65(7):2094-9.
19. Redondo MJ, Geyer S, Steck AK, Sharp S, Wentworth JM, Weedon MN, et al. A Type 1 Diabetes Genetic Risk Score Predicts Progression of Islet Autoimmunity and Development of Type 1 Diabetes in Individuals at Risk. Diabetes care. 2018;41(9):1887-94.
20. Jones A, McDonald T, Shields B, Hattersley A. Latent Autoimmune Diabetes of Adults (LADA) represents a mixed population of autoimmune (type 1) and non-autoimmune (type 2) diabetes rather than an intermediate phenotype. Currently in review.
21. Cerolsaletti K, Hao W, Greenbaum CJ. Genetics Coming of Age in Type 1 Diabetes. Diabetes care. 2019;42(2):189.
22. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocrine Reviews. 2019.
23. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CNA, et al. Clinical and Genetic Determinants of Progression of Type 2 Diabetes: A DIRECT Study. Diabetes care. 2014;37(3):718.
24. Kostev K, Gölz S, Scholz B-M, Kaiser M, Pscherer S. Time to Insulin Initiation in Type 2 Diabetes Patients in 2010/2011 and 2016/2017 in Germany. Journal of Diabetes Science and Technology. 2019:1932296819835196.
25. Bottazzo GF, Bosi E, Cull CA, Bonifacio E, Locatelli M, Zimmet P, et al. IA-2 antibody prevalence and risk assessment of early insulin requirement in subjects presenting with type 2 diabetes (UKPDS 71). Diabetologia. 2005;48(4):703-8.
26. Desai M, Cull CA, Horton VA, Christie MR, Bonifacio E, Lampasona V, et al. GAD autoantibodies and epitope reactivities persist after diagnosis in latent autoimmune diabetes in adults but do not predict disease progression: UKPDS 77. Diabetologia. 2007;50(10):2052-60.
27. Khunti K, Wolden ML, Thorsted BL, Andersen M, Davies MJ. Clinical inertia in people with type 2 diabetes: a retrospective cohort study of more than 80,000 people. Diabetes care. 2013;36(11):3411-7.
28. Nichols GA, Koo YH, Shah SN. Delay of insulin addition to oral combination therapy despite inadequate glycemic control: delay of insulin therapy. Journal of general internal medicine. 2007;22(4):453-8.
29. Pilla SJ, Yeh H-C, Juraschek SP, Clark JM, Maruthur NM. Predictors of Insulin Initiation in Patients with Type 2 Diabetes: An Analysis of the Look AHEAD Randomized Trial. Journal of general internal medicine. 2018;33(6):839-46.

30. Hayes RP, Bowman L, Monahan PO, Marrero DG, McHorney CA. Understanding diabetes medications from the perspective of patients with type 2 diabetes: prerequisite to medication concordance. The Diabetes educator. 2006;32(3):404-14.
31. Vaag A, Lund SS. Insulin initiation in patients with type 2 diabetes mellitus: treatment guidelines, clinical evidence and patterns of use of basal vs premixed insulin analogues. European journal of endocrinology. 2012;166(2):159-70.
32. Kostev K, Dippel F-W, Rathmann W. Predictors of insulin initiation in metformin and sulfonylurea users in primary care practices: the role of kidney function. Journal of diabetes science and technology. 2014;8(5):1023-8.
33. American Diabetes Association. 8. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2018. Diabetes care. 2018;41(Supplement 1):S73 - S85.
34. Viberti G, Kahn SE, Greene DA, Herman WH, Zinman B, Holman RR, et al. A Diabetes Outcome Progression Trial (ADOPT). Diabetes care. 2002;25(10):1737.
35. Best JD, Drury PL, Davis TME, Taskinen M-R, Kesäniemi YA, Scott R, et al. Glycemic control over 5 years in 4,900 people with type 2 diabetes: real-world diabetes therapy in a clinical trial cohort. Diabetes care. 2012;35(5):1165-70.
36. Turner RC, Cull CA, Frighi V, Holman RR. Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus: progressive requirement for multiple therapies (UKPDS 49). UK Prospective Diabetes Study (UKPDS) Group. Jama. 1999;281(21):2005-12.
37. Schrijnders D, Hartog LC, Kleefstra N, Groenier KH, Landman GWD, Bilo HJG. Within-Sulfonylurea-Class Evaluation of Time to Intensification with Insulin (ZODIAC-43). PloS one. 2016;11(6):e0157668-e.
38. Carney GA, Bassett K, Wright JM, Dormuth CR. Is thiazolidinediones use a factor in delaying the need for insulin therapy in type 2 patients with diabetes? A population-based cohort study. BMJ open. 2012;2(6):e001910.
39. Fu AZ, Qiu Y, Davies MJ, Engel SS. Initial sulfonylurea use and subsequent insulin therapy in older subjects with type 2 diabetes mellitus. Diabetes Ther. 2012;3(1):12-.
40. Calvert MJ, McManus RJ, Freemantle N. Management of type 2 diabetes with multiple oral hypoglycaemic agents or insulin in primary care: retrospective cohort study. The British journal of general practice : the journal of the Royal College of General Practitioners. 2007;57(539):455-60.
41. Blumenthal KJ, Larkin ME, Winning G, Nathan DM, Grant RW. Changes in glycemic control from 1996 to 2006 among adults with type 2 diabetes: a longitudinal cohort study. BMC health services research. 2010;10:158-.
42. Royston P. Flexible Parametric Alternatives to the Cox Model, and more. The Stata Journal. 2001;1(1):1-28.
43. Royston P, Lambert P. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. USA: Stata Press; 2011.
44. Koivula, R. W., et al. (2019). "Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium." Diabetologia 62(9): 1601-1615.
45. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes

and impaired glucose regulation for use in a multiethnic UK setting. Diabetic Medicine. 2010;27(8):887-95.
46. Shields BM, McDonald TJ, Ellard S, Campbell MJ, Hyde C, Hattersley AT. The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. Diabetologia. 2012;55(5):1265-72.
47. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal. 2014;35(29):1925-31.

Appendix 1:

R Code for creating diabetes classification model shiny app

```
#####################################################################
## Create classification model shiny app
## Original author - Anita Lynam - g26482@hotmail.co.uk - January 2019
#####################################################################
#Load libraries
library(shiny)
library(shinyjs)
# Define UI for application that returns a probability
ui <- fluidPage(
useShinyjs(),
titlePanel("Type 1/Type 2 diabetes classification model BETA Version"),
h4("This model is designed to differentiate type 1 from type 2 diabetes. If a
diagnosis of monogenic diabetes is being considered please use the ",
tags$a(href="https://www.diabetesgenes.org/mody-probability-calculator/",
target="_blank", "MODY calculator")),
h4("Please enter the age and BMI, other biomarkers are optional"),
div(id="form",
# Sidebar layout with a input and output definitions
sidebarLayout(
sidebarPanel(
#Numeric input control for Age
textInput(inputId = "age", label = "Enter age at diagnosis (yrs) (min 18, max
50)", placeholder = "min is 18, max is 50"
 ),
h5(tags$strong("Enter BMI OR enter height and weight in boxes below, then
press 'Use Height and Weight'")),
# Set BMI using text box
numericInput(inputId = "BMI",value = NULL,label = tags$div(HTML(paste("Enter
BMI (kg/m",tags$sup(2),") (min 17.5, max 70)",sep =""))), min = 17.5, max = 70,
step = 0.5),
#inputs to calculate BMI if required
#height
textInput(inputId ="Height", label ="Height (cm)"),
#weight
textInput(inputId ="Weight", label = "Weight (kg)"),
#allows the user to select BMI input type
actionButton("runBMIInputs", "Use Height and Weight"),
```

```
h5(""),

selectInput("Ethnicity", "Ethnicity: Model is currently only available for white-ethnicities", choices = c("White-European")),

radioButtons("GADA", "Select GADA status:", c("Positive" = "1", "Negative" = "0","Not tested" = ""), selected = ""),

# Selector for choosing IA2

radioButtons("IA2", "Select IA-2 status:", c("Positive" = "1", "Negative" = "0", "Not tested" = ""), selected = ""),

#Numeric input control for T1D GRS

textInput(inputId = "GRS",value = "", label = "Enter T1D GRS centile of type 1 diabetes population*", placeholder = "Enter centile between 0 and 1" ), width = 5),

# Outputs

#the first condition panel is for model 1

mainPanel(

h4(textOutput("txt1")),

h4(textOutput("txt4")),

h4(HTML(paste(textOutput("txt5"),textOutput("txt7")))),

h4(textOutput("txt2")),

h4(textOutput("txt3")),

h4(textOutput("txt6")),

actionButton("resetAll", "Reset all"),

tags$br(),

tabsetPanel(

tabPanel("Model predictions",

conditionalPanel(condition =  "(input.GADA == '') && (input.IA2 == '') && (input.GRS ==')" ,textOutput("mod1Prob"),

tags$head(tags$style("#mod1Prob{color: black; font-size: 20p"

)

)

),

conditionalPanel(

condition =  "(input.GADA != '') && (input.IA2 == '') && (input.GRS ==')" ,

textOutput("mod2Prob"),

tags$head(tags$style("#mod2Prob{color: black; font-size: 20px;}"
```

```
)

)

),

conditionalPanel(condition =  "(input.GADA != '') && (input.IA2 != '') &&
(input.GRS ==')" ,

textOutput("mod3Prob"),

tags$head(tags$style("#mod3Prob{color: black;font-size: 20px;}"

)

)

),

conditionalPanel(condition =  "(input.GADA != '') && (input.IA2 != '') &&
(input.GRS !=')" ,

textOutput("mod4Prob"),

tags$head(tags$style("#mod4Prob{color: black; font-size: 20px;}"

)

)

),

conditionalPanel(condition =  "(input.GADA == '') && (input.IA2 != '') &&
(input.GRS==')",

textOutput("mod5Prob"),

tags$head(tags$style("#mod5Prob{color: black; font-size: 20px;}"

)

)

),

conditionalPanel( condition =  "(input.GADA == '') && (input.IA2 == '') &&
(input.GRS !=')",

textOutput("mod8Prob"),

tags$head(tags$style("#mod8Prob{color: black; font-size: 20px;}"

)

)

),

conditionalPanel(condition =  "(input.GADA == '') && (input.IA2 != '') &&
(input.GRS !=')",

textOutput("mod6Prob"),

tags$head(tags$style("#mod6Prob{color: black;font-size: 20px;}"
```

```
)

)

),

conditionalPanel(condition =  "(input.GADA != '') && (input.IA2 == '') &&
(input.GRS !='')",

textOutput("mod7Prob"),

tags$head(tags$style("#mod7Prob{color: black; font-size: 20px; }"

)

)

)

),

tabPanel("Model information", paste("This model is designed to assist
classification of diabetes in patients diagnosed aged 18 to 50. It was developed
in a white UK population and therefore predictions may not be applicable to
other populations. Type 1/type 2 diabetes is defined using a gold standard
based on measured endogenous insulin secretion (C-peptide) and early insulin
requirement (see diabetes definition tab). The development of the model is
described in XXXX")),

tabPanel("Diabetes definition", h5(" For model development diabetes type was
defined as follows:",tags$br(), "Type 1 - Insulin requirement with 3 years of
diagnosis and severe endogenous insulin deficiency (non-fasting C-peptide
<200pmol/L)", tags$br(), "Type 2 - Absence of insulin requirement within 3
years of diagnosis, or (where early insulin treatment) substantial retained
endogenous insulin secretion after 5 years diabetes duration (non-fasting C-
peptide >600pmol/L)"))

),

width = 7

)

)

),

h5("*Type 1 Diabetes Genetic Risk Score (30 SNP), Oram RA, Patel K, Hill A,
Shields B, McDonald TJ, Jones A, Hattersley AT, Weedon MN: A Type 1
diabetes genetic risk score can aid discrimination between Type 1 and Type 2
diabetes in young adults. Diabetes care 2016;39:337-344")

)

# Define server function

server <- function(input, output, session) {

output$txt1 = renderText({

paste("You have selected: ")
```

```
})
output$txt2 = renderText({
if(input$GADA =="1") {
paste("GADA: Positive")
} else if (input$GADA =="0") {
paste("GADA: Negative")
}
})
output$txt3 = renderText({
if(input$IA2 =="1") {
paste("IA-2: Positive")
} else if (input$IA2 =="0") {
paste("IA-2: Negative")
}
})
output$txt4 = renderText({
if (!(is.null(input$age)||input$age =="")){
paste("Age at diagnosis: ", input$age)}
})
output$txt5 = renderText({
if(!(is.null(input$BMI)||input$BMI ==" ||is.na(input$BMI))){
paste("BMI: ", round(input$BMI,1)) }
})
observeEvent(input$resetAll, {
reset("form")
output$txt7 = renderText({
paste("")
})
})
output$txt6 = renderText({
if(input$GRS !=""){
paste("T1D GRS centile: ", as.numeric(input$GRS))
}
```

```
})

#bmi calculation

observeEvent(input$runBMIInputs,{

updateNumericInput(session, "BMI", value =
paste(round(as.numeric(input$Weight)/((as.numeric(input$Height)/100)^2),1)))

output$txt7 = renderText({

paste("(based on height",input$Height," (cms) and weight ", input$Weight, " (kg)
inputs)")

})

})

observeEvent(input$BMI,if(!(is.null(input$Height)||input$Height
==''||is.na(input$Height)) & !(is.null(input$Weight)||input$Weight
==''||is.na(input$Weight)) & (round(input$BMI)) !=
round(as.numeric(input$Weight)/((as.numeric(input$Height)/100)^2))){
output$txt7 = renderText({

paste("")

})}

)

model1pred = function (){37.9391 + (-5.085444 *log(as.numeric(input$age))) +
(-6.342471 * log(input$BMI))

}

output$mod1Prob = renderText({

if (is.null(input$age)||input$age ==" ) { paste ("Please enter age between 18 and
50 (inclusive)")

}

else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")

}

else if (is.null(input$BMI)||input$BMI ==''||is.na(input$BMI) ) { paste ("Please
enter valid BMI OR use the height and weight inputs")

}

else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model1pred())/(1+exp(model1pred())))*100)>99){

paste("The probability of type 1 diabetes based on your selected inputs is >
99%")
```

```
}
else if(round((exp(model1pred())/(1+exp(model1pred())))*100)<1)
{
paste("The probability of type 1 diabetes based on your selected inputs is <
1%")
}
else {
paste("The probability of type 1 diabetes based on your selected inputs is",
round((exp(model1pred())/(1+exp(model1pred())))*100), "%")
}
}
})
model2pred = function () { -0.9833514 + (0.9433088*model1pred()) + (
3.113623*as.numeric(input$GADA))
}
output$mod2Prob = renderText({if (is.null(input$age)||input$age ==" ) { paste
("Please enter age between 18 and 50 (inclusive)")
}
else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")
}
else if (is.null(input$BMI)||input$BMI ==''||is.na(input$BMI) ) { paste ("Please
enter valid BMI OR use the height and weight inputs")
}
else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")
}
else {if (round((exp(model2pred())/(1+exp(model2pred())))*100)>99)
{paste("The probability of type 1 diabetes based on your selected inputs is >
99%")
} else if(round((exp(model2pred())/(1+exp(model2pred())))*100)<1){paste("The
probability of type 1 diabetes based on your selected inputs is < 1%")
}
else {paste("The probability of type 1 diabetes based on your selected inputs
is",          round((exp(model2pred())/(1+exp(model2pred())))*100), "%")
}
}
```

```
})
AntiStatus1 = function () {
if (input$GADA == 1 && input$IA2 ==0 ){
1
}
else {
0
}
}
AntiStatus2 = function (){
if (input$GADA == 0 && input$IA2 ==1 ){
1
}
else {
0
}
}
AntiStatus3 = function (){
if (input$GADA == 1 && input$IA2 ==1 ){
1
}
else {0
}
}
model3pred = function (){ -1.280086  + (0.9166205*model1pred()) + (3.082366
* AntiStatus1()) + (3.494462* AntiStatus2()) + (4.350717 * AntiStatus3())
}
output$mod3Prob = renderText({ if (is.null(input$age)||input$age ==" )
{paste ("Please enter age between 18 and 50 (inclusive)")
}
else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")
}
```

```
else if (is.null(input$BMI)||input$BMI ==" ||is.na(input$BMI)) { paste ("Please enter valid BMI OR use the height and weight inputs")

}

else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model3pred())/(1+exp(model3pred())))*100)>99){ paste("The probability of type 1 diabetes based on your selected inputs is > 99%")

}

else if(round((exp(model3pred())/(1+exp(model3pred())))*100)<1){paste("The probability of type 1 diabetes based on your selected inputs is < 1%")

}

else {paste("The probability of type 1 diabetes based on your selected inputs is",          round((exp(model3pred())/(1+exp(model3pred())))*100), "%")

}

}

})

model4pred = function (){-7.7859  + (0.8766028*model3pred()) + (30.11052 * (((qnorm(as.numeric(input$GRS) ))*0.025569)+0.278778))

}

output$mod4Prob = renderText({if (is.null(input$age)||input$age ==" ) { paste ("Please enter age between 18 and 50 (inclusive)")

}

else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please enter valid age, minimum age is 18, maximum is 50")

}

else if (is.null(input$BMI)||input$BMI =="||is.na(input$BMI) ) { paste ("Please enter valid BMI OR use the height and weight inputs")

}

else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model4pred())/(1+exp(model4pred())))*100)>99){paste("The probability of type 1 diabetes based on your selected inputs is > 99%")

}

else if(round((exp(model4pred())/(1+exp(model4pred())))*100)<1){ paste("The probability of type 1 diabetes based on your selected inputs is < 1%")
```

```
}
else {paste("The probability of type 1 diabetes based on your selected inputs
is",          round((exp(model4pred())/(1+exp(model4pred())))*100), "%")
}
}
})
model5pred = function (){ -0.3553344 + (3.194096*as.numeric(input$IA2)) +
(0.9916812*model1pred())
}
output$mod5Prob = renderText({if (is.null(input$age)||input$age ==" ) { paste
("Please enter age between 18 and 50 (inclusive)")
}
else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")
}
else if (is.null(input$BMI)||input$BMI ==''||is.na(input$BMI)) { paste ("Please
enter valid BMI OR use the height and weight inputs")
}
else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")
}
else {if (round((exp(model5pred())/(1+exp(model5pred())))*100)>99){paste("The
probability of type 1 diabetes based on your selected inputs is > 99%")
}
else if(round((exp(model5pred())/(1+exp(model51pred())))*100)<1){paste("The
probability of type 1 diabetes based on your selected inputs is < 1%")
}
else {paste("The probability of type 1 diabetes based on your selected inputs
is",          round((exp(model5pred())/(1+exp(model5pred())))*100), "%")
}
}
})
model6pred = function () {-9.9304 + ( 2.953142*as.numeric(input$IA2)) +
(0.8736316*model1pred())+(37.40205*(((qnorm(as.numeric(input$GRS)
))*0.025569)+0.278778))
}
```

```
output$mod6Prob = renderText({if (is.null(input$age)||input$age =='' ) { paste
("Please enter age between 18 and 50 (inclusive)")

}

else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")

}

else if (is.null(input$BMI)||input$BMI =='' ||is.na(input$BMI)) { paste ("Please
enter valid BMI OR use the height and weight inputs")

}

else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model6pred())/(1+exp(model6pred())))*100)>99){paste("The
probability of type 1 diabetes based on your selected inputs is > 99%")

}

else if(round((exp(model6pred())/(1+exp(model6pred())))*100)<1){paste("The
probability of type 1 diabetes based on your selected inputs is < 1%")

}

else {paste("The probability of type 1 diabetes based on your selected inputs
is",

        round((exp(model6pred())/(1+exp(model6pred())))*100), "%")

}

}

})

model7pred = function (){-8.868744 + (2.63093*as.numeric(input$GADA)) +
(0.8454927*model1pred())+(31.22606*(((qnorm(as.numeric(input$GRS)
))*0.025569)+0.278778))

}

output$mod7Prob = renderText({if (is.null(input$age)||input$age =='' ) { paste
("Please enter age between 18 and 50 (inclusive)")

}

else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")

}

else if (is.null(input$BMI)||input$BMI =='||is.na(input$BMI)) { paste ("Please
enter valid BMI OR use the height and weight inputs")

}
```

```
else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model7pred())/(1+exp(model7pred())))*100)>99){
paste("The probability of type 1 diabetes based on your selected inputs is >
99%")

}

else if(round((exp(model7pred())/(1+exp(model7pred())))*100)<1){paste("The
probability of type 1 diabetes based on your selected inputs is < 1%")

}

else {paste("The probability of type 1 diabetes based on your selected inputs
is",

        round((exp(model7pred())/(1+exp(model7pred())))*100), "%")

}

}

})

model8pred = function () { -8.659769 +
(0.8729876*model1pred())+(33.93968*(((qnorm(as.numeric(input$GRS)
))*0.025569)+0.278778))

}

output$mod8Prob = renderText({if (is.null(input$age)||input$age ==" ) { paste
("Please enter age between 18 and 50 (inclusive)")

}

else if ((as.numeric(input$age)<18 || as.numeric(input$age)>50)){ paste("Please
enter valid age, minimum age is 18, maximum is 50")

}

else if (is.null(input$BMI)||input$BMI ==" ||is.na(input$BMI)) { paste ("Please
enter valid BMI OR use the height and weight inputs")

}

else if (input$BMI <17.5 || input$BMI >70) { paste ("Please enter valid BMI
value, minimum BMI is 17.5, maximum is 70")

}

else {if (round((exp(model8pred())/(1+exp(model8pred())))*100)>99){paste("The
probability of type 1 diabetes based on your selected inputs is > 99%")

}

else if(round((exp(model8pred())/(1+exp(model8pred())))*100)<1){paste("The
probability of type 1 diabetes based on your selected inputs is < 1%")

}
```

```
else {paste("The probability of type 1 diabetes based on your selected inputs
is", round((exp(model8pred())/(1+exp(model8pred())))*100), "%")

}

}

})

}

# Create the Shiny app object

shinyApp(ui = ui, server = server)
```

Appendix 2:

R Code for machine learning and logistic regression comparison

```
#################################################################

## Create combined machine learning comparison

## original author - ferratlauric@gmail.com - September 2018

## adapted by Anita Lynam - g26482@hotmail.co.uk - July 2019

#################################################################

# useful web links

# https://www.r-project.org/conferences/useR-2013/Tutorials/kuhn/user_caret_2up.pdf

# https://medium.com/all-things-ai/in-depth-parameter-tuning-for-gradient-boosting-3363992e9bae

# https://stackoverflow.com/questions/15613332/using-caret-package-to-find-optimal-parameters-of-gbm

# https://topepo.github.io/caret/available-models.html

# https://astro.temple.edu/~msobel/courses_files/StochasticBoosting(gradient).pdf

# https://topepo.github.io/caret/random-hyperparameter-search.html

# https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

# https://topepo.github.io/caret/variable-importance.html

# https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

# https://topepo.github.io/caret/variable-importance.html

# https://www.rdocumentation.org/packages/caret/versions/6.0-81/topics/varImp

# https://www.rdocumentation.org/packages/caret/versions/6.0-81/topics/train

#################################################################
####

# install libraries

install.packages("recipes")

install.packages("caret")

install.packages("DMwR")

install.packages("kernlab")

install.packages("randomForest")

install.packages("pROC")

install.packages("gbm")

install.packages("gbm")

install.packages("purrr")

install.packages("PRROC")
```

```
install_github("ddsjoberg/dca")
install.packages("corrplot")
install.packages("plyr")
install.packages("dplyr")
install.packages("Rcpp")
install.packages("rlang")
install.packages("readstata13")

#Load libraries
library(recipes)
library(caret)
library(knitr)
library(kernlab)
library(DMwR)
library(randomForest)
library(pROC)
library(ggplot2)
library(readstata13)
library(rpart)
library(rpart.plot)
library(gbm)
library(gridExtra)
library(plyr)
library(dplyr) # for data manipulation
library(purrr) # for functional programming (map)
library(PRROC) # for Precision-Recall curve calculations
library(dca)
library(corrplot)
library(Rcpp)
library(rlang)
################################
# 1 - Source file
################################
```

```
setwd("Your file path here")
#name the data files
dataFile1 <- "Your stata test data file.dta"
dataFile2 <- "Your stata validation.dta"
#load Stata datasets
dataset_test <- read.dta13(dataFile1)
dataset_val <-  read.dta13(dataFile2)
#create a datset containing a subset of variables to include in the model
myvars <- c("Your outcome variable","Your covariate 1","Your covariate
2","Your covariate 3")
dataset_test <- dataset_test[myvars]
################################
# 2 - Set up the model training
################################
#set seed for reproductivity
seedchoice <- 7
# model formulas
# Add as many covariates as required
formula.model4 <- formula("Your outcome variable ~ Your covariate 1 + Your
covariate 2 + Your covariate3 + ....")
# Data need to be put in a good shape to be used in the caret framework:
# factorise string data and mumeric data which need to be factorised
# no missing data (always possible to impute when it is not the case)
#identify factor variables and view the levels
is.fact2 <- sapply(dataset_test, is.factor)
factors2.df <- dataset_test[, is.fact2]
lapply(factors2.df, levels)
is.fact3 <- sapply(dataset_val, is.factor)
factors3.df <- dataset_val[, is.fact3]
lapply(factors3.df, levels)
#amend class to factor for training and validation datasets
Yourcovariate3 <- myvars[4]
Youroutcomevariable <- myvars[1]
dataset_test[,Yourcovariate3] <- as.factor(dataset_test[,Yourcovariate3])
```

```
dataset_test[,Youroutcomevariable] <-
as.factor(dataset_test[,Youroutcomevariable])

dataset_val[,Yourcovariate3] <- as.factor(dataset_val[,Yourcovariate3])

dataset_val[,Youroutcomevariable] <-
as.factor(dataset_val[,Youroutcomevariable])

#rename the levels of the facor variable. This is required to run the training
models for each of the five imputed datasets

feature6.names <- names(dataset_test)

for (f in feature6.names)    {

  if (class(dataset_test[[f]]) == "factor")    {

    levels6 <- unique(c(dataset_test[[f]]))

    dataset_test[[f]] <- factor(dataset_test[[f]],

                                labels = make.names(levels6))

                                                    }

                                        }

feature7.names <- names(dataset_val)

for (f in feature7.names)    {

  if (class(dataset_val[[f]]) == "factor")     {

    levels7 <- unique(c(dataset_val[[f]]))

    dataset_val[[f]] <- factor(dataset_val[[f]],

                               labels = make.names(levels7))

                                                    }

                                        }

#create standardised variables for the continuous variables

Yourcovariate1 <- myvars[2]

Yourcovariate2 <- myvars[3]

Std_Yourcovariate1 <- paste0("Std_",Yourcovariate1)

Std_Yourcovariate2 <- paste0("Std_",Yourcovariate2)


dataset_test[,Std_Yourcovariate1] <- (dataset_test[,Yourcovariate1] -
mean(dataset_test[,Yourcovariate1]))/sd(dataset_test[,Yourcovariate1])

dataset_test[,Std_Yourcovariate2] <- (dataset_test[,Yourcovariate2] -
mean(dataset_test[,Yourcovariate2]))/sd(dataset_test[,Yourcovariate2])

dataset_val[,Std_Yourcovariate1] <- (dataset_val[,Yourcovariate1] -
mean(dataset_val[,Yourcovariate1]))/sd(dataset_val[,Yourcovariate1])
```

```
dataset_val[,Std_Yourcovariate2] <- (dataset_val[,Yourcovariate2] -
mean(dataset_val[,Yourcovariate2]))/sd(dataset_val[,Yourcovariate2])

# prepare training scheme

#routines, fits each model and calculates a resampling based performance
measure.

# The traincontrol function controls the computational nuances of the train
function

# repeatedcv (repeated cross validation) method is a resampling method
Control that creates multiple versions of the folds and aggregates the results

# number is the k number of folds for the repeatedcv

# repeats is the number of complete sets of folds to compute

# The summmary funcion is a function to compute performance metrics across
resamples.

# twoClassSummary computes sensitivity, specificity and the area under the
ROC curve

# sampling is the type of additional sampling that is conducted after resampling

# (usually to resolve class imbalances).

# SMOTE (Chawla et. al. 2002) is a well-known algorithm to fight unbalanced
classification problem.

# The general idea of this method is to artificially generate new examples of the
minority class using

# the nearest neighbors of these cases. Furthermore, the majority class
examples are also under-sampled,

# leading to a more balanced dataset.

#for use in default and grid search optimised models

control <- trainControl(method = "repeatedcv", number = 10, repeats =
5,classProbs = TRUE,summaryFunction = twoClassSummary, sampling =
"smote", savePredictions = TRUE)

#for use in random search optimised models

control_Rand_Search <- trainControl(method = "repeatedcv", number = 10,
repeats = 5,classProbs = TRUE,summaryFunction = twoClassSummary,
sampling = "smote", savePredictions = TRUE, search = "random")

################################

# 3 - Train the models

################################

# The train function sets up a grid of tuning parameters for a number of
classification and regression

# ROC  used to select the optimal model using the largest value.
```

```
# train the Gradient bootstrap Machine model (Stochastic Gradient Boosting)

# utils::browseVignettes("gbm")

# verbose is an argument of the gmb package, indicating whether or not to print
out progress and

# performance indicators

#build all the Gbm (Stochastic gradient boosting model) models

#tuning parameters: n.trees (number of  iterations), interaction depth
(complexity), shrinkage (learning rate), n.minobsinnode (min number of training
det damples in a node to commence splitting)

#learning rate shrinks the contribution of each tree by learning_rate

getModelInfo()$gbm$parameters

#Shrinkage: the smaller the number, the better the predictive value, the more
trees required, and the more computational cost.

#the smaller the shrinkage, the more trees you should have

# Fetch max Value for interaction.depth

floor(sqrt(NCOL(dataset_test)))

#set up the grid

gbmGrid <-  expand.grid(interaction.depth = c(1, 3, 6, 9, 10),

                        n.trees = c(10, 50,100,150,500),

                        shrinkage = seq(from = 0.01, to = 0.1, by = 0.01),

                        n.minobsinnode = c(5,10,15,20))

#tune the hyper-parameters using  Grid Search

set.seed(seedchoice)

modelGbm_CC_GADA_Lr <- train(formula.model4, data = dataset_test, method
= "gbm", trControl = control, verbose = FALSE,metric =
'ROC',tuneGrid=gbmGrid)

#random search independently draws from a uniform density from the same
configuration space as would be spanned by a regular grid,

#we do not use random hyperparameter search for gbm models as it may be
inefficients

# train the SVM model

# Support Vector Machines with Radial Basis Function Kernel (SVM classifier
using a non-linear kernel)

#RBF is a reasonable first choice, it can handle nonlinear relationships

#C is the penalty parameter of the error term. It controls the trade off between
smooth decision boundary (small c) and classifying the training points correctly.
```

```
#larger values of C focus attention more on (correctly classified) points near the decision boundary (wiggly boundary), while smaller values involve data further away (wider margins).

#sigma the radius/spread/decision boundary of the kernel

#When gamma is low, the 'curve' of the decision boundary is very low and thus the decision region is very broad.

#When gamma is high, the 'curve' of the decision boundary is high, which creates islands of decision-boundaries around data points.

#using training dataset and default parameters

getModelInfo()$svmRadial$parameters

svmGrid <-  expand.grid(sigma = c(0.01, 0.1, 1, 10, 100),

                        C = seq(from = 0.1, to = 1, by = 0.05))

#using training dataset and tune the hyper-parameters using Caret Grid Search

set.seed(seedchoice)

modelSvm_CC_GADA_Lr <- train(formula.model4, data = dataset_test, method = "svmRadial", trControl = control, verbose = TRUE,metric = 'ROC',tuneGrid=svmGrid)

# train the Random forest model

#parameter mtry is the number of variables available for splitting at each tree node

#The default is the square root of the number of predictor variables (rounded down)

#as we are only using three variables we do not optimise the parameters

#For mtry refer to http://code.env.duke.edu/projects/mget/export/HEAD/MGET/Trunk/PythonPackage/dist/TracOnlineDocumentation/Documentation/ArcGISReference/RandomForestModel.FitToArcGISTable.html

#using training dataset and default parameters

set.seed(seedchoice)

modelRf_CC_GADA <- train(formula.model4, data = dataset_test, method = 'rf', trControl = control,metric = 'ROC')

# train a logistic regression model

#using training dataset

#there are no tuning parameters for glm method within caret

set.seed(seedchoice)

modelLG_CC_GADA <- train(formula.model4, data = dataset_test, method = "glm", family = "binomial", trControl = control,metric = 'ROC')

# train neural network
```

```
getModelInfo()$nnet$parameters

#size parameter is the number of units in hidden layer (nnet fit a single hidden layer neural network)

#decay parameter is the regularization parameter to avoid over-fitting

nnetGrid <-  expand.grid(size = seq(from = 1, to = 10, by = 1), decay = c(0.5, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001))

#tune the hyper-parameters using Caret Grid Search

set.seed(seedchoice)

modelnnet_CC_GADA_Lr <- train(formula.model4, data = dataset_test, method = "nnet", trControl = control,metric = 'ROC', tuneGrid = nnetGrid)

# train a k-nearest-neighbours

#based on euclidean distance

getModelInfo()$knn$parameters

#k parameter is the number of neighbours.

knnGrid <-  expand.grid(k = seq(from = 1, to = 100, by = 1))

#tune the hyper-parameters using Caret Grid Search

set.seed(seedchoice)

modelknn_CC_GADA_Lr <- train(formula.model4, data = dataset_test, method = "knn", trControl = control,metric = 'ROC', tuneGrid = knnGrid)

# view the final models

summary(modelknn_CC_GADA_Lr)

summary(modelnnet_CC_GADA_Lr)

summary(modelLG_CC_GADA)

summary(modelRf_CC_GADA)

summary(modelSvm_CC_GADA_Lr)

summary(modelGbm_CC_GADA_Lr)

# collect resamples

#compare the models for the CC existing GADA model (hyperparameter grid search) for comparison

#no grid search for RF or LG models

results_grid_CC_GADA <- resamples(list(LogisticRegression = modelLG_CC_GADA, StochasticGradientBoosting = modelGbm_CC_GADA_Lr, SupportVectorMachine = modelSvm_CC_GADA_Lr, NeuralNetwork = modelnnet_CC_GADA_Lr,RandomForest = modelRf_CC_GADA,              KNearestNeighbours = modelknn_CC_GADA_Lr))

summary(results_grid_CC_GADA)
```

```
#check character string for the performance measure used to sort or computing the between-model correlations
results_grid_CC_GADA$metrics
#visualizing resampling results across models
xyplot(results_grid_CC_GADA, what = "BlandAltman")
# boxplots of results and save as pdf
pdf("Your file name.pdf")
bwplot(results_grid_CC_GADA)
dev.off()
# dot plots of results (includes 95% CI)
# average performance value (with two-sided confidence limits) for each model
pdf("Your file name.pdf")
dotplot(results_grid_CC_GADA)
dev.off()
#trellis scatterplot of results
pdf("Your file name.pdf")
splom(results_grid_CC_GADA)
dev.off()
#test for a difference in the average resampled area under the ROC curve
diffs <- diff(results_grid_CC_GADA, metric = "ROC")
summary(diffs)
#calculate the 95% CI for the resampling ROC AUC
test  <- results_grid_CC_GADA$values

m <- mean(test$`RandomForest~ROC`)
s <- sd(test$`RandomForest~ROC`)
l <- length(test$`RandomForest~ROC`)
m+c(-1.96,1.96)*s/sqrt(length(l))

m <- mean(test$`LogisticRegression~ROC`)
s <- sd(test$`LogisticRegression~ROC`)
l <- length(test$`LogisticRegression~ROC`)
m + (c(-1.96,1.96)*(s/sqrt(length(l))))
```

```
m <- mean(test$`StochasticGradientBoosting~ROC`)
s <- sd(test$`StochasticGradientBoosting~ROC`)
l <- length(test$`StochasticGradientBoosting~ROC`)
m + (c(-1.96,1.96)*(s/sqrt(length(l))))

m <- mean(test$`SupportVectorMachine~ROC`)
s <- sd(test$`SupportVectorMachine~ROC`)
l <- length(test$`SupportVectorMachine~ROC`)
m + (c(-1.96,1.96)*(s/sqrt(length(l))))

m <- mean(test$`NeuralNetwork~ROC`)
s <- sd(test$`NeuralNetwork~ROC`)
l <- length(test$`NeuralNetwork~ROC`)
m + (c(-1.96,1.96)*(s/sqrt(length(l))))

m <- mean(test$`KNearestNeighbours~ROC`)
s <- sd(test$`KNearestNeighbours~ROC`)
l <- length(test$`KNearestNeighbours~ROC`)
m + (c(-1.96,1.96)*(s/sqrt(length(l))))
###################################
# 4 - Perform external validation
###################################
#for the Gbm grid search model (validation)
probsTestGbmGridVal <- predict(modelGbm_CC_GADA_Lr,
newdata=dataset_val, type = "prob")
dataset_val <- data.frame(dataset_val,probsTestGbmGridVal$X2)
predTestGbmGridVal <- log(as.numeric(probsTestGbmGridVal$X2)/(1-
as.numeric(probsTestGbmGridVal$X2)))
#then create a roc object and calculate the ROC on the validation dataset
roc_objTestGbmGridval <- roc(dataset_val[,Youroutcomevariable],
predTestGbmGridVal)
AUC_objTestGbmGridval <- auc(roc_objTestGbmGridval)
AUC_objTestGbmGridval
```

```
ci.auc(roc_objTestGbmGridval)

#for the Svm grid search model (validation)

probsTestSvmGridVal <- predict(modelSvm_CC_GADA_Lr,
newdata=dataset_val, type = "prob")

dataset_val <- data.frame(dataset_val,probsTestSvmGridVal$X2)

predTestSvmGridVal <- log(as.numeric(probsTestSvmGridVal$X2)/(1-
as.numeric(probsTestSvmGridVal$X2)))

#then create a roc object and calculate the ROC on the validation dataset

roc_objTestSvmGridval <- roc(dataset_val[,Youroutcomevariable],
predTestSvmGridVal)

AUC_objTestSvmGridval <- auc(roc_objTestSvmGridval)

AUC_objTestSvmGridval

ci.auc(roc_objTestSvmGridval)

#for the knn grid search model (validation)

probsTestknnGridVal <- predict(modelknn_CC_GADA_Lr,
newdata=dataset_val, type = "prob")

dataset_val <- data.frame(dataset_val,probsTestknnGridVal$X2)

probsTestknnGridVal$X2[probsTestknnGridVal$X2 == 1] <- 0.999999

probsTestknnGridVal$X2[probsTestknnGridVal$X2 == 0] <- 0.000001

predTestknnGridVal <- log(as.numeric(probsTestknnGridVal$X2)/(1-
as.numeric(probsTestknnGridVal$X2)))

#then create a roc object and calculate the ROC on the validation dataset

roc_objTestknnGridval <- roc(dataset_val[,Youroutcomevariable],
predTestknnGridVal)

AUC_objTestknnGridval <- auc(roc_objTestknnGridval)

AUC_objTestknnGridval

ci.auc(roc_objTestknnGridval)

#for the nnet grid search model (validation)

probsTestnnetGridVal <- predict(modelnnet_CC_GADA_Lr,
newdata=dataset_val, type = "prob")

dataset_val <- data.frame(dataset_val,probsTestnnetGridVal$X2)

predTestnnetGridVal <- log(as.numeric(probsTestnnetGridVal$X2)/(1-
as.numeric(probsTestnnetGridVal$X2)))

#then create a roc object and calculate the ROC on the validation dataset

roc_objTestnnetGridval <- roc(dataset_val[,Youroutcomevariable],
predTestnnetGridVal)
```

```
AUC_objTestnnetGridval <- auc(roc_objTestnnetGridval)

AUC_objTestnnetGridval

ci.auc(roc_objTestnnetGridval)

#for the RF model (validation)

probsTestRfVal <- predict(modelRf_CC_GADA, newdata=dataset_val, type =
"prob")

dataset_val <- data.frame(dataset_val,probsTestRfVal$X2)

probsTestRfVal$X2[probsTestRfVal$X2 == 1] <- 0.999999

probsTestRfVal$X2[probsTestRfVal$X2 == 0] <- 0.000001

predTestRfVal <- log(as.numeric(probsTestRfVal$X2)/(1-
as.numeric(probsTestRfVal$X2)))

#then create a roc object and calculate the ROC on the validation dataset

roc_objTestRfval <- roc(dataset_val[,Youroutcomevariable], predTestRfVal)

AUC_objTestRfval <- auc(roc_objTestRfval)

AUC_objTestRfval

ci.auc(roc_objTestRfval)

#for the logistic regression model (validation)

probsTestlgVal <- predict(modelLG_CC_GADA, newdata=dataset_val, type =
"prob")

dataset_val <- data.frame(dataset_val,probsTestlgVal$X2)

predTestLGVal <- log(as.numeric(probsTestlgVal$X2)/(1-
as.numeric(probsTestlgVal$X2)))

#then create a roc object and calculate the ROC on the validation dataset

roc_objTestLGval <- roc(dataset_val[,Youroutcomevariable], predTestLGVal)

AUC_objTestLGval <- auc(roc_objTestLGval)

AUC_objTestLGval

ci.auc(roc_objTestLGval)

#plot the roc curves

plot(roc_objTestRfval, col = "gray85",main = "",add=FALSE)

plot(roc_objTestLGval, col = "gray45", add = TRUE)

plot(roc_objTestSvmGridval, co = "black", add = TRUE)

plot(roc_objTestGbmGridval, col = "gray85",  lty = 3, add = TRUE)

plot(roc_objTestnnetGridval, col = "black",lty = 3, add = TRUE)

plot(roc_objTestknnGridval, col = "gray45", lty = 3, add = TRUE)
```

```
model=c('LG','GBM','SVM','RF','Nnet','Knn')

AUC <- c(AUC_objTestLGval, AUC_objTestGbmGridval,
AUC_objTestSvmGridval, AUC_objTestRfval, AUC_objTestnnetGridval,
AUC_objTestknnGridval)

ValResults <- data.frame(model, AUC)

#use prediction-recall curve to validate the models

#calculate the AUPRC for the validation dataset

prRFVal <- pr.curve(1-
dataset_val$probsTestRfVal.X2,dataset_val$probsTestRfVal.X2, curve =
TRUE)

prLGVal <-  pr.curve(1-
dataset_val$probsTestlgVal.X2,dataset_val$probsTestlgVal.X2, curve = TRUE)

prSVMVal <- pr.curve(1-
dataset_val$probsTestSvmGridVal.X2,dataset_val$probsTestSvmGridVal.X2,
curve = TRUE)

prGBMVal <- pr.curve(1-
dataset_val$probsTestGbmGridVal.X2,dataset_val$probsTestGbmGridVal.X2,
curve = TRUE)

prNNVal <- pr.curve(1-
dataset_val$probsTestnnetGridVal.X2,dataset_val$probsTestnnetGridVal.X2,
curve = TRUE)

prKNNVal <- pr.curve(1-
dataset_val$probsTestknnGridVal.X2,dataset_val$probsTestknnGridVal.X2,
curve = TRUE)

#return the AUPRC

prRFVal

prLGVal

prSVMVal

prGBMVal

prNNVal

prKNNVal

# plot PR curve for the test curve in red, without legend

plot(prRFVal, color = "gray85",auc.main=FALSE, main = "")

plot( prLGVal, color = "gray45", add = TRUE)

plot( prSVMVal, color = "black", add = TRUE)

plot( prGBMVal, color = "gray85", lty = 3, add = TRUE)

plot( prNNVal, color = "black", lty = 3, add = TRUE)

plot( prKNNVal, color = "gray45", lty = 3,add = TRUE)
```

```
#plot the calibration plots with loess smoother

#for logistic regression

#create 10 risk groups

dataset_val %>% mutate(quintile=ntile(dataset_val$probsTestlgVal.X2,10)) ->
dataset_val_10

Youroutcomevariable_num <- paste(Youroutcomevariable,"num")

dataset_val_10[as.numeric(dataset_val_10[,Youroutcomevariable])==
1,Youroutcomevariable_num] <- 0

dataset_val_10[as.numeric(dataset_val_10[,Youroutcomevariable])==
2,Youroutcomevariable_num] <- 1

#average the observed and expected probabilities of patients in each risk group

obs <- aggregate(as.numeric(dataset_val_10[,Youroutcomevariable_num]),
list(dataset_val_10$quintile),mean)

exptd <- aggregate(dataset_val_10$probsTestlgVal.X2,
list(dataset_val_10$quintile),mean)

obsn <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10, length)

#CIs for scatter points

lci <- obs - (1.96*(((obs*(1-obs))/obsn[,Youroutcomevariable])^.5))

lci[lci<0]<- 0

uci <- obs + (1.96*(((obs*(1-obs))/obsn[,Youroutcomevariable])^.5))

uci[uci>1]<- 1

LR_Cali_Plot = data.frame(exptd$x,obs$x, uci$x, lci$x)

ggplot(LR_Cali_Plot, aes(x= exptd$x, y=obs$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

         scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci$x, ymax=uci$x), width=0.02) +

        theme_bw()

#for SVM

#create 10 risk groups
```

```
dataset_val %>%
mutate(quintile=ntile(dataset_val$probsTestSvmGridVal.X2,10)) ->
dataset_val_10_SVM

dataset_val_10_SVM[as.numeric(dataset_val_10_SVM[,Youroutcomevariable])
== 1,Youroutcomevariable_num] <- 0

dataset_val_10_SVM[as.numeric(dataset_val_10_SVM[,Youroutcomevariable])
== 2,Youroutcomevariable_num] <- 1

#average the observed and expected probabilities of patients in each risk group

obs_SVM <-
aggregate(as.numeric(dataset_val_10_SVM[,Youroutcomevariable_num]),
list(dataset_val_10_SVM$quintile),mean)

exptd_SVM <- aggregate(dataset_val_10_SVM$probsTestSvmGridVal.X2,
list(dataset_val_10_SVM$quintile),mean)

obsn_SVM <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10_SVM, length)

#CIs for scatter points

lci_SVM <- obs_SVM- (1.96*(((obs_SVM*(1-
obs_SVM))/obsn_SVM[,Youroutcomevariable])^.5))

lci_SVM[lci_SVM<0]<- 0

uci_SVM <- obs_SVM + (1.96*(((obs_SVM*(1-
obs_SVM))/obsn_SVM[,Youroutcomevariable])^.5))

uci_SVM[uci_SVM>1]<- 1

SVM_Cali_Plot <- data.frame(exptd_SVM$x,obs_SVM$x, uci_SVM$x,
lci_SVM$x)

ggplot(SVM_Cali_Plot, aes(x= exptd_SVM$x, y=obs_SVM$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

        scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci_SVM$x, ymax=uci_SVM$x), width=0.02) +

         theme_bw()

#for Random Forest

#create 10 risk groups

dataset_val %>% mutate(quintile=ntile(dataset_val$probsTestRfVal.X2,10)) ->
dataset_val_10_RF
```

```
dataset_val_10_RF[as.numeric(dataset_val_10_RF[,Youroutcomevariable])==
1,Youroutcomevariable_num] <- 0

dataset_val_10_RF[as.numeric(dataset_val_10_RF[,Youroutcomevariable])==
2,Youroutcomevariable_num] <- 1

#average the observed and expected probabilities of patients in each risk group

obs_RF <-
aggregate(as.numeric(dataset_val_10_RF[,Youroutcomevariable_num]),
list(dataset_val_10_RF$quintile),mean)

exptd_RF <- aggregate(dataset_val_10_RF$probsTestRfVal.X2,
list(dataset_val_10_RF$quintile),mean)

obsn_RF <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10_RF, length)

#CIs for scatter points

lci_RF <- obs_RF- (1.96*(((obs_RF*(1-
obs_RF))/obsn_RF[,Youroutcomevariable])^.5))

lci_RF[lci_RF<0]<- 0

uci_RF = obs_RF + (1.96*(((obs_RF*(1-
obs_RF))/obsn_RF[,Youroutcomevariable])^.5))

uci_RF[uci_RF>1]<- 1

RF_Cali_Plot <- data.frame(exptd_RF$x,obs_RF$x, uci_RF$x, lci_RF$x)

ggplot(RF_Cali_Plot, aes(x= exptd_RF$x, y=obs_RF$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

        scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci_RF$x, ymax=uci_RF$x), width=0.02) +

        theme_bw()

#for GBM

#create 10 risk groups

dataset_val %>%
mutate(quintile=ntile(dataset_val$probsTestGbmGridVal.X2,10)) ->
dataset_val_10_GBM

dataset_val_10_GBM[as.numeric(dataset_val_10_GBM[,Youroutcomevariable])
== 1,Youroutcomevariable_num] <- 0
```

```
dataset_val_10_GBM[as.numeric(dataset_val_10_GBM[,Youroutcomevariable])
== 2,Youroutcomevariable_num] <- 1

#average the observed and expected probabilities of patients in each risk group

obs_GBM <-
aggregate(as.numeric(dataset_val_10_GBM[,Youroutcomevariable_num]),
list(dataset_val_10_GBM$quintile),mean)

exptd_GBM <- aggregate(dataset_val_10_GBM$probsTestGbmGridVal.X2,
list(dataset_val_10_GBM$quintile),mean)

obsn_GBM <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10_GBM, length)

#CIs for scatter points

lci_GBM <- obs_GBM- (1.96*(((obs_GBM*(1-
obs_GBM))/obsn_GBM[,Youroutcomevariable])^.5))

lci_GBM[lci_GBM<0]<- 0

uci_GBM <- obs_GBM + (1.96*(((obs_GBM*(1-
obs_GBM))/obsn_GBM[,Youroutcomevariable])^.5))

uci_GBM[uci_GBM>1]<-1

GBM_Cali_Plot = data.frame(exptd_GBM$x,obs_GBM$x, uci_GBM$x,
lci_GBM$x)

ggplot(GBM_Cali_Plot, aes(x= exptd_GBM$x, y=obs_GBM$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

        scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci_GBM$x, ymax=uci_GBM$x), width=0.02) +

        theme_bw()

#for KNN

#create 10 risk groups

dataset_val %>%
mutate(quintile=ntile(dataset_val$probsTestknnGridVal.X2,10)) ->
dataset_val_10_KNN

dataset_val_10_KNN[as.numeric(dataset_val_10_KNN[,Youroutcomevariable])
== 1,Youroutcomevariable_num] <- 0

dataset_val_10_KNN[as.numeric(dataset_val_10_KNN[,Youroutcomevariable])
== 2,Youroutcomevariable_num] <- 1
```

```
#average the observed and expected probabilities of patients in each risk group

obs_KNN <-
aggregate(as.numeric(dataset_val_10_KNN[,Youroutcomevariable_num]),
list(dataset_val_10_KNN$quintile),mean)

exptd_KNN <- aggregate(dataset_val_10_KNN$probsTestknnGridVal.X2,
list(dataset_val_10_KNN$quintile),mean)

obsn_KNN <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10_KNN, length)

#CIs for scatter points

lci_KNN <- obs_KNN- (1.96*(((obs_KNN*(1-
obs_KNN))/obsn_KNN[,Youroutcomevariable])^.5))

lci_KNN[lci_KNN<0]<- 0

uci_KNN = obs_KNN + (1.96*(((obs_KNN*(1-
obs_KNN))/obsn_KNN[,Youroutcomevariable])^.5))

uci_KNN[uci_KNN>1]<- 1

KNN_Cali_Plot <- data.frame(exptd_KNN$x,obs_KNN$x, uci_KNN$x,
lci_KNN$x)

ggplot(KNN_Cali_Plot, aes(x= exptd_KNN$x, y=obs_KNN$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

        scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci_KNN$x, ymax=uci_KNN$x), width=0.02) +

        theme_bw()

#for NN

#create 10 risk groups

dataset_val %>%
mutate(quintile=ntile(dataset_val$probsTestnnetGridVal.X2,10)) ->
dataset_val_10_NN

dataset_val_10_NN[as.numeric(dataset_val_10_NN[,Youroutcomevariable])==
1,Youroutcomevariable_num] <- 0

dataset_val_10_NN[as.numeric(dataset_val_10_NN[,Youroutcomevariable])==
2,Youroutcomevariable_num] <- 1

#average the observed and expected probabilities of patients in each risk group
```

```
obs_NN <-
aggregate(as.numeric(dataset_val_10_NN[,Youroutcomevariable_num]),
list(dataset_val_10_NN$quintile),mean)

exptd_NN <- aggregate(dataset_val_10_NN$probsTestnnetGridVal.X2,
list(dataset_val_10_NN$quintile),mean)

obsn_NN <- aggregate(as.formula(paste0(Youroutcomevariable ,"~ quintile")),
dataset_val_10_NN, length)

#CIs for scatter points

lci_NN <- obs_NN- (1.96*(((obs_NN*(1-
obs_NN))/obsn_NN[,Youroutcomevariable])^.5))

lci_NN[lci_NN<0]<- 0

uci_NN <- obs_NN + (1.96*(((obs_NN*(1-
obs_NN))/obsn_NN[,Youroutcomevariable])^.5))

uci_NN[uci_NN>1]<- 1

NN_Cali_Plot <- data.frame(exptd_NN$x,obs_NN$x, uci_NN$x, lci_NN$x)

ggplot(NN_Cali_Plot, aes(x= exptd_NN$x, y=obs_NN$x)) +

        geom_point(size = 2) +

        geom_smooth(method=loess, se=FALSE, col = "black", lwd = 1) +

        geom_abline(slope=1, intercept=0, lty=2 ) +

        scale_x_continuous(name = "Expected", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        scale_y_continuous(name = "Observed", breaks = c(0.0,
        0.2,0.4,0.6,0.8,1.0),limits = c(0,1)) +

        geom_errorbar(aes(ymin=lci_NN$x, ymax=uci_NN$x), width=0.02) +

        theme_bw()

#calcluate Calibration slope for each model

glm(formula(paste0(Youroutcomevariable," ~ predTestGbmGridVal")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~ predTestSvmGridVal")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~ predTestknnGridVal")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~ predTestnnetGridVal")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~ predTestRfVal")), family=binomial,
data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~ predTestLGVal")),
family=binomial, data=dataset_val)
```

```
#calcluate Calibration in the large for each model

#predicted risks are understated if _b[_cons] > 0 or overstated if _b[_cons] < 0

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestGbmGridVal)")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestSvmGridVal)")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestknnGridVal)")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestnnetGridVal)")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestRfVal)")),
family=binomial, data=dataset_val)

glm(formula(paste0(Youroutcomevariable," ~  offset(predTestLGVal)")),
family=binomial, data=dataset_val)

#calcluate overall misCalibration for each model

#the slope coefficient beta of the linear predictors reflects the deviations from
the ideal slope of 1.

#If p is significant then there is deviation from zero

mc1 <- glm(formula(paste0(Youroutcomevariable," ~  predTestGbmGridVal+
offset(predTestGbmGridVal)")), family=binomial, data=dataset_val)

mc2 <- glm(formula(paste0(Youroutcomevariable," ~  predTestSvmGridVal +
offset(predTestSvmGridVal)")), family=binomial, data=dataset_val)

mc3 <- glm(formula(paste0(Youroutcomevariable," ~  predTestknnGridVal +
offset(predTestknnGridVal)")), family=binomial, data=dataset_val)

mc4 <- glm(formula(paste0(Youroutcomevariable," ~  predTestnnetGridVal +
offset(predTestnnetGridVal)")), family=binomial, data=dataset_val)

mc5 <- glm(formula(paste0(Youroutcomevariable," ~  predTestRfVal +
offset(predTestRfVal)")), family=binomial, data=dataset_val)

mc6 <- glm(formula(paste0(Youroutcomevariable," ~  predTestLGVal +
offset(predTestLGVal)")), family=binomial, data=dataset_val)

summary(mc1)

summary(mc2)

summary(mc3)

summary(mc4)

summary(mc5)

summary(mc6)

#correlation matrix of predictions - validation dataset
```

```
predMatrixVal <- data.frame(dataset_val$probsTestGbmGridVal.X2
,dataset_val$probsTestSvmGridVal.X2, dataset_val$probsTestknnGridVal.X2 ,
dataset_val$probsTestnnetGridVal.X2 ,dataset_val$probsTestRfVal.X2,
dataset_val$probsTestlgVal.X2)

names(predMatrixVal)[1] <-"GBM"

names(predMatrixVal)[2] <-"SVM"

names(predMatrixVal)[3] <-"KNN"

names(predMatrixVal)[4] <-"NN"

names(predMatrixVal)[5] <-"RF"

names(predMatrixVal)[6] <-"LR"

MVal <- cor(predMatrixVal)

corrplot(MVal, method="number",tl.cex = 1)

#create a variable importance dataframe

# Svm and KNN do not have built-in variable importance score

Model <- c('Logistic Regression','Stochastic Gradient Boosting', 'Neural
Network', 'Random Forest')

# calculate the variable importance scores

# varImp function provides the variable importance

LGImp <- varImp(modelLG_CC_GADA, scale = FALSE)

LGImp

gmbImp <- varImp(modelGbm_CC_GADA_Lr, scale = FALSE)

gmbImp

nnetImp <- varImp(modelnnet_CC_GADA_Lr, scale = FALSE)

nnetImp

rfImp <- varImp(modelRf_CC_GADA, scale = FALSE)

rfImp

#manually divide each variable importance scores by max to scale

Yourcovariate1 <- c(insert your variance importance scores here)

Yourcovariate2 <- c(insert your variance importance scores here)

Yourcovariate3 <- c(insert your variance importance scores here)

#build the DF with the scaled variable importance scores

varImpDF = data.frame(Model,Yourcovariate1,Yourcovariate2,Yourcovariate3)

#build the plots of the variable importance ranks

plotVarImp1 <- ggplot(data = varImpDF, aes(x = Model, y =
varImpDF$Yourcovariate1))+geom_bar(stat="identity",width=0.06)+
```

```
coord_flip()+ ylab("Scaled variable importance score")+  xlab("") + ggtitle("Your covariate 1") + scale_y_continuous(expand = c(0, 0)) +theme(axis.text.y = element_blank(),axis.ticks.y = element_blank() )

plotVarImp2 <- ggplot(data = varImpDF, aes(x = Model, y = varImpDF$Yourcovariate2))+geom_bar(stat="identity",width=0.06)+ coord_flip()+ ylab("Scaled variable importance score")+ xlab("") + ggtitle("Your covariate 2") + scale_y_continuous(expand = c(0, 0)) +theme(axis.text.y = element_blank(),axis.ticks.y = element_blank() )

plotVarImp3 <- ggplot(data = varImpDF, aes(x = Model, y = varImpDF$Yourcovariate3))+geom_bar(stat="identity",width=0.06)+ coord_flip()+ ylab("Scaled variable importance score")+  xlab("") + ggtitle("Your covariate 3") + scale_y_continuous(expand = c(0, 0)) +theme(axis.text.y = element_blank(),axis.ticks.y = element_blank() )

#plot the charts on one row

par(mfrow<-c(1,3))

plot(plotVarImp1)

plot(plotVarImp2)

plot(plotVarImp3)

grid.arrange(plotVarImp1, plotVarImp2, plotVarImp3,ncol = 3)

###################################

# 5 - save the objects for future use

###################################

save(dataset_test,dataset_val,ValResults,control,modelLG_CC_GADA,modelRf_CC_GADA, modelSvm_CC_GADA_Lr, modelGbm_CC_GADA_Lr, modelnnet_CC_GADA_Lr,modelknn_CC_GADA_Lr, varImpDF, results_grid_CC_GADA,gmbImp, rfImp, nnetImp,dataset_val_10, file = "Your Machine Learning Objects.RData")
```