

Identifying type 1 and 2 diabetes in population level data: assessing the accuracy of published approaches

Nicholas J Thomas^{1,2}, Andrew McGovern^{1,2}, Katherine G Young¹, Seth A Sharp¹, Michael N Weedon¹, Andrew T Hattersley^{1,2}, John Dennis¹, 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: 304

Manuscript: 4279

Tables: 4

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract:

Aims

Population datasets are increasingly used to study type 1 or 2 diabetes, and inform clinical practice. However, correctly classifying diabetes type, when insulin treated, in population datasets is challenging. Many different approaches have been proposed, ranging from simple age or BMI cut offs, to complex algorithms, and the optimal approach is unclear. We aimed to compare the performance of approaches for classifying insulin treated diabetes for research studies, evaluated against two independent biological definitions of diabetes type.

Method

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

Results

Accuracy of approaches ranged from 71%-88% in UKBB and 68%-88% in DARE. All approaches were improved by combining with requirement for early insulin treatment (<1 year from diagnosis). When classifying all participants, combining early insulin requirement with a type 1 diabetes probability model incorporating continuous clinical features (diagnosis age and BMI only) consistently achieved high accuracy, (UKBB 87%, DARE 85%). Self-reported diabetes type alone had high accuracy (UKBB 87%, DARE 88%) but was available in just 15% of UKBB participants. For identifying type 1 diabetes with minimal misclassification, using models with high thresholds or young age at diagnosis (<20 years) had the highest performance. An online tool developed from all UKBB findings allows the optimum approach of those tested to be selected based on variable availability and the research aim.

Conclusion

Self-reported diagnosis and models combining continuous features with early insulin requirement are the most accurate methods of classifying insulin treated diabetes in research datasets without measured classification biomarkers.

Background

Robustly classifying diabetes type in population level data is challenging

Large population level 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 $\geq 98\%$ 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]. The clinical diagnosis of insulin treated diabetes cases is frequently not available in population data, and if available will include substantial misclassification and miscoding ($\approx 15\%$)[4-8]. In population data, biomarkers which can help improve classification, such as C-peptide or islet autoantibodies [9], are rarely available.

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

The optimum approach for classifying T1D and T2D in population data remains unclear. Previously published approaches vary and include: clinician or self-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 [10-21]. Where the performance of these approaches has been assessed, it is normally against a clinical assessment of T1D or T2D diagnosis [7, 10, 11, 16-21]. 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 [12, 15, 22]. 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 studies using large datasets, we aimed to compare the performance of a number of published approaches for classifying insulin treated diabetes in two population level 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 cohorts 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 [23] 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 population cohort with diabetes (the DARE cohort) against diabetes type defined by C-peptide level measured after a median 14 years duration.

Study design and participants

UK Biobank

UK Biobank recruited a population cohort of more than 500,000 people aged between 40 and 70 years registered with the UK National Health Service [24]. We evaluated a subset of 26,399 unrelated individuals self-reporting diabetes. To allow direct comparison of classification approaches in the same cohort, individuals were excluded where missing BMI measurement ($n=237$) or self-reported 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 [25, 26]. 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 [27]. A separate secondary exploratory analysis was undertaken including all 23,098 participants of all ethnicities. Clinical history was self-reported 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 through 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) [5]. We evaluated 1,296 participants (22% (1296/5991) of the DARE cohort) with measured C-peptide. C-peptide was measured on stored non fasting EDTA at DARE recruitment after January 2010 as previously described (see ESM) [5]. 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 [9]. 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 [5].

Assessment of population approaches for classifying diabetes type in insulin treated individual

Overall we compared ten different approaches for the classification of insulin treated diabetes with the variables required for each approach 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 [7, 10-12, 15]. 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). The approaches evaluated were: age at diagnosis alone ((approach 1: *age*), BMI alone (approach 2: *BMI*), a T1D probability clinical features model, incorporating BMI and age at diagnosis (approach 3: *clinical model*) [12, 22], a T1D probability model incorporating BMI and age at diagnosis, sex, HDL, triglycerides and total cholesterol (approach 4: *lipid model*) [15], a tree structured model using hospital episode statistics (HES) data including ICD10 and ICD9 diabetes codes and history of diabetic ketoacidosis (DKA) (approach 5: *ICD Algorithm*) [18] and an algorithm developed for classifying diabetes type within UK Biobank (approach 6: *Biobank algorithm*) [10]. The UK Biobank algorithm defines diabetes as probable or possible T1D or T2D using three different flow charts incorporating age at diagnosis, time to Insulin, non-metformin oral hypoglycaemic agents (OHA), self-report of type 1 diabetes and ethnicity. In our study the UK Biobank algorithm was evaluated with and without the addition of self-reported diabetes type. For the model incorporating lipids, in the 26% (1262/4845) of UK Biobank participants where lipids were not available, instead of excluding these cases, their calculated clinical model probability as above (BMI and age at diagnosis) was substituted, allowing classification of all participants. While both the clinical and lipid models can also incorporate islet antibodies and T1DGRS this was not assessed due to lack of islet autoantibodies and the use of T1DGRS for outcome definition in UK biobank, and the performance incorporating these markers in DARE being already published [12, 15]. Self-reported diabetes type as a single marker was also evaluated (approach 7: *Self*).

In the subset of UK Biobank participants with linked primary care data, recorded diagnostic codes were used to classify diabetes type, within the following approaches: a classification algorithm incorporating diagnosis codes and prescriptions for non-metformin OHA, glucagon or urine acetone (ketone) strips [11] ((approach 8: *diagnosis codes algorithm*), diagnosis codes used alone [17] (approach 9: *diagnostic codes alone*) and diagnostic code in combination with age of diagnosis [16] (approach 10: *diagnosis codes and age*). Linked primary care data were not available in DARE. The code for all approaches is provided in ESM appendix 2.

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 [12]: T1D $\geq 80\%$ 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. Rapid insulin requirement and OHA treatment are well reported to associate with T1D and T2D respectively. Therefore as an additional analysis performance of approaches was further evaluated with the addition of knowledge of rapid insulin requirement defined as insulin treatment within a year of diagnosis, or also by current treatment with any OHA.

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 population with diabetes can allow the proportion of T1D in that population to be determined [23, 28]. Genetic risk cannot be used in isolation to define diabetes type at an individual level as a high genetic susceptibility for T1D does not prevent a person having T2D [29]. However, within a cohort the average distribution of T1DGRS will reflect the proportion of T1D to T2D cases within it. A T1D cohort defined by an accurate classification approach will have a high average genetic susceptibility to T1D, closely mirroring the genetic risk of a reference cohort with confirmed T1D [3]. Conversely a T1D cohort defined by an inaccurate classification approach will have a lower average genetic susceptibility to T1D, as it will also include T2D cases with neutral genetic susceptibility for T1D. This concept allows the proportion of T1D to T2D within cohorts defined by different classification approaches to be estimated and the relative performance of different approaches to be compared. The main analysis was restricted to participants of white European descent, where the T1DGRS is validated, with a separate secondary exploratory analysis undertaken in all ethnicities.

We generated a 67 SNP type 1 diabetes T1DGRS as previously described [30], see ESM methods for further detail.

DARE:

T1D was defined as severe insulin deficiency: measured non-fasting C-peptide <200pmol/L. Type 2 diabetes was defined as participants currently insulin treated with a C-peptide \geq 200pmol/L. All analysed participants had a duration of diabetes at C-peptide measurement of over three years [9]. C-peptide was measured on stored non-fasted EDTA at DARE recruitment and as part of routine care ESM and as previously described [5].

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 [31]. 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 calculated according to the normalised difference of each clinical definitions mean T1DGRS ($Approach_{Called (T1D/T2D)}$) and the mean T1DGRS of the two reference populations ($Reference_{T1D}$ and $Reference_{T2D}$) in the equations below and as described previously [28, 32]. 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 $1 - Proportion_{T1D}$.

$$Proportion (PPV)_{T1D} = \left| \frac{Approach_{T1DGRS \text{ Called } T1D} - Reference_{T1DGRS \text{ } T2D}}{Reference_{T1DGRS \text{ } T1D} - Reference_{T1DGRS \text{ } T2D}} \right|$$

$$Proportion (PPV)_{T2D} = 1 - \left| \frac{Approach_{T1DGRS \text{ Called } T2D} - Reference_{T1DGRS \text{ } T2D}}{Reference_{T1DGRS \text{ } T1D} - Reference_{T1DGRS \text{ } T2D}} \right|$$

Calculating accuracy in UK Biobank and DARE

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

$$Accuracy = \left| \frac{(PPV_{T1D} \times n_{T1D}) + (PPV_{T2D} \times n_{T2D})}{(n_{T1D} + n_{T2D})} \right|.$$

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.

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 [31] mean T1DGRS (10.37 SD 2.26) suggesting little to no T1D in this group. The genetically assessed 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 rapid insulin requirement combined with the *clinical model*, overall correctly classifying 87% and *lipid model* overall correctly classifying 88%. Self-reported diabetes type with or without the addition of rapid insulin had accuracy of 87% but was available in just 15% (519/3534) of all cases. For the majority of approaches, adding rapid insulin treatment requirement 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 ($\leq 25\text{kg}/\text{m}^2$)* 71%. In the 47% (1644/3534) of the insulin treated cohort with linked primary care data, diabetes *diagnosis codes algorithm* alongside rapid insulin requirement 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, results were broadly similar in this subset, again models with and without lipids, combined with early insulin treatment having the highest overall accuracy of 88% and 87% respectively.

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 rapid insulin requirement had accuracy of 85%. Self-reported diabetes type alone gave the highest accuracy 88%. The Biobank algorithm (incorporating self-reported diabetes type) with rapid insulin had accuracy of 87%. This reduced to 84% when self-reported diabetes type was not included within the algorithm. Again all methods were improved by adding rapid insulin requirement. 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 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 rapid insulin requirement was combined with either age at diagnosis ≤ 20 years (PPV 100%) or a clinical model probability $\geq 80\%$ (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 rapid insulin requirement identified 69% of all T1D cases, with a PPV of 90%. This was similar to using a lower clinical model probability of $\geq 20\%$ 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 approaches to optimally identify type 2 diabetes within insulin treated participants with diabetes

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 rapid insulin requirement 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, 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 to when undertaken in just White Europeans, median accuracy 85% (range 75% to 91%) with the best accuracy achieved using probability models combined with rapid insulin requirement: 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 population datasets, based on findings in UK Biobank: [Classifying Diabetes for Research: Method Selector \(newcastlelse.github.io\)](https://newcastlelse.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 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 rapid insulin requirement 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 ($\geq 85\%$). Self-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 when available self-reported diabetes type can be used to classify population cohorts with insulin treated diabetes. When self-reported diabetes type is unavailable, as seen in UK Biobank in the majority of cases, probability models combined with rapid insulin requirement provide a highly accurate simple alternative. Approaches combining different variables had higher accuracy than cut-offs applied to single variables such as BMI or age at diagnosis. All approaches are improved by adding variables capturing either rapid insulin requirement or current OHA treatment. It was possible to identify pure T1D cohorts in both datasets though 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) [4, 5, 7]. 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 [7]. Age at diagnosis correctly classified 83% of all cases in their study comparable with in our study: 82% in UK Biobank and 79% in DARE. This remained comparable when age of diagnosis was combined with rapid insulin requirement: 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 in the NPOD consortium [22]. Because knowledge of time to insulin treatment is not always available, we also evaluated absence of oral hypoglycemic treatment to define T1D showing this was only slightly less accurate than rapid insulin requirement. The importance of insulin treatment in helping initially determine diabetes type in population 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 despite use of a pragmatic random sample, mirroring previous studies defining diabetes type using C-peptide [33].

Limitations of our study include that T1DGRS is known to modestly reduce with increasing age of T1D diagnosis, [34-37] and our T1D reference T1DGRS cohort were all diagnosed with T1D under 17 years of age [30]. In previous studies the mean T1DGRS of those with confirmed type 1 diabetes diagnosed over 30 years of age was 2.5% lower than those diagnosed <18 years [38]. In UK Biobank the majority (90% (3194/3534)) of insulin treated participants were aged >17 years at diagnosis, making our genetic estimates of T1D proportions slight underestimates. However, as all approaches were evaluated within the same dataset the comparative performance results remain robust and reassuringly similar results were also found defining diabetes type by C-peptide in DARE. 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 [12, 15]. 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 [15]. Using genetic predisposition to T1D can be helpful in diabetes classification; in the original development of the clinical model adding T1DGRS improved performance [12] 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 [12], but are rarely available in population data 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. Finally we have only tested one of a number of published algorithms for using primary care electronic healthcare records (available only in a subset of UK Biobank participants, and unavailable in the DARE cohort). The approach proposed by Klompas using primary care data [11] has been extensively used, and therefore was included as an example of an approach that makes use of the additional data available in electronic healthcare records.

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 [39]. 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 population 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. Self-reported diagnosis and models combining continuous features are the most accurate methods of classifying insulin treated diabetes in research datasets without measured classification biomarkers.

Contributors

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.

Declaration of interest

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.

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.

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.

Data Availability

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/>

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 [7, 10-12, 16-18]. * For the models previously published 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 [12].

Reference name (approach number)	Clinical information required	Cut offs used and reference code		
		Whole Cohort For defining type 1 diabetes remainder type 2 diabetes	For defining type 1 diabetes only	For defining type 2 diabetes only
Age (1)	Age at diagnosis	<35 years ⁽⁷⁾	≤20 years ⁽¹¹⁾	≥40 years ⁽⁷⁾
BMI (2)	Current BMI	≤25 kg/m ² ⁽¹¹⁾	≤23 kg/m ² ⁽⁷⁾	≥28 kg/m ² ⁽⁷⁾
Clinical model (3)	Current BMI, age at diagnosis	Model probability ≥12% ⁽¹²⁾	Model probability ≥80% *	Model probability <5% *
Lipid model (4)	Current BMI, age at diagnosis, Sex, HDL, triglyceride, total Cholesterol	Model probability ≥12% ⁽¹²⁾	Model probability ≥80% *	Model probability <5% *
ICD codes (5)	(ICD 10 or 9 code), OHA, Age at diagnosis, DKA episode history	Algorithm Type 1 diabetes ⁽¹⁸⁾	N/A	N/A
UK Biobank Algorithm (6)	Age at diagnosis, time to Insulin, non metformin OHA, self-report of type 1 diabetes, ethnicity	Possible and probable type 1 diabetes ⁽¹⁰⁾	Probable type 1 diabetes ⁽¹⁰⁾	Probable type 2 Diabetes ⁽¹⁰⁾
Self (7)	Self reported diabetes type	Self reported type 1 diabetes	N/A	N/A
Diagnosis Codes algorithm (8)	Diabetes diagnosis codes, non metformin OHA, prescription for glucagon, prescription for urine acetone strip	Ratio of type 1 diabetes to type 2 diabetes diagnosis codes >0.5 with either glucagon or non metformin OHA prescription or prescription of urine acetone strip alone ⁽¹¹⁾	N/A	N/A
Diagnosis code + age (9)	Diabetes diagnosis codes and age at diagnosis.	Any diagnosis code of Type 1 diabetes or age at diagnosis ⁽¹⁶⁾	N/A	N/A
Majority diagnosis codes (10)	Diabetes diagnosis codes	Ratio of type 1 diabetes to type 2 diabetes diagnosis codes >0.5 ⁽¹⁷⁾	N/A	N/A

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).

Approach	Called Type 1 Diabetes			Called Type 2 Diabetes			Accuracy
	(n)	PPV	Sensitivity	(n)	PPV	Sensitivity	
Lipid model probability $\geq 12\%$ and insulin within year of diagnosis	1169	87% (84-90)	79% (77-81)	2365	88% (86-91)	93% (92-94)	88%
Clinical model probability $\geq 12\%$ and insulin a within year of diagnosis	1047	89% (86-92)	72% (70-75)	2487	86% (83-88)	95% (94-96)	87%
Self reported diabetes type (n=519 available) and insulin within a year of diagnosis	224	85% (77-92)	86% (81-91)	295	89% (82-97)	89% (85-92)	87%
Self reported diabetes type (n=519 available)	253	80% (73-87)	92% (88-95)	266	93% (86-101)	83% (79-87)	87%
UK Biobank probable & possible type 1 diabetes and insulin within a year of diagnosis	988	90% (87-93)	69% (66-71)	2546	84% (82-87)	96% (95-96)	86%
ICD algorithm and insulin within a year of diagnosis	1025	89% (86-92)	71% (68-73)	2509	85% (82-87)	95% (94-96)	86%
UK Biobank probable & possible type 1 diabetes and insulin within a year of diagnosis (no self report)	918	93% (89-96)	66% (63-68)	2616	83% (81-85)	97% (96-98)	85%
ICD algorithm	1184	82% (79-85)	75% (73-78)	2350	86% (84-89)	91% (90-92)	85%
Age diabetes diagnosed <35 years and insulin within a year of diagnosis	867	93% (89-96)	62% (59-65)	2667	82% (79-84)	97% (96-98)	84%
Lipid model probability $\geq 12\%$	1501	74% (71-77)	86% (84-88)	2033	91% (89-94)	83% (81-84)	84%
UK Biobank probable & possible type 1 diabetes (no self report)	1142	80% (77-83)	70% (68-73)	2392	84% (81-87)	90% (88-91)	83%
UK Biobank probable & possible type 1 diabetes	1231	78% (75-81)	74% (72-77)	2303	86% (83-88)	88% (87-89)	83%
Clinical model probability $\geq 12\%$	1325	76% (73-79)	78% (76-80)	2209	87% (84-90)	86% (84-87)	83%
Age diabetes diagnosed <35 years	1065	80% (77-84)	66% (64-69)	2469	82% (80-85)	91% (89-92)	82%
BMI ≤ 25 (kg/m ²) and insulin within a year of diagnosis	511	80% (75-85)	32% (29-34)	3023	71% (68-73)	95% (95-96)	72%
BMI ≤ 25 (kg/m ²)	658	70% (65-74)	35% (33-38)	2876	71% (68-73)	91% (90-92)	71%

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.

Approach	Called Type 1 Diabetes			Called Type 2 Diabetes			Accuracy
	(n)	PPV	Sensitivity	(n)	PPV	Sensitivity	
Self reported diabetes type and insulin within a year of diagnosis	310	89% (85-92)	82% (78-86)	474	88% (85-91)	92% (90-95)	88%
Self reported diabetes type	335	86% (83-90)	87% (83-90)	449	90% (87-93)	90% (87-93)	88%
UK Biobank probable & possible type 1 diabetes and insulin within a year of diagnosis	325	86% (82-90)	84% (80-88)	459	88% (85-91)	90% (87-93)	87%
Clinical model probability $\geq 12\%$ and insulin a within year of diagnosis	278	90% (86-93)	75% (70-79)	506	83% (80-86)	94% (91-96)	85%
UK Biobank probable & possible type 1 diabetes and insulin within a year of diagnosis (no self report)	257	90% (87-94)	69% (65-74)	527	81% (77-84)	94% (92-97)	84%
UK Biobank probable & possible type 1 diabetes	392	76% (72-80)	89% (86-92)	392	91% (88-93)	79% (75-83)	83%
Age diabetes diagnosed <35 years and insulin within a year of diagnosis	242	90% (86-94)	65% (60-70)	542	79% (75-82)	95% (93-97)	82%
Clinical model probability $\geq 12\%$	346	78% (74-82)	81% (77-85)	438	85% (82-89)	83% (80-87)	82%
UK Biobank probable & possible type 1 diabetes (no self report)	300	81% (77-85)	73% (68-78)	484	81% (78-85)	87% (84-90)	81%
Age diabetes diagnosed <35 years	280	80% (76-85)	67% (62-72)	504	78% (75-82)	88% (85-91)	79%
BMI ≤ 25 (kg/m ²) and insulin within a year of diagnosis	140	87% (82-93)	37% (31-42)	644	67% (63-71)	96% (94-98)	71%
BMI ≤ 25 (kg/m ²)	187	72% (66-79)	40% (35-46)	597	67% (63-70)	88% (85-91)	68%

Table 4: Comparative performance of approaches classifying type 1 diabetes 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. Brackets signify 95% CI, Positive predictive value (PPV).

Approach	UK BIOBANK		DARE	
	PPV of cases called type 1 diabetes	Sensitivity for identifying type 1 diabetes	PPV of cases called type 1 diabetes	Sensitivity for identifying type 1 diabetes
Age diabetes diagnosed ≤ 20 years and insulin within a year of diagnosis	100% (99-100)	33% (30-35)	96% (93-100)	40% (32-49)
Clinical model probability $\geq 80\%$ and insulin within a year of diagnosis	99% (98-100)	37% (34-39)	96% (93-99)	47% (39-54)
Lipid model probability $\geq 80\%$ and insulin within a year of diagnosis	97% (95-98)	40% (38-43)	n/a	n/a
Lipid model probability $\geq 80\%$	92% (90-94)	42% (39-45)	n/a	n/a
Age diabetes diagnosed ≤ 20 years	92% (90-95)	34% (31-36)	96% (92-99)	40% (32-49)
Clinical model probability $\geq 20\%$ and insulin within a year of diagnosis	91% (89-92)	67% (65-70)	91% (88-95)	70% (65-76)
Clinical model probability $\geq 80\%$	91% (88-93)	37% (35-40)	93% (90-97)	47% (39-55)
UK Biobank probable type 1 diabetes and insulin within a year of diagnosis	90% (88-92)	69% (66-71)	86% (82-90)	84% (80-88)
Lipid model probability $\geq 20\%$ and insulin within a year of diagnosis	89% (87-91)	75% (73-77)	n/a	n/a
UK Biobank probable type 1 diabetes	89% (87-91)	70% (67-72)	84% (80-88)	88% (84-91)
BMI ≤ 23 (kg/m ²) and insulin within a year of diagnosis	82% (78-87)	16% (14-18)	90% (83-97)	19% (10-28)
Self reported type 1 diabetes	80% (75-85)	92% (88-95)	86% (83-90)	87% (83-90)
Lipid model probability $\geq 20\%$	80% (78-82)	81% (79-84)	n/a	n/a
Clinical model probability $\geq 20\%$	80% (78-83)	71% (69-74)	84% (80-88)	74% (69-79)
BMI ≤ 23 (kg/m ²)	75% (70-80)	17% (15-19)	79% (70-87)	20% (11-28)

References

- [1] Group SDD (2019) Scottish Diabetes Survey 2019. In:
- [2] American Diabetes A (2017) 2. Classification and Diagnosis of Diabetes. *Diabetes care* 40(Suppl 1): S11-S24. 10.2337/dc17-S005
- [3] 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. *Lancet Diabetes Endocrinol* 6(2): 122-129. 10.1016/S2213-8587(17)30362-5
- [4] Foteinopoulou E, Clarke CAL, Pattenden RJ, et al. (2020) Impact of routine clinic measurement of serum C-peptide in people with a clinician-diagnosis of type 1 diabetes. *Diabetic medicine : a journal of the British Diabetic Association*: e14449. 10.1111/dme.14449
- [5] Thomas NJ, Lynam AL, Hill AV, et al. (2019) Type 1 diabetes defined by severe insulin deficiency occurs after 30 years of age and is commonly treated as type 2 diabetes. *Diabetologia* 62(7): 1167-1172. 10.1007/s00125-019-4863-8
- [6] Munoz C, Floreen A, Garey C, et al. (2019) Misdiagnosis and Diabetic Ketoacidosis at Diagnosis of Type 1 Diabetes: Patient and Caregiver Perspectives. *Clinical diabetes : a publication of the American Diabetes Association* 37(3): 276-281. 10.2337/cd18-0088
- [7] 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. *Brit J Gen Pract* 66(646): E315-E322. 10.3399/bjgp16X684961
- [8] Stone MA, Camosso-Stepinovic J, Wilkinson J, de Lusignan S, Hattersley AT, Khunti K (2010) Incorrect and incomplete coding and classification of diabetes: a systematic review. *Diabetic medicine : a journal of the British Diabetic Association* 27(5): 491-497. 10.1111/j.1464-5491.2009.02920.x
- [9] Jones AG, Hattersley AT (2013) The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabetic medicine : a journal of the British Diabetic Association* 30(7): 803-817. 10.1111/dme.12159
- [10] Eastwood SV, Mathur R, Atkinson M, et al. (2016) Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PloS one* 11(9): e0162388. 10.1371/journal.pone.0162388
- [11] Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R (2013) Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes care* 36(4): 914-921. 10.2337/dc12-0964
- [12] Lynam A, McDonald T, Hill A, et al. (2019) Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18-50 years. *BMJ open* 9(9): e031586. 10.1136/bmjopen-2019-031586
- [13] Practitioners RCoG (2011) Coding, classification and diagnosis of diabetes A review of the coding, classification and diagnosis of diabetes in primary care in England with recommendations for improvement. In:
- [14] Hippisley-Cox J, Coupland C (2015) Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: cohort study. *BMJ* 351: h5441. 10.1136/bmj.h5441
- [15] Lynam AL, Dennis JM, Owen KR, et al. (2020) Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn Progn Res* 4: 6. 10.1186/s41512-020-00075-2
- [16] Lethebe BC, Williamson T, Garies S, et al. (2019) Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. *CMAJ Open* 7(2): E246-E251. 10.9778/cmajo.20180142
- [17] Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA (2018) Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data.

Pharmacoepidemiol Drug Saf 27(10): 1053-1059. 10.1002/pds.4377

[18] Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA (2011) Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. *J Diabetes Sci Technol* 5(3): 486-493. 10.1177/193229681100500303

[19] Weisman A, Tu K, Young J, et al. (2020) Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. *BMJ Open Diabetes Res Care* 8(1). 10.1136/bmjdr-2020-001224

[20] Zhong VW, Pfaff ER, Beavers DP, et al. (2014) Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study. *Pediatr Diabetes* 15(8): 573-584. 10.1111/pedi.12152

[21] Sharma M, Petersen I, Nazareth I, Coton SJ (2016) An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clin Epidemiol* 8: 373-380. 10.2147/CLEP.S113415

[22] Carr ALJ, Perry DJ, Lynam AL, et al. (2020) Histological validation of a type 1 diabetes clinical diagnostic model for classification of diabetes. *Diabetic medicine : a journal of the British Diabetic Association* 37(12): 2160-2168. 10.1111/dme.14361

[23] Thomas NJ, Jones SE, Weedon MN, Shields BM, Oram RA, Hattersley AT (2017) Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol*. 10.1016/S2213-8587(17)30362-5

[24] Allen NE, Sudlow C, Peakman T, Collins R, Biobank UK (2014) UK biobank data: come and get it. *Science translational medicine* 6(224): 224ed224. 10.1126/scitranslmed.3008601

[25] Oram RA, Patel K, Hill A, et al. (2015) A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes care* 39(3): 337-344. 10.2337/dc15-1111

[26] Patel KA, Oram RA, Flanagan SE, et al. (2016) Type 1 Diabetes Genetic Risk Score: A Novel Tool to Discriminate Monogenic and Type 1 Diabetes. *Diabetes* 65(7): 2094-2099. 10.2337/db15-1690

[27] Tyrrell J, Jones SE, Beaumont R, et al. (2016) Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BMJ* 352: i582. 10.1136/bmj.i582

[28] Evans BD, Słowiński PM, Hattersley AT, et al. (2020) Estimating population level disease prevalence using genetic risk scores. *medRxiv: 2020.2002.2020.20025528*. 10.1101/2020.02.20.20025528

[29] Mishra R, Chesi A, Cousminer DL, et al. (2017) Relative contribution of type 1 and type 2 diabetes loci to the genetic etiology of adult-onset, non-insulin-requiring autoimmune diabetes. *BMC Med* 15(1): 88. 10.1186/s12916-017-0846-0

[30] Sharp SA, Rich SS, Wood AR, et al. (2019) Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes care* 42(2): 200-207. 10.2337/dc18-1785

[31] Rich SS, Akolkar B, Concannon P, et al. (2009) Overview of the Type 1 Diabetes Genetics Consortium. *Genes Immun* 10 Suppl 1: S1-4. 10.1038/gene.2009.84

[32] Sukcharoen K, Sharp SA, Thomas NJ, et al. (2020) IgA Nephropathy Genetic Risk Score to Estimate the Prevalence of IgA Nephropathy in UK Biobank. *Kidney Int Rep* 5(10): 1643-1650. 10.1016/j.ekir.2020.07.012

[33] 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(11): e009088. 10.1136/bmjopen-2015-009088

[34] Leslie RD, Lernmark A (2018) Genetic risk scores in adult-onset type 1 diabetes. *Lancet Diabetes Endocrinol* 6(3): 168-169. 10.1016/S2213-8587(18)30039-1

[35] Graham J, Hagopian WA, Kockum I, et al. (2002) Genetic effects on age-dependent onset and islet cell autoantibody markers in type 1 diabetes. *Diabetes* 51(5): 1346-1355.

10.2337/diabetes.51.5.1346

[36] Howson JM, Rosinger S, Smyth DJ, Boehm BO, Group A-ES, Todd JA (2011) Genetic analysis of adult-onset autoimmune diabetes. *Diabetes* 60(10): 2645-2653.

10.2337/db11-0364

[37] Perry DJ, Wasserfall CH, Oram RA, et al. (2018) Application of a Genetic Risk Score to Racially Diverse Type 1 Diabetes Populations Demonstrates the Need for Diversity in Risk-Modeling. *Sci Rep* 8(1): 4529. 10.1038/s41598-018-22574-5

[38] Thomas NJ, Walkey HC, Kaur A, et al. (2021) The absence of islet autoantibodies in clinically diagnosed older-adult onset type 1 diabetes suggests an alternative pathology, advocating for routine testing in this age group. medRxiv: 2021.2003.2022.21252507.

10.1101/2021.03.22.21252507

[39] Jones AG, McDonald TJ, Shields BM, Hagopian W, Hattersley AT (2021) Latent Autoimmune Diabetes of Adults (LADA) Is Likely to Represent a Mixed Population of Autoimmune (Type 1) and Nonautoimmune (Type 2) Diabetes. *Diabetes care*.

10.2337/dc20-2834