Identifying type 1 and 2 diabetes in research datasets where classification biomarkers are unavailable: assessing the accuracy of published approaches

Nicholas J Thomas 1,2, Andrew McGovern 1,2, Katherine G Young 1, Seth A Sharp 1, Michael N Weedon 1, Andrew T Hattersley 1,2, John Dennis 1, Angus G Jones 1,2

Author details:
1. Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK
2. Department of Diabetes and Endocrinology, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK, Exeter, UK

Corresponding author

Angus Jones
Institute of Biomedical and Clinical Science, University of Exeter, College of Medicine and Health, Exeter, UK
Email: angus.jones@exeter.ac.uk
Tel 01392408538

Abstract: 199
Manuscript: 4043
Tables: 4
Abstract:

Objective

We aimed to compare the performance of approaches for classifying insulin treated diabetes within research datasets without measured classification biomarkers, evaluated against two independent biological definitions of diabetes type.

Study design and Setting

We compared accuracy of ten reported approaches for classifying insulin treated diabetes into type 1 (T1D) and type 2 (T2D) diabetes in two cohorts: UK Biobank (UKBB) n=26,399 and DARE n=1,296. Overall performance for classifying T1D and T2D were assessed using: a T1D genetic risk score and genetic stratification method (UKBB); C-peptide measured at >3 years diabetes duration (DARE).

Results

Approaches accuracy ranged from 71%-88% (UKBB) and 68%-88% (DARE). When classifying all participants, combining early insulin requirement with a T1D probability model (incorporating diagnosis age and BMI), and interview reported diabetes type (UKBB available in only 15%) consistently achieved high accuracy, (UKBB 87% and 87%, DARE 85% and 88% respectively). For identifying T1D with minimal misclassification, models with high thresholds or young diagnosis age (<20 years) had highest performance. Findings were incorporated into an online tool identifying optimum approaches based on variable availability.

Conclusion

Models combining continuous features with early insulin requirement are the most accurate methods for classifying insulin treated diabetes in research datasets without measured classification biomarkers.
Keywords

Diabetes Classification
Population Studies
Cohort stratification
Type 1 Diabetes
Type 2 Diabetes
Diabetes Epidemiology

Running Title: Identifying type 1 and 2 diabetes in research datasets
What is new

Key findings:

Across two different datasets classification models incorporating continuous clinical features combined with early insulin requirement, or (where available) interview reported diabetes type, consistently achieved high accuracy (≥85%).

When identifying a type 1 diabetes cohort with minimal misclassification, young age at diagnosis (<20 years) or models with high thresholds had very high predictive value, but modest sensitivity.

What this adds to what is known:

The best approaches for classifying diabetes type in research datasets without measured classification biomarkers was previously unclear. This work allows researchers to identify the optimum classification approach for their dataset and research question.

Implications

The optimal method for identifying diabetes subtypes in observational data will depend on available data and research question. Researchers can select the optimum approach using an online tool devised using the study findings (Classifying Diabetes for Research: Method Selector (newcastle-rse.github.io)).
Background

Robustly classifying diabetes type in research datasets without measured classification biomarkers is challenging

Large population level research datasets are widely used for clinical studies of people with diabetes, however for results to be robust, accurate diabetes classification is fundamental. Together type 1 diabetes (T1D) and type 2 diabetes (T2D) account for ≥98% of all diabetes cases, (1) but these two subtypes have marked differences in aetiology, pathophysiology and management (2). While absence of insulin treatment in longstanding diabetes is highly specific for T2D (2, 3), classifying currently insulin treated diabetes cases is challenging (3-6). Clinical diagnosis is frequently unavailable in research datasets, and if available will include substantial misclassification and miscoding (≈15%)(7-12). In research datasets, biomarkers which can help improve classification, such as C-peptide or islet autoantibodies (4, 13), are rarely available. The rarity of T2D in children makes young age of diabetes onset specific for T1D, but the over half of T1D cases occurring in adults will be missed (3-5, 14).

The comparative performance of approaches to classify insulin treated diabetes in epidemiological studies is unknown.

The optimum approach for classifying T1D and T2D in research datasets remains unclear. Previously published approaches vary and include: clinician or interview reported diabetes type, diabetes treatment, billing codes or using specific cut offs of diabetes related features for example body mass index (BMI) or age at diabetes diagnosis (15-25). Where the performance of these approaches has been assessed, it is normally against a clinical based assessment of T1D or T2D diagnosis (10, 15-18, 22-25). These assessments will not only suffer from the inaccuracies of clinical diagnosis and coding, but also a circularity bias that features favored by clinicians for determining diabetes type will appear to be most discriminatory. While prediction models for classification have been developed and tested against C-peptide and histology defined diabetes type these have not been compared to other approaches (19, 20, 26). To date, there has not been an evaluation of the comparative performance of existing classification approaches against a robust independent biomarker.

Aim

To help researchers to choose the optimum diabetes classification approach for research datasets without measured classification biomarkers, we aimed to compare the performance of a number of published approaches for classifying insulin treated diabetes in two population level research datasets. Classification approaches were evaluated against two independent biological definitions of diabetes type based on type 1 diabetes genetic risk scores (T1GRS), and measured C-peptide.
Method

Within two population research datasets we assessed the performance of different published approaches for classifying insulin treated diabetes into T1D and T2D against biomarker defined diabetes subtypes. In UK Biobank we used a type 1 diabetes genetic risk score (T1DGRS) within a previously published genetic stratification method (3, 27) to compare the proportion of T1D and T2D cases correctly and incorrectly classified by each approach. We also assessed the performance of these approaches in a large unselected research dataset with diabetes (the DARE cohort) against diabetes type defined by C-peptide level (a measure of endogenous insulin secretion) measured after a median 14 years duration (13).

Study design and participants

UK Biobank

We evaluated a subset of 26,399 unrelated individuals self-reporting diabetes from the UK Biobank (28). To allow direct comparison of classification approaches in the same cohort, individuals were excluded where missing BMI measurement (n=237) or age at diabetes diagnosis (n=1,675). A further 1,389 participants were excluded where it was not possible to generate a T1DGRS. Overall 23,098 participants met study eligibility criteria, a study flowchart is shown in electronic supplementary materials (ESM) Figure 1a. A subset of 45% (10,491/23,098) of participants had linkage to their primary care record.

The main analysis was restricted to the 72% (16,619/23,098) of participants of white European descent, as the T1DGRS used to define diabetes type has not been validated in non-white ethnicities (29, 30). People of white European descent were those who self-identified as white European and were confirmed as ancestrally white by use of principal components analyses of genome-wide genetic information (31). A secondary exploratory analysis was undertaken including all 23,098 participants of all ethnicities. Clinical history was interview reported diabetes type via an interactive questionnaire and nurse led interview, further details of clinical features and lipid assessment are given in ESM.

DARE cohort

The DARE study recruited, predominantly though primary care in the South West of England, an unselected population of adults with diabetes (regardless of age of onset or diabetes type, gestational diabetes excluded) (8). We evaluated 1,296 participants (22% (1296/5991) of the DARE cohort) receiving insulin treatment. C-peptide was measured on stored non fasting EDTA at DARE recruitment after January 2010 as previously described (see ESM) (8). Participants were excluded where BMI measurement was missing (n=6) or if diabetes duration at recruitment was ≤3 years (n=49) due to the limitations of C-peptide assessment in short duration diabetes (13). A study flow chart is shown in ESM Figure 1b. Although all ethnicities were recruited to DARE, 99% were white (1224/1241). In DARE all clinical history was self-reported by participants in an interview with a research nurse as reported previously (8).

Assessment of population level approaches for classifying diabetes type in insulin treated
Overall we compared ten different approaches for the classification of insulin treated diabetes selected based on those commonly used in the literature (10, 15-20, 22, 32). The variables required for each approach are listed in Table 1. For all approaches using continuous variables, cut offs to classify either T1D or T2D were selected based on previously proposed values where available (10, 15, 16, 19, 20). Different cut offs were used where the aim was to classify all insulin treated participants or select a T1D or T2D cohort with minimal misclassification (Table 1). For identifying ‘pure’ type 1 and 2 diabetes using prediction models, no previous cut off has been recommended, therefore cut offs were chosen prior to analysis based on probability thresholds that gave high positive predictive value (PPV) for type 1 or 2 diabetes in previous literature (19): T1D ≥80% probability and T2D <5% probability, for defining T1D a further cut-off of 20% probability was evaluated to give a high PPV whilst aiming to capture a high percentage of all T1D cases. Insulin within a year of diagnosis and OHA treatment are well reported to associate with T1D and T2D respectively (10). Therefore as an additional analysis performance of approaches was further evaluated with the addition of knowledge of insulin within a year of diagnosis defined as insulin treatment within a year of diagnosis, or also by current treatment with any OHA. Full details for each approach given in ESM methods.

Biological definitions of diabetes type approaches evaluated against:

UK Biobank:

We have recently shown that measuring the average polygenic susceptibility to T1D (captured by a T1D genetic risk score (T1DGRS)) of a cohort with diabetes can allow the proportion of T1D in that cohort to be estimated, based on enrichment for genetic susceptibility to type 1 diabetes over and above population susceptibility, as described in statistical analysis below (3, 27). Important at an individual level a high genetic susceptibility for T1D does not prevent a person having T2D and those developing T1D can do so without T1D genetic risk (33). Therefore this analysis is evaluated within a cohort as on average those with type 1 diabetes will have significantly higher genetic predisposition to type 1 diabetes than those without (30, 34). Calculations of proportions with and without type 1 diabetes using this method are estimates but have been previously shown to be robust with the accuracy and precision of these estimates discussed in detail elsewhere (27). Full details of T1DGRS generation used are given in ESM methods.

DARE:

T1D was defined as severe insulin deficiency: measured non-fasting C-peptide <200pmol/L. T2D was defined as participants currently insulin treated with a C-peptide ≥200pmol/L. All analysed participants had a duration of diabetes at C-peptide measurement of over three years (13)

Statistical analysis

When classifying all insulin treated cases, approaches were ranked by the overall accuracy of each definition, defined as the proportion of all T1D and T2D cases correctly classified relative to the total
number of all cases classified. For each approach the PPV of cases called T1D and T2D (percent of those identified who have the condition as defined by the biological standard) and sensitivity for detecting T1D and T2D (percentage of cases with the condition identified) were also calculated. Where aiming to classify just a T1D or T2D cohort, approaches were ranked firstly based on PPV and then secondly by sensitivity.

**UK Biobank**

For each classification approach the mean T1DGRS for cases classified as T1D \((Approach_{Called} \ T1D)\) and T2D \((Approach_{Called} \ T2D)\) were separately evaluated against mean T1DGRS for reference T1D cases \((Reference_{T1D})\) \(n=6483\) mean T1DGRS = 14.50 and reference Type 2 diabetes equivalent cohort \((Reference_{T2D})\) \(n= 9246\) mean T1DGRS = 10.37 both taken from the Type 1 Diabetes Genetics Consortium (35). Reference T1D cases were white European, clinically diagnosed and aged <17 years at diagnosis. The higher the proportion of diabetes cases correctly defined by a classification approach the more the T1DGRS of the groups classified as T1D or T2D will respectively genetically resemble true T1D and T2D reference populations (method shown in ESM figure 2). The proportion of T1D within groups, defined by each classification approach, is then estimated according to the normalised difference of each clinical definitions mean T1DGRS \((Approach_{Called} \ (T1D/T2D))\) and the mean T1DGRS of the two reference populations \((Reference_{T1D} \text{ and } Reference_{T2D})\) in the equations below and as described previously (27, 36). For cases defined as having T1D by each classification approach, PPV for T1D is equivalent to \(Proportion_{T1D}\). For cases defined as having T2D by each classification approach, positive predictive value (PPV) for T2D is calculated as \(1 - Proportion_{T1D}\).

\[
Proportion (PPV)_{T1D} = \frac{Approach \ T1DGRS_{Called \ T1D} - Reference \ T1DGRS_{T2D}}{Reference \ T1DGRS_{T1D} - Reference \ T1DGRS_{T2D}}
\]

\[
Proportion (PPV)_{T2D} = 1 - \frac{Approach \ T1DGRS_{Called \ T2D} - Reference \ T1DGRS_{T2D}}{Reference \ T1DGRS_{T1D} - Reference \ T1DGRS_{T2D}}
\]

Sensitivity was estimated as:

\[
Sensitivity \ T1D = \left| \frac{(PPV_{T1D} \times n_{T1D})}{(PPV_{T1D} \times n_{T1D}) + ((1 - PPV_{T2D}) \times n_{T2D})} \right|
\]

Where \(n_{T1D}\) is the number of cases called as having T1D and \(n_{T2D}\) is the number of cases called as having T2D by each approach.

**Determining accuracy in UK Biobank and DARE**

Where all insulin treated participants were classified as having either T1D or T2D, accuracy was calculated as:

\[
Accuracy = \left| \frac{(PPV_{T1D} \times n_{T1D}) + (PPV_{T2D} \times n_{T2D})}{(n_{T1D} + n_{T2D})} \right|
\]
All analyses were performed using Stata 16 (StataCorp LP, College Station, TX).
Results:

Performance of approaches to classifying all insulin treated white European participants with diabetes in UK Biobank

Within the UK Biobank, of the white European participants meeting eligibility criteria, 21% (3534/16,619) were insulin treated. The clinical characteristics of all participants split by insulin treatment status are shown in ESM Table 1. In the 13,085 participants with diabetes not currently insulin treated the mean T1DGRS (10.32, SD 2.38) was consistent with a classical non-T1D reference population (35) mean T1DGRS (10.37, SD 2.26) suggesting little to no T1D in this group. The genetically assessed estimated performance of classification approaches to classify all insulin treated diabetes cases as either T1D or T2D ranked by accuracy are shown in Table 2. The median classification accuracy was 85%, and varied substantially by approach (range 71% to 88%). The highest accuracy overall was insulin within a year of diagnosis combined with the clinical model, overall correctly classifying 87% and lipid model overall correctly classifying 88%. Interview reported diabetes type with or without insulin within a year of diagnosis had accuracy of 87% but was available in just 15% (519/3534) of all cases. For the majority of approaches, adding Insulin within a year of diagnosis to define T1D substantially improved accuracy with absence of OHA treatment only slightly less accurate, ESM table 2. The lowest accuracy was seen in approaches using simple cut-offs for individual variables: age of diagnosis (<35 years) 82% and BMI (≤25kg/m2) 71%.

In the 47% (1644/3534) of the insulin treated cohort with linked primary care data, 

diabetes diagnosis codes algorithm alongside insulin within a year of diagnosis gave the highest accuracy of approaches that incorporate electronic healthcare record data and diagnosis codes at 85%. For direct comparison ESM table 3 gives the performance of other classification approaches in this reduced subset of the dataset with linked primary care records, with results broadly similar in this subset.

Performance of approaches to classifying all insulin treated participants with diabetes in DARE

In the DARE cohort we identified 1241 people with diabetes who met our inclusion criteria, 63% (784/1241) were insulin treated with 42% (333/784) having a C-peptide <200 pmol/l consistent with T1D, at a median duration of 18 years. Table 3 gives the performance of classification approaches to classify all insulin treated diabetes cases as either T1D or T2D against a C-peptide definition of diabetes type. Accuracy values and overall ranking of approaches were similar to when diabetes type was defined genetically in UK Biobank, median accuracy 83% (range 68%-88%). In DARE the clinical model combined with insulin within a year of diagnosis had accuracy of 85%. Interview reported diabetes type alone gave the highest accuracy 88%. The Biobank algorithm (incorporating interview reported diabetes type) with Insulin within a year of diagnosis had accuracy of 87%. This reduced to 84% when interview reported diabetes type was not included within the algorithm. Again all methods were improved by adding insulin within a year of diagnosis. In the 451 non-insulin treated participants with C-peptide measured 99.6% (449/451) had a C-peptide ≥200 consistent with T2D.

Performance of approaches to optimally identify type 1 and type 2 diabetes amongst insulin treated participants with diabetes
The performance of methods to optimally identify T1D, ranked by PPV in UK Biobank (percent of those identified as T1D who have the condition genetically) are shown in Table 4. A pure T1D cohort was generated when insulin within a year of diagnosis was combined with either age at diagnosis ≤ 20 years (PPV 100%) or a clinical model probability ≥ 80% (PPV 99%). However, these approaches had low sensitivity respectively only identifying 33% and 37% of all T1D cases. Using probable T1D in the Biobank algorithm combined with insulin within a year of diagnosis identified 69% of all T1D cases, with a PPV of 90%. This was similar to using a lower clinical model probability of ≥ 20% identifying 67% of all T1D cases with a PPV of 91%. Comparable results for the majority of approaches for both PPV and sensitivity of T1D identified were achieved in DARE, using C-peptide defined diabetes type, Table 4.

Performance of methods to optimally identify T2D, ranked by PPV in UK Biobank (percent of those identified as T2D who have the condition genetically) are shown in ESM Table 4. A pure T2D cohort was generated using probable T2D within the Biobank Algorithm, PPV 100% but this had low sensitivity capturing just 17% of all insulin treated T2D cases. A clinical model probability < 5% gave a T2D PPV of 94% and captured 67% of all T2D cases. Adding absence of insulin within a year of diagnosis to all definitions of T2D increased T2D PPV in all approaches but resulted in a lower proportion of all T2D cases being captured. Comparable results for both PPV and sensitivity for each approach were achieved in DARE, using C-peptide defined diabetes type, ESM Table 4.

Performance of approaches to classifying all insulin treated participants with diabetes in UK Biobank

As an exploratory analysis we evaluated the performance of approaches to classify all participants with diabetes in UK Biobank regardless of ethnicity. Within the 4,845 insulin treated participants the overall performance of approaches was similar when the analysis was undertaken in just White Europeans, median accuracy 85% (range 75% to 91%) with the best accuracy achieved using probability models combined with insulin within a year of diagnosis: lipid model 91%, clinical features only model 90%, ESM Table 5.

Development of algorithm for optimal approach selection.

We developed a pragmatic online tool for researchers to select the optimum approach of those evaluated for classifying insulin treated diabetes cases in research datasets, based on findings in UK Biobank: Classifying Diabetes for Research: Method Selector (newcastlerse.github.io). The optimal approach varies based on the research question being asked and the diabetes outcomes available in the dataset being used. ESM appendix 2 provides researchers the R code to implement all methods which is also provided within the online tool.
Discussion

We evaluated the performance of approaches for classifying diabetes type in two different population level research datasets: UK Biobank and DARE. Results were consistent across datasets despite using two different biological definitions of diabetes type. The impact of classification approach selection on study results and conclusions is highlighted by the marked variation in accuracy observed in our study. Across the two different datasets combining insulin within a year of diagnosis with T1D models incorporating BMI and age at diagnosis (clinical model) and these features with lipids (lipid model) consistently achieved the highest accuracy for classifying all insulin treated participants (≥85%). Interview reported diabetes type showed similar accuracy in both UK Biobank and the DARE cohort but was only recorded in the minority (15%) of UK Biobank participants limiting its utility.

Our results suggest that probability models combined with insulin within a year of diagnosis provide a highly accurate approach to classifying research cohorts with insulin treated diabetes. As a simple alternative interview reported diabetes type can be used although this was only available in 15% of UK Biobank participants. Why a low percentage of participants reported diabetes type in UK Biobank is unclear. To explore this we compared those with and without an interview reported diabetes type which suggested a trend towards more T1D in those reporting a diabetes type at interview but no systematic bias (ESM table 6). Furthermore with recent changes to guidance for biomarker testing in national and international guidance (37-40) it is possible clinical diagnosis and therefore interview reported diagnosis may become more accurate over time. This study also highlights the limitations of using single cutoffs particularly age of diagnosis likely reflecting the finding that nearly half of all T1D cases occur after 30 years of age (3-5, 14). All approaches are improved by adding variables capturing either insulin within a year of diagnosis or current OHA treatment. It was possible to identify pure T1D cohorts in both datasets through use of a combination of early insulin treatment and either high model probability or very young age at diagnosis.

A key strength of our study was that performance was evaluated against biological definitions of diabetes type. This reduces the potential for inaccuracies and bias if testing against clinical definitions which are subject to both error and circularity (with features accurate for clinical classification reflecting features clinicians consider to be important) (7, 8, 10). The main analysis in UK Biobank was restricted to white European participants where the T1DGRS has been validated. As an exploratory analysis we evaluated all participants to show that the ranking of approaches remained similar (meaning the optimum approach remains valid) even if the absolute accuracy of approaches in all non-white European ethnicities should be interpreted with caution. Whilst all ethnicities were included in DARE 99% of participants were white European.

Few studies have compared different classification methods to robust biomarker defined diabetes types. In a cohort with insulin treated diabetes, Hope et al evaluated the performance of age of diagnosis <35 to classify diabetes cases with T1D defined by C-peptide deficiency and cases with preserved C-peptide defined as T2D (10). Age at diagnosis correctly classified 83% of all cases in their study comparable with our study: 82% in UK Biobank and 79% in DARE. This remained comparable when age of diagnosis was combined with insulin within a year of diagnosis: Hope et al study accuracy of 85% versus 84% UK Biobank and 82% DARE. Model performance was also high when
previously assessed against diabetes subtypes defined by pancreatic histology (26). The importance of insulin treatment in helping initially determine diabetes type in research datasets is emphasized by the genetic susceptibility of all participants not currently insulin treated being consistent with little to no T1D in this group. In DARE absence of insulin treatment was also almost never associated with C-peptide deficiency mirroring previous studies defining diabetes type using C-peptide (41).

Limitations of our study include the fact that both the Biobank algorithm (developed in UK Biobank) and the T1D clinical model (developed in a cohort that included DARE) were evaluated in the same cohorts they were developed in. Reassuringly both methods performed comparatively well in the alternative data set they were not developed in suggesting any bias was minimal. Despite using both T1D probability models in all participants even though they were developed in adults aged 18-50 they were consistently high performing approaches in both datasets (19, 20). It is possible accuracy could have been further improved by varying cutoffs in older adults however this would have risked being over fitted. Lipids in UK Biobank were also unfasted, in contrast to the model development dataset, and it is therefore possible performance would increase where fasted lipids are available (20). Using genetic predisposition to T1D can be helpful in diabetes classification; in the original development of the clinical model adding T1DGRS improved performance (19) and we would recommend using this when genetic data is available, however as T1DGRS was our outcome we were unable to evaluate this approach. Islet autoantibodies used in combination with clinical models also improve performance (19), but are rarely available in research datasets as is the case in UK Biobank. Classifying diabetes as only being T1D or T2D will miss other types of diabetes. Reassuringly in DARE just 2% (29/1241) of the cohort had a clinician diagnosis which was not T1D or T2D. T1DGRS is known to modestly reduce with increasing age of T1D diagnosis (42-44) and our T1D reference cohort were diagnosed <17 years of age. In previous studies the mean T1DGRS of those with confirmed T1D diagnosed over 18 years of age was 2.5% lower than those diagnosed <18 years (45). Given that over half of type 1 diabetes develops in adults this means our genetically estimated type 1 diabetes prevalence will be a slight underestimate. However the comparative performance results as in the same datasets remains robust and reassuringly in DARE defining diabetes type by C-peptide similar results were found. It is possible interview reported diabetes type could be influenced by the research staff conducting the interviews and there appears a subtle suggestion of bias towards T1D in those reporting versus not reporting diabetes type in UK Biobank. While other methods of collecting self-report may potentially have lower accuracy recent research has found similar PPV of 83% and sensitivity of 92% for type 1 diabetes when assessing self-reported diabetes type via a telephone survey’ (32). It has also previously been reported that UK Biobank is not truly representative of the UK population: participants being from less deprived areas, and more predominantly of white ethnicity than the general population (46). These issues with recruited population level research datasets are not unique to UK Biobank but caution should be used applying these finding to non-white or low income populations (47).

Our results are important for all researchers studying type 1 or 2 diabetes. The considerable differences in pathophysiology, treatment and associated risks of T1D and T2D means inadvertently studying mixed cohorts could lead to misleading study findings (48). Our results allow determination of the optimal approach for classifying insulin treated diabetes cases whilst also confirming that non-insulin treated cases of over three years duration can confidently be labelled as having T2D. Approaches can be selected based on which diabetes specific outcomes are available and the
research question being asked. An added advantage of our study is that researchers can understand the accuracy of the approach used and how this might impact their results and their relatability to other studies where different approaches may have been used. For ease our findings have been translated into an online tool allowing researchers to determine and then implement the optimal approach for their research question and dataset.

**Conclusion:**

Within two separate research datasets and using two different biological definitions of diabetes we show the performance of approaches for classifying insulin treated diabetes type for research studies and translate this into an online tool for optimal approach selection for researchers. Interview reported diabetes type diagnosis and models combining continuous features are the most accurate methods of classifying insulin treated diabetes in research datasets without measured classification biomarkers.
List of abbreviations

UKBB: UK Biobank
DARE: Diabetes Alliance for Research in England
T1D: Type 1 Diabetes
T2D: Type 2 Diabetes
T1DGRS: Type 1 Diabetes genetic risk score
BMI: Body mass index
Declarations

Ethics approval and consent to participate

Ethics for the Diabetes Alliance for Research in England (DARE) study was granted by the Devon & Torbay Research Ethics Committee, ref: 2002/7/118.

Availability of data and materials

UK Biobank data are available through a procedure described at http://www.ukBiobank.ac.uk/using-the-resource/.
DARE data are available through application to the Peninsula Research Bank https://exetercrfnihr.org/about/exeter-10000-prb/

Competing interests

AGJ contributed to the development of the two classification models assessed in this work. Other authors declare that there are no relationships or activities that might bias, or be perceived to bias, their work.

Funding

The Diabetes Alliance for Research in England (DARE) study was funded by the Wellcome Trust and supported by the Exeter NIHR Clinical Research Facility.

NJT is funded by a Wellcome Trust funded GW4 PhD. AM is supported by a National Institute for Health Research (NIHR) Academic Clinical Fellowship. M.N.W. is supported by the Wellcome Trust Institutional Support Fund (WT097835MF). SAS is supported by a Diabetes UK PhD studentship (17/0005757). JMD is supported by an Independent Fellowship funded by Research England’s Expanding Excellence in England (E3) fund. KGY is supported by Research England’s Expanding Excellence in England (E3) fund. ATH is supported by the NIHR Exeter Clinical Research Facility and a Wellcome Senior Investigator award and an NIHR Senior Investigator award. AGJ was supported by an NIHR Clinician Scientist award (CS-2015-15-018). The views given in this article do not necessarily represent those of the National Institute for Health Research, the National Health Service, or the Department of Health.

Authors Contributions

NJT, AM and AGJ designed the study. SAS, KGY and MNW acquired the data and SAS and MNW generated the T1DGRS. NJT, JD, AM and AGJ analysed the data. NJT wrote the first draft of the report. All authors reviewed the draft, contributed to the revision of the report and gave final approval for publication. AGJ and NJT are the guarantors of this work.

Acknowledgments

The authors thank participants who took part in the study and the research teams who undertook cohort recruitment.
This research has in part been conducted using UK Biobank Resource. Biobank application 9055.

The authors are grateful to Mike Simpson of Newcastle University for developing the online tool.
**Figures**

**Table 1:** *Diabetes specific factors required for each approach and the different cut offs required for classifying all cases, or defining type 1 diabetes or type 2 diabetes.* Where available cut offs were taken from existing literature (10, 15-19, 22). * For the previously published models, cut offs were not available for selecting pure T1D and T2D cohorts so pragmatic values were chosen from published data aiming for 100% and >90% PPV for T1D classification and 100% PPV for T2D classification (19).

<table>
<thead>
<tr>
<th>Reference name (approach number)</th>
<th>Clinical information required</th>
<th>Whole Cohort For defining type 1 diabetes remainder type 2 diabetes</th>
<th>For defining type 1 diabetes only</th>
<th>For defining type 2 diabetes only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (1)</td>
<td>Age at diagnosis &lt;35 years (10)</td>
<td>≤20 years (16)</td>
<td>≥40 years (10)</td>
<td></td>
</tr>
<tr>
<td>BMI (2)</td>
<td>Current BMI ≤25 kg/m² (14)</td>
<td>≤23 kg/m² (10)</td>
<td>≥28 kg/m² (10)</td>
<td></td>
</tr>
<tr>
<td>Clinical model (3)</td>
<td>Current BMI, age at diagnosis</td>
<td>Model probability ≥12% (19)</td>
<td>Model probability ≥80% *</td>
<td>Model probability &lt;5% *</td>
</tr>
<tr>
<td>Lipid model (4)</td>
<td>Current BMI, age at diagnosis, Sex, HDL, triglyceride, total Cholesterol</td>
<td>Model probability ≥12% (19)</td>
<td>Model probability ≥80% *</td>
<td>Model probability &lt;5% *</td>
</tr>
<tr>
<td>ICD codes (5)</td>
<td>(ICD 10 or 9 code), OHA, Age at diagnosis, DKA episode history</td>
<td>Algorithm Type 1 diabetes (18)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>UK Biobank Algorithm (6)</td>
<td>Age at diagnosis, time to Insulin, non metformin OHA, interview-report of type 1 diabetes, ethnicity</td>
<td>Possible and probable type 1 diabetes (15)</td>
<td>Probable type 1 diabetes (13)</td>
<td>Probable type 2 Diabetes (13)</td>
</tr>
<tr>
<td>Interview reported (7)</td>
<td>Interview reported diabetes type</td>
<td>Interview reported diabetes type 1 diabetes (32)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Diagnosis Codes algorithm (8)</td>
<td>Diabetes diagnosis codes, non metformin OHA, prescription for glucagon, prescription for urine acetone strip</td>
<td>Ratio of type 1 diabetes to type 2 diabetes diagnosis codes &gt;0.5 with either glucagon or non metformin OHA prescription or prescription of urine acetone strip alone (16)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Diagnosis code + age (9)</td>
<td>Diabetes diagnosis codes and age at diagnosis</td>
<td>Any diagnosis code of Type 1 diabetes or age at diagnosis &lt;22 years (17)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Majority diagnosis codes (10)</td>
<td>Diabetes diagnosis codes</td>
<td>Ratio of type 1 diabetes to type 2 diabetes diagnosis codes &gt;0.5 (22)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Table 2: Comparative performance of approaches classifying all insulin treated white European participants with diabetes in UK Biobank. Cases are classified as type 1 diabetes if they meet the stated criteria, and are otherwise classified as type 2 diabetes. Results ranked by accuracy (total correctly classified) then type 1 diabetes PPV. Brackets signify 95% CI, Positive predictive value (PPV).

<table>
<thead>
<tr>
<th>Approach</th>
<th>(n)</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>(n)</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipid model probability ≥12% and insulin within year of diagnosis</td>
<td>1169</td>
<td>87% (84-90)</td>
<td>79% (77-81)</td>
<td>2365</td>
<td>88% (86-91)</td>
<td>93% (92-94)</td>
<td>88%</td>
</tr>
<tr>
<td>Clinical model probability ≥12% and insulin a within year of diagnosis</td>
<td>1047</td>
<td>89% (86-92)</td>
<td>72% (70-75)</td>
<td>2487</td>
<td>86% (83-88)</td>
<td>95% (94-96)</td>
<td>87%</td>
</tr>
<tr>
<td>Interview reported diabetes type (n=519 available) and insulin within a year of diagnosis</td>
<td>224</td>
<td>85% (77-92)</td>
<td>86% (81-91)</td>
<td>295</td>
<td>89% (82-97)</td>
<td>89% (85-92)</td>
<td>87%</td>
</tr>
<tr>
<td>Interview reported diabetes type (n=519 available)</td>
<td>253</td>
<td>80% (73-87)</td>
<td>92% (88-95)</td>
<td>266</td>
<td>93% (86-101)</td>
<td>83% (79-87)</td>
<td>87%</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes and insulin within a year of diagnosis</td>
<td>988</td>
<td>90% (87-93)</td>
<td>69% (66-71)</td>
<td>2546</td>
<td>84% (82-87)</td>
<td>96% (95-96)</td>
<td>86%</td>
</tr>
<tr>
<td>ICD algorithm and insulin within a year of diagnosis</td>
<td>1025</td>
<td>89% (86-92)</td>
<td>71% (68-73)</td>
<td>2509</td>
<td>85% (82-87)</td>
<td>95% (94-96)</td>
<td>86%</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes and insulin within a year of diagnosis (no interview report)</td>
<td>918</td>
<td>93% (89-96)</td>
<td>66% (63-68)</td>
<td>2616</td>
<td>83% (81-85)</td>
<td>97% (96-98)</td>
<td>85%</td>
</tr>
<tr>
<td>ICD algorithm</td>
<td>1184</td>
<td>82% (79-85)</td>
<td>75% (73-78)</td>
<td>2350</td>
<td>86% (84-89)</td>
<td>91% (90-92)</td>
<td>85%</td>
</tr>
<tr>
<td>Age diabetes diagnosed &lt;35 years and insulin within a year of diagnosis</td>
<td>867</td>
<td>93% (89-96)</td>
<td>62% (59-65)</td>
<td>2667</td>
<td>82% (79-84)</td>
<td>97% (96-98)</td>
<td>84%</td>
</tr>
<tr>
<td>Lipid model probability ≥12%</td>
<td>1501</td>
<td>74% (71-77)</td>
<td>86% (84-88)</td>
<td>2033</td>
<td>91% (89-94)</td>
<td>83% (81-84)</td>
<td>84%</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes (no interview report)</td>
<td>1142</td>
<td>80% (77-83)</td>
<td>70% (68-73)</td>
<td>2392</td>
<td>84% (81-87)</td>
<td>90% (88-91)</td>
<td>83%</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes</td>
<td>1231</td>
<td>78% (75-81)</td>
<td>74% (72-77)</td>
<td>2303</td>
<td>86% (83-88)</td>
<td>88% (87-89)</td>
<td>83%</td>
</tr>
<tr>
<td>Clinical model probability ≥12%</td>
<td>1325</td>
<td>76% (73-79)</td>
<td>78% (76-80)</td>
<td>2209</td>
<td>87% (84-90)</td>
<td>86% (84-87)</td>
<td>83%</td>
</tr>
<tr>
<td>Age diabetes diagnosed &lt;35 years</td>
<td>1065</td>
<td>80% (77-84)</td>
<td>66% (64-69)</td>
<td>2469</td>
<td>82% (80-85)</td>
<td>91% (89-92)</td>
<td>82%</td>
</tr>
<tr>
<td>BMI ≤25 (kg/m2) and insulin within a year of diagnosis</td>
<td>511</td>
<td>80% (75-85)</td>
<td>32% (29-34)</td>
<td>3023</td>
<td>71% (68-73)</td>
<td>95% (95-96)</td>
<td>72%</td>
</tr>
<tr>
<td>BMI ≤25 (kg/m2)</td>
<td>658</td>
<td>70% (65-74)</td>
<td>35% (33-38)</td>
<td>2876</td>
<td>71% (68-73)</td>
<td>91% (90-92)</td>
<td>71%</td>
</tr>
</tbody>
</table>
Table 3: Comparative performance of approaches classifying all insulin treated participants with diabetes in DARE. Cases are classified as type 1 diabetes if they meet the stated criteria, and are otherwise classified as type 2 diabetes. Results ranked by accuracy (total correctly classified) then type 1 diabetes PPV. Brackets signify 95% CI, Positive predictive value (PPV). Lipid model and Diagnosis codes not evaluated as unavailable in DARE.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Called Type 1 Diabetes</th>
<th>Called Type 2 Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n)</td>
<td>PPV</td>
</tr>
<tr>
<td>Interview reported diabetes type and insulin within a year of diagnosis</td>
<td>310</td>
<td>89% (85-92)</td>
</tr>
<tr>
<td>Interview reported diabetes type</td>
<td>335</td>
<td>86% (83-90)</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes and insulin within a year</td>
<td>325</td>
<td>86% (82-90)</td>
</tr>
<tr>
<td>(including interview report)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical model probability ≥12% and insulin a within year of diagnosis</td>
<td>278</td>
<td>90% (86-93)</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes and insulin within a year</td>
<td>257</td>
<td>90% (87-94)</td>
</tr>
<tr>
<td>(no interview report)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes (including interview report)</td>
<td>392</td>
<td>76% (72-80)</td>
</tr>
<tr>
<td>Age diabetes diagnosed &lt;35 years and insulin within a year of diagnosis</td>
<td>242</td>
<td>90% (86-94)</td>
</tr>
<tr>
<td>Clinical model probability ≥12%</td>
<td>346</td>
<td>78% (74-82)</td>
</tr>
<tr>
<td>UK Biobank probable &amp; possible type 1 diabetes (no interview report)</td>
<td>300</td>
<td>81% (77-85)</td>
</tr>
<tr>
<td>Age diabetes diagnosed &lt;35 years</td>
<td>280</td>
<td>80% (76-85)</td>
</tr>
<tr>
<td>BMI ≤25 (kg/m2) and insulin within a year of diagnosis</td>
<td>140</td>
<td>87% (82-93)</td>
</tr>
<tr>
<td>BMI ≤25 (kg/m2)</td>
<td>187</td>
<td>72% (66-79)</td>
</tr>
</tbody>
</table>
Table 4: Comparative performance of approaches classifying type 1 diabetes with minimal misclassification in UK Biobank and DARE in insulin treated participants. Cases are classified as type 1 diabetes if they meet the stated criteria, and are otherwise classified as type 2 diabetes. Results ranked in UK Biobank by type 1 diabetes PPV then sensitivity for identifying type 1 diabetes. Analysis in UK Biobank restricted to White Europeans. Cut offs chosen to give high T1D PPV as per table 1. Brackets signify 95% CI, Positive predictive value (PPV).

<table>
<thead>
<tr>
<th>Approach</th>
<th>UK BIObANK</th>
<th>DARE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPV of cases called type 1 diabetes</td>
<td>Sensitivity for identifying type 1 diabetes</td>
</tr>
<tr>
<td>Age diabetes diagnosed ≤20 years and insulin within a year of diagnosis</td>
<td>100% (99-100)</td>
<td>33% (30-35)</td>
</tr>
<tr>
<td>Clinical model probability ≥80% and insulin within a year of diagnosis</td>
<td>99% (98-100)</td>
<td>37% (34-39)</td>
</tr>
<tr>
<td>Lipid model probability ≥80% and insulin within a year of diagnosis</td>
<td>97% (95-98)</td>
<td>40% (38-43)</td>
</tr>
<tr>
<td>Lipid model probability ≥80%</td>
<td>92% (90-94)</td>
<td>42% (39-45)</td>
</tr>
<tr>
<td>Age diabetes diagnosed ≤20 years</td>
<td>92% (90-95)</td>
<td>34% (31-36)</td>
</tr>
<tr>
<td>Clinical model probability ≥20% and insulin within a year of diagnosis</td>
<td>91% (89-92)</td>
<td>67% (65-70)</td>
</tr>
<tr>
<td>Clinical model probability ≥80%</td>
<td>91% (88-93)</td>
<td>37% (35-40)</td>
</tr>
<tr>
<td>UK Biobank probable type 1 diabetes and insulin within a year of diagnosis</td>
<td>90% (88-92)</td>
<td>69% (66-71)</td>
</tr>
<tr>
<td>Lipid model probability ≥20% and insulin within a year of diagnosis</td>
<td>89% (87-91)</td>
<td>75% (73-77)</td>
</tr>
<tr>
<td>UK Biobank probable type 1 diabetes</td>
<td>89% (87-91)</td>
<td>70% (67-72)</td>
</tr>
<tr>
<td>BMI ≤23 (kg/m2) and insulin within a year of diagnosis</td>
<td>82% (78-87)</td>
<td>16% (14-18)</td>
</tr>
<tr>
<td>Interview reported type 1 diabetes</td>
<td>80% (75-85)</td>
<td>92% (88-95)</td>
</tr>
<tr>
<td>Lipid model probability ≥20%</td>
<td>80% (78-82)</td>
<td>81% (79-84)</td>
</tr>
<tr>
<td>Clinical model probability ≥20%</td>
<td>80% (78-83)</td>
<td>71% (69-74)</td>
</tr>
<tr>
<td>BMI ≤23 (kg/m2)</td>
<td>75% (70-80)</td>
<td>17% (15-19)</td>
</tr>
</tbody>
</table>
References

Highlights

What is new

Key findings:

Across two different datasets classification models incorporating continuous clinical features combined with early insulin requirement, or (where available) interview reported diabetes type, consistently achieved high accuracy (≥85%).

When identifying a type 1 diabetes cohort with minimal misclassification, young age at diagnosis (<20 years) or models with high thresholds had very high predictive value, but modest sensitivity.

What this adds to what is known:

The best approaches for classifying diabetes type in population datasets was previously unclear. This work allows researchers to identify the optimum classification approach for their dataset and research question.

Implications

The optimal method for identifying diabetes subtypes in observational data will depend on available data and research question. Researchers can select the optimum approach using an online tool devised using the study findings (Classifying Diabetes for Research: Method Selector (newcastlerse.github.io).
Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

| AGJ contributed to the development of the two classification models assessed in this work. Other authors declare that there are no relationships or activities that might bias, or be perceived to bias, their work. |
Nicholas J Thomas: Conceptualization, Data curation, Formal analysis, Visualization, Writing Original Draft
Andrew McGovern Conceptualization, visualization, Katherine G Young Data curation, Software, Seth A Sharp Data curation Formal Analysis, Michael N Weedon Data curation, Formal Analysis, Andrew T Hattersley Visualization, Supervision John Dennis Conceptualization, visualization, Angus G Jones Conceptualization, Visualization, Writing Original Draft, Supervision. All authors: Writing Review & Editing