Predicting type 1 diabetes in children using electronic health records in primary care in the UK: development and validation of a machine-learning algorithm

Rhian Daniel, Hywel Jones, John W Gregory, Ambika Shetty, Nick Francis, Shantini Paranjothy, Julia Townson

Lancet Digit Health 2024; 6: e386–95

See **Comment** page e375

Division of Population Medicine, School of Medicine, Cardiff University, Cardiff, UK (Prof R Daniel PhD, H Jones PGDip, Prof J W Gregory MD); **The Noah's Ark Children's Hospital for Wales, Department of Paediatric Diabetes and Endocrinology, Cardiff and Vale University Health Board, Cardiff, UK** (A Shetty MD); **Primary Care Research Centre, University of Southampton, Southampton, UK** (Prof N Francis PhD); **Public Health Directorate, NHS Grampian, Aberdeen, UK** (Prof S Paranjothy PhD); **Centre for Trials Research, Cardiff University, Cardiff, UK** (J Townson PhD)

Correspondence to: Dr Julia Townson, Centre for Trials Research, Cardiff University, Cardiff, UK **townson@cardiff.ac.uk**

Summary

Background Children presenting to primary care with suspected type 1 diabetes should be referred immediately to secondary care to avoid life-threatening diabetic ketoacidosis. However, early recognition of children with type 1 diabetes is challenging. Children might not present with classic symptoms, or symptoms might be attributed to more common conditions. A quarter of children present with diabetic ketoacidosis, a proportion unchanged over 25 years. Our aim was to investigate whether a machine-learning algorithm could lead to earlier detection of type 1 diabetes in primary care.

Methods We developed the predictive algorithm using Welsh primary care electronic health records (EHRs) linked to the Brecon Dataset, a register of children newly diagnosed with type 1 diabetes. Children were included from their first primary care record within the study period of Jan 1, 2000, to Dec 31, 2016, until either type 1 diabetes diagnosis, they turned 15 years of age, or study end. We developed an ensemble learner (SuperLearner) using 26 potential predictors. Validation of the algorithm was done in English EHRs from the Clinical Practice Research Datalink (primary care) and Hospital Episode Statistics, focusing on the ability of the algorithm to identify children who went on to develop type 1 diabetes and the time by which diagnosis could be anticipated.

Findings The development dataset comprised 34 754 400 primary care contacts, relating to 952 402 children, and the validation dataset comprised 43 089 103 primary care contacts, relating to 1 493 328 children. Of these, 1829 (0·19%) children younger than 15 years in the development dataset, and 1516 (0·10%) in the validation dataset had a reliable date of type 1 diabetes diagnosis. If set to give an alert in 10% of contacts, an estimated 71·6% (95% CI 68·8–74·4) of the children with type 1 diabetes would receive an alert by the algorithm in the 90 days before diagnosis, with diagnosis anticipated, on average, by an estimated 9·34 days (95% CI 7·77–10·9).

Interpretation If implemented into primary care settings, this predictive algorithm could substantially reduce the proportion of patients with new-onset type 1 diabetes presenting in diabetic ketoacidosis. Acceptability of alert thresholds should be explored in primary care.

Funding Diabetes UK.

Copyright © 2024 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC 4.0 license.

Introduction

Type 1 diabetes, caused by autoimmune destruction of insulin-secreting β cells, is diagnosed in approximately 3000 children younger than 15 years in England and Wales each year.[1] The UK National Institute for Health and Care Excellence guidelines state that general practitioners should make an immediate emergency referral to secondary care if they suspect a child has type 1 diabetes, to avoid the risk of life-threatening diabetic ketoacidosis caused by insulin deficiency.[2] However, identifying a child with type 1 diabetes in primary care is challenging, because of the relative rarity of the condition. Children might not present to primary care with the four classical symptoms of undiagnosed type 1 diabetes, which are excessive urination, thirst, weight loss, and tiredness[3] (known as toilet, thirsty, thinner, and tired [the 4Ts])[4] and many of their symptoms can be attributed to more common childhood conditions.

Globally, the rate of children presenting with diabetic ketoacidosis at onset of type 1 diabetes has been increasing,[5] and in the UK, it has been unacceptably high,[6] at 25% for the past 25 years,[7] causing substantial morbidity and mortality.[8] In addition, children who present with diabetic ketoacidosis are at risk of poorer glycaemic control, leading to a greater risk of adverse long-term outcomes,[9] increased costs due to hospital admission,[10] and greater psychological effects on families.[11]

Previous studies, using routinely collected electronic health records, show that children with type 1 diabetes have substantially more encounters with primary care than children without type 1 diabetes in the 1 month to 6 months before diagnosis.[3]

Studies in other disease areas have shown that machine-learning approaches using routinely collected electronic health records can predict specific health conditions, improve diagnostic accuracy, and support

Research in context

Evidence before this study

We searched PubMed, Google Scholar, and Web of Science on Sept 4, 2023, using the terms "type 1 diabetes" AND ("newly diagnosed" OR "new onset") AND ("child" OR "children" OR "pediatric") AND ("machine learning" OR "predict") AND ("diagnosis"), with no restrictions on dates or language. This search identified 13 papers, either concerned with predicting type 1 diabetes diagnosis through a combination of biomarkers and genetic risk, or with predicting quality of metabolic control following diagnosis. We did not find any studies using machine learning and routinely collected electronic health records to predict type 1 diabetes in childhood. We then searched without the terms ("child" OR "children" OR "pediatrics"). This search identified one paper that used machine learning to distinguish between type 1 and type 2 diabetes, within a population that had already been identified with diabetes. To our knowledge, our study is the first to develop and validate a predictive algorithm to identify type 1 diabetes in childhood on the basis of primary care interactions.

Added value of this study

By validating our algorithm in an independent dataset, we provide evidence that a proportion of children younger than 15 years who will go on to develop type 1 diabetes can be identified from their primary care electronic health records, and how this proportion varies with alert threshold. Our results suggest that time to diagnosis would have been reduced for several children had a predictive tool based on this algorithm been in use. This finding illustrates the potential of such a tool to reduce the risk of children presenting in diabetic ketoacidosis, and the associated mortality and morbidity.

Implications of all the available evidence

Globally, the timely diagnosis of type 1 diabetes in childhood is recognised as inadequate, despite extensive publicity campaigns to raise awareness with members of the public and primary care practitioners. This proof-of-concept study shows that it would be possible to identify children earlier using a machine-learning algorithm. To evaluate the efficacy of the algorithm in practice, feasibility and acceptability of the tool needs to be assessed in primary care.

clinical decision making.[12–14] To our knowledge, there have not been any studies using machine learning and routinely collected electronic health records to predict the onset of unrecognised type 1 diabetes.

Methods

Methods for developing the algorithm

Development data source

Two datasets held by the Secured Anonymised Information Linkage (SAIL) databank at Swansea University, Swansea, UK were linked to provide the development dataset.[15,16] These datasets comprised routinely collected electronic Welsh primary health-care records and a secondary care register of children diagnosed with type 1 diabetes in Wales (Brecon Dataset).[17] The study period was Jan 1, 2000, to Dec 31, 2016. Primary care data included information on symptoms, diagnoses, tests, prescriptions, and referrals, provided by approximately 75% of all primary care practices in Wales (appendix p 2).

See Online for appendix

Data linkage was done within SAIL, using the UK National Health Service (NHS) number, name, gender, age, and postcode of the individuals.[15] Once linked, individuals were attributed an Anonymous Linking Field allowing them to be tracked over time and across datasets, while ensuring researchers only had access to non-identifiable data.

Participants

Participants were included if, at any time during the study period, they were younger than 15 years, were registered at a general practitioner practice contributing data to SAIL, and had at least one contact with primary care while also being younger than 15 years and being registered with a practice contributing to SAIL. Primary care contacts occurring after the date of diagnosis were excluded (appendix p 3).

Predictors

Potential predictors (table 1) were selected based on the factors that are known to be associated with developing type 1 diabetes, informed by the literature, clinical experience, and our previous study.[3] Read codes are provided in the appendix (pp 14–42).

Statistical analysis

To respect the time-to-event nature of the outcome, and to allow for the possibility that different predictors could be important over different time horizons, while also making use of machine-learning algorithms for binary classification, the task of predicting the diagnosis of incident type 1 diabetes was split into seven components (appendix p 5). All the included contacts, from all eligible patients, were stacked vertically, and seven binary outcomes attached to each contact, encoding whether there was a type 1 diabetes diagnosis recorded on the same day as the contact, between 1 day and 7 days after the contact, between 8 days and 14 days after the contact, between 15 days and 30 days after the contact, between 31 days and 90 days after the contact, between 91 days and 180 days after the contact, and between

181 days and 365 days after the contact. Each of the seven binary outcomes was then predicted in turn, with those that received a diagnosis of type 1 diabetes in any window excluded from the dataset for subsequent windows.

Each of the seven binary outcomes was predicted using an ensemble machine learner, the SuperLearner[18] in R. 11 algorithms were included in the library (appendix p 5). Three versions of each of the 11 algorithms were included, using all predictors, using the predictors selected by univariate correlation screening, and using the predictors selected by random forest screening, leading to 34 algorithms in total (including the simple mean).

For each of the seven outcomes, cross-validation was implemented with five folds, respecting the clustering of contacts within children; that is, fold membership was by child, not by contact. The area under the receiver operating characteristic curve (AUROC) was used as the performance measure for each algorithm in the ensemble. The SuperLearner then selected, for each of the seven binary outcomes in turn, the optimal (in terms of cross-validated AUROC) convex combination of all 34 algorithms; these convex combinations are the seven fitted SuperLearners. Finally, the best (again in terms of cross-validated AUROC) convex combination of the seven fitted SuperLearners was found and constituted the final fitted algorithm.

To compare our algorithm with simpler approaches, such as logistic regression, we developed an algorithm based on logistic regression. Because of the relatively low numbers of events and the high number of predictors, the logistic-regression models did not include lagged and twice-lagged contact features, but otherwise included the same predictors as the SuperLearner. An additional benchmark against a chance algorithm was included.

Sample size

We used data from our previous study[3] in conjunction with person-time reweighting (to correct for case–control sampling) to explore the feasibility of the type of analysis planned and the relationship between generalisation error (as measured by the mean-squared error of predictions in new data) and sample size (appendix p 9).

Our investigation estimated that a power-law threshold[19] would be reached at a random (as opposed to case-control) sample of around 250 000 children in the development dataset. Our planned study definition led to around four times this sample size, which allowed us to include the additional planned predictors and a richer ensemble of learners.

	Further details on the predictor
Demographics	
Sex	..
Age	Used as continuous and categorical variables for the algorithm (age ≤6 years, 7–12 years, and ≥13 years)
Clinical predictors	
Non-oral antibiotics	Prescriptions
Oral antibiotics	Prescriptions
Antipyretics	Symptoms, diagnosis, or prescription
Atopic or allergy	Symptoms, diagnosis, or prescription
Behavioural concerns	Symptoms or diagnosis
Bloods	Tests or results
Blurred vision	Symptoms, diagnosis, or referral
Breathlessness	Symptoms or diagnosis
Non-specific contact	Symptoms or diagnosis
Constipation	Symptoms, diagnosis, or prescription
Family history	Recorded
Fungal	Symptoms, diagnosis, or prescription
Gastrointestinal	Symptoms, diagnosis, or prescription
Headache	Symptoms or diagnosis
Obesity	Symptoms, diagnosis, or measure
Polyuria	Symptoms or diagnosis
Prednisolone	Prescriptions
Rash	Symptoms, diagnosis, or prescription
Lower respiratory tract infection	Symptoms, diagnosis, or prescription
Upper respiratory tract infection	Symptoms, diagnosis, or prescription
Skin infections	Symptoms, diagnosis, or prescription
Thirst	Symptoms or diagnosis
Tiredness	Symptoms or diagnosis
Urinary (excluding polyuria)	Symptoms, diagnosis, prescription, test, or results
Vomiting or nausea	Symptoms or diagnosis
Weight loss	Symptoms or diagnosis
Contact timings	
Days elapsed between current and first previous contacts, relative to usual contact frequency (difference and log ratio)	Precisely, if D_1 was the number of days that elapsed between the current and most recent previous contact and D_A the average number of days between all past consecutive consultations, not including the most recent two consultations, then both $D_1 - D_A$ and $\log(D_1/D_A)$ were included
Days elapsed between first and second previous contacts, relative to usual contact frequency (difference and log ratio)	Precisely, if D_2 is the number of days that elapsed between the most recent previous contact and the one before that, then both $D_2 - D_A$ and $\log(D_2/D_A)$ were included

We present details of how the predictors were defined, where applicable. Full details on the read codes used for each of the symptoms, diagnoses, and prescriptions can be found in the appendix (pp 14–42).

Table 1: **Predictors used in the SuperLearner algorithms**

Methods for validating the algorithm

Validation data source

The validation dataset comprised English primary care patient records during the same study period (Jan 1, 2000, to Dec 31, 2016), obtained from the UK Clinical Practice Research Datalink (CPRD) GOLD,[20] linked to secondary-care data from the Hospital Episode Statistics Admitted Patient Care database (HES-APC; Independent Scientific Advisory Committee protocol 20_023R2). CPRD GOLD captures data from approximately 6·9% of UK general practices, records symptoms, diagnoses, prescriptions, tests, and referrals, and has been shown to be largely representative of the UK population, in terms of age, sex, and ethnicity.[20] HES-APC holds hospital admission data, from all hospitals in England funded by the NHS.[21]

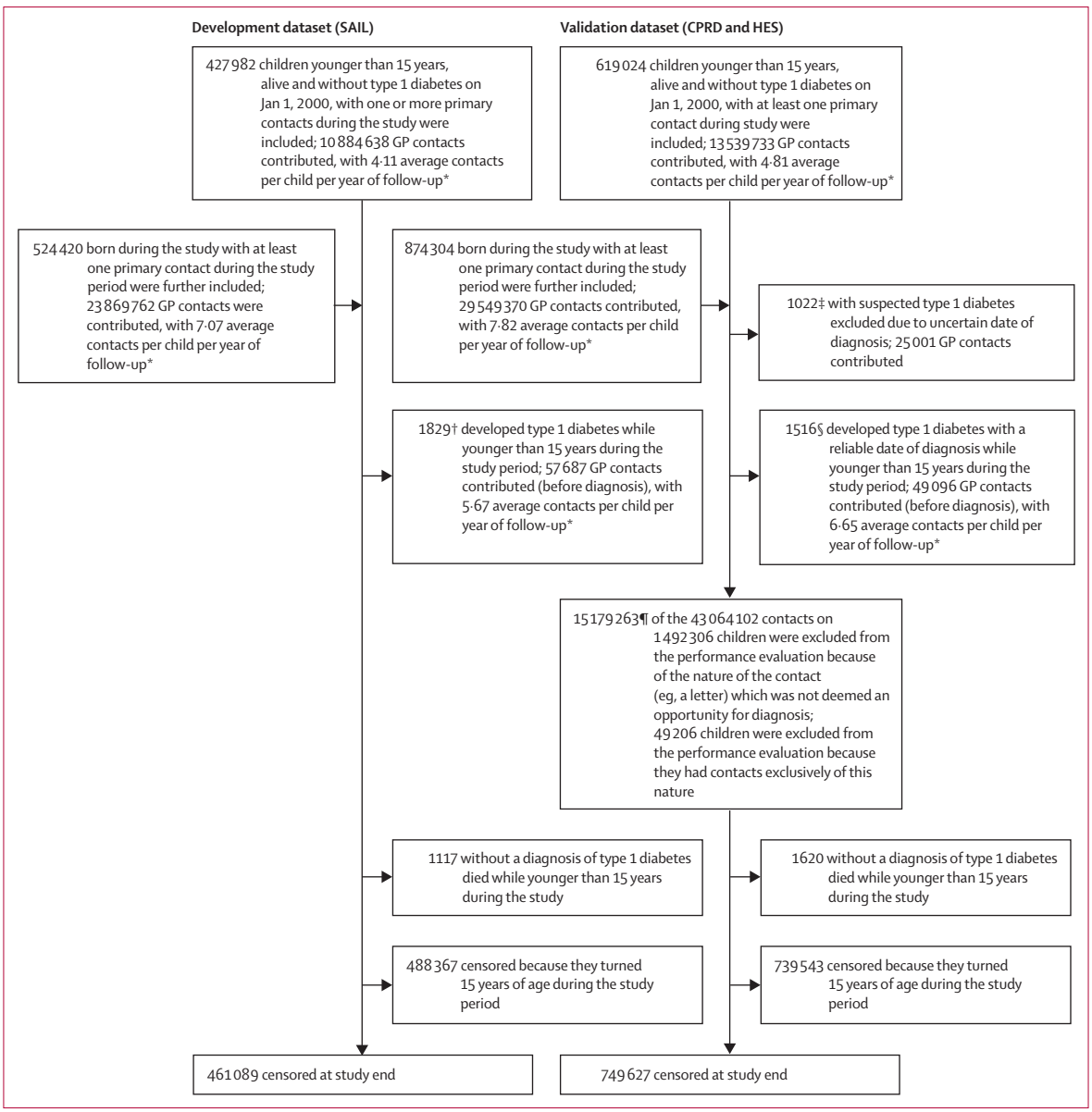

Figure 1: **Flowchart of the progression of children through the constructed development and validation datasets**
CPRD=Clinical Practice Research Datalink. HES=Hospital Episode Statistics. SAIL=Secured Anonymised Information Linkage. *Length of follow-up for each child estimated as time between their first and last included primary care contact. †452 (25%) of the 1829 children had diabetic ketoacidosis at diagnosis, 538 (29%) were aged 6 years or younger, 1004 (55%) were aged 7–12 years, and 287 (16%) were aged 13 years or older at diagnosis. Median age at diagnosis was 9·66 years (IQR 6·30–12·19). 941 (51%) were male and 888 (49%) were female. ‡551 (54%) of 1022 children excluded because they had no hospital code for type 1 diabetes, 303 (30%) excluded because they were missing one or both of the primary care codes (diagnosis, product, or both), and the remaining 168 (16%) had all three codes but with dates that did not agree sufficiently closely. §390 (26%) of 1516 children had diabetic ketoacidosis at diagnosis, 472 (31%) were aged 6 years or younger, 809 (53%) were aged 7–12 years, and 235 (16%) were aged 13 years or older at diagnosis. Median age at diagnosis was 9·37 years (IQR 5·94–11·94). 830 (55%) were male and 686 (45%) were female. ¶15 804 (0·10%) of the 15 179 263 excluded contacts occurred in children who would go on to receive a type 1 diabetes diagnosis and 12 of these children were excluded because they had contacts exclusively of this nature.

Participants

The inclusion and exclusion of participants and their primary care contacts was mirrored in the development dataset. However, the dates for type 1 diabetes diagnosis were inferred, leading to three categories of participants: those diagnosed with type 1 diabetes during the study period without a reliable date of diagnosis; those suspected of having type 1 diabetes with an unknown date of diagnosis; and those without a diagnosis of type 1 diabetes within the study period. Participants suspected of having type 1 diabetes during the study period without a reliable date of diagnosis were excluded from the validation dataset for the primary analyses but were reintroduced in the sensitivity analyses (appendix p 13).

To deduce the date of diagnosis we used the HES-APC data on inpatient admissions, assuming the provisional date of diagnosis to be the date of the start of the earliest hospital spell in which an International Classification of Diseases tenth revision code for type 1 diabetes was recorded (appendix p 4).

We confirmed this provisional date as the reliable date of diagnosis if first, the child also had a type 1 diabetes code in their primary care records (appendix p 42), and this date did not precede the provisional date of diagnosis by more than a day, and second, the child also had a prescription code relating to type 1 diabetes (appendix pp 42–46), not preceding the provisional date of diagnosis. Unless both the first and second conditions were satisfied, the date of diagnosis was classified as unknown (appendix pp 4–5).

Performance evaluation

We evaluated the performance of the algorithm in terms of the effort the alert would cause in primary care when set at a particular threshold, namely, the proportion of all actionable primary care contacts (appendix p 47) with children younger than 15 years in which an alert would be raised.

The benefit of the algorithm was measured in two ways. First, as the proportion of all children with an incident type 1 diabetes diagnosis included in the validation dataset for whom the algorithm, when set to a given threshold, would raise an alert in one or more of their contacts during a set window of time leading up to the observed date of diagnosis. That is, the effort was measured as a proportion of contacts, whereas the benefit was measured as a proportion of children with type 1 diabetes. For our primary analysis, we used a 90-day window leading to diagnosis. In the sensitivity analyses we used 14 days, 45 days, and 180 days (appendix pp 11–12). Children contributed to the denominator of this proportion irrespective of whether they had a primary care contact during the relevant window (appendix pp 6–7).

The second measure of benefit related to the number of days by which diagnosis would be anticipated (within the defined window) with the algorithm set at a particular threshold. We investigated both the mean and 75th quantile of its distribution among children with a type 1 diabetes diagnosis for whom an alert would be raised during the relevant window.

In each of these measures, alerts raised on the date of diagnosis were counted as successes for the algorithm. Our rationale for using this method is outlined together with further discussion of the challenges associated with developing a prediction model for a low-prevalence outcome in the appendix (pp 6–7).

We selected thresholds that would lead to a 5% and 10% effort, and determined a third threshold corresponding to the proportion of actionable contacts in which a general practitioner records any one of the four key diagnostic symptoms of type 1 diabetes (the 4Ts).[4] This third threshold allowed us to benchmark our algorithm against a potential alert system on the basis of only the 4Ts, which somewhat reflects the current advice to general practitioners.

Benchmarking against chance (ie, a system that produces alerts at a given frequency but purely at random) required more care than would typically be the case for diagnostic performance comparisons because benefit is defined per child and effort per contact. That is, a system that produces alerts at random in 10% of all actionable contacts, for example, would still achieve benefit greater than 10% given that each child typically has more than one actionable contact in the lead-up to diagnosis. We thus estimated the performance of such a chance algorithm in the same way as our other algorithms, that is by simulating its effect in the validation data.

We also looked separately at children diagnosed in diabetic ketoacidosis and separately at those in the three age categories, to investigate differential predictive performance for different subgroups.

This study is reported using the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis prediction model guidelines (TRIPOD; see appendix p 48).

Individual participant consent was not required for this study because the data is fully anonymised, and therefore does not fall within the scope of UK Data Protection Regulations.

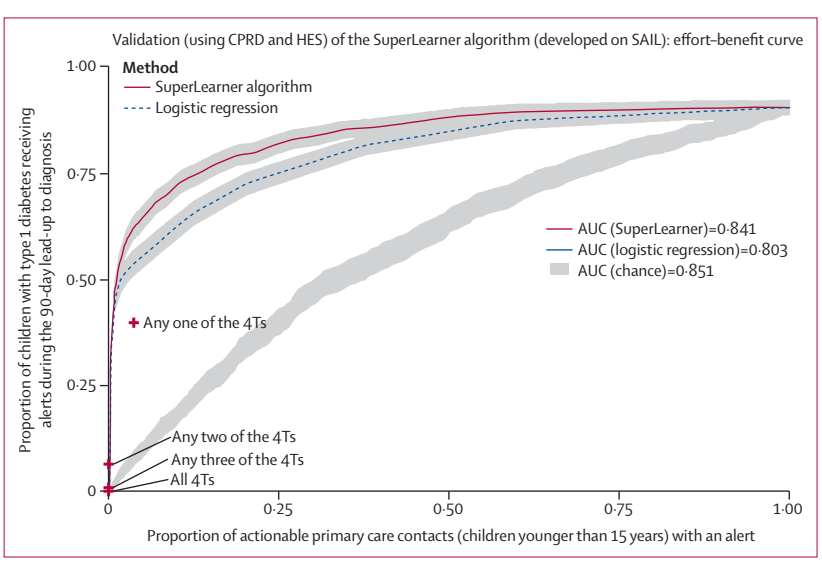

Figure 2: **Estimated discriminatory benefit of the final fitted algorithm based on the SuperLearner against the effort**

The estimated discriminatory benefit against the effort for SuperLearner is compared with logistic regression, chance, and, at the relevant single thresholds, alert systems based on all 4Ts, any three of the 4Ts, any two of the 4Ts, or one of the 4Ts. Discriminatory benefit indicates the proportion of children younger than 15 years who develop type 1 diabetes during the study period (with a reliable date of diagnosis) in the validation dataset (CPRD and HES) who would have received an alert from the algorithm in at least one actionable primary care contact during the 90-day window leading up to diagnosis. Note that the maximum benefit on the right-hand side is 0·905, given that 144 (9·5%) of 1516 children who developed type 1 diabetes in our study did not have a primary care contact less than 90 days before diagnosis. Effort indicates the proportion of all included actionable primary care contacts in which an alert is raised. 4T=toilet, thirsty, thinner, and tired. AUC=area under the curve. CPRD=Clinical Practice Research Datalink. HES=Hospital Episode Statistics. SAIL=Secured Anonymised Information Linkage.

	10% threshold		5% threshold		3·1% threshold	
	Estimated sensitivity	95% CI	Estimated sensitivity	95% CI	Estimated sensitivity	95% CI
Main analysis						
All eligible children with reliable diagnosis date for type 1 diabetes						
SuperLearner method	71·6%	68·8–74·4	64·2%	61·2–67·2	59·8%	56·7–62·9
Logistic-regression method	65·8%	62·8–68·7	58·6%	55·5–61·7	53·4%	50·3–56·5
Chance method	19·6%	17·1–22·1	10·9%	8·92–12·8	6·50%	4·96–8·04
Any of the 4Ts	NA	NA	NA	NA	40·0%	37·6–42·5
Subgroup analyses						
Only children diagnosed with diabetic ketoacidosis						
SuperLearner method	53·4%	47·0–59·8	44·5%	38·2–50·8	43·2%	36·9–49·5
Logistic-regression method	47·9%	41·5–54·3	41·1%	34·8–47·4	38·6%	32·3–44·8
Chance method	20·3%	15·2–25·5	10·2%	6·31–14·0	6·78%	3·57–9·99
Any of the 4Ts	NA	NA	NA	NA	28·6%	24·1–33·2
Children diagnosed when younger than 7 years						
SuperLearner method	62·5%	56·7–68·3	57·0%	51·1–62·9	54·0%	48·1–60·0
Logistic-regression method	48·5%	42·6–54·5	44·9%	38·9–50·8	42·6%	36·8–48·5
Chance method	24·3%	19·2–29·4	13·2%	9·21–17·3	6·99%	3·96–10·0
Any of the 4Ts	NA	NA	NA	NA	39·7%	35·2–44·1
Children diagnosed between age 7 years and 12 years						
SuperLearner method	74·8%	71·1–78·4	65·6%	61·6–69·6	60·6%	56·5–64·7
Logistic-regression method	75·0%	71·3–78·6	64·8%	60·8–68·8	58·6%	54·4–62·7
Chance method	18·2%	15·0–21·5	8·47%	6·13–10·8	4·60%	2·84–6·37
Any of the 4Ts	NA	NA	NA	NA	39·3%	35·9–42·7
Children diagnosed when aged 13 years and older						
SuperLearner method	75·9%	69·5–82·3	71·2%	64·4–78·0	66·5 %	59·4–73·6
Logistic-regression method	64·1%	56·9–71·3	60·6%	53·2–67·9	54·1%	46·6–61·6
Chance method	17·6 %	11·9–23·4	6·47%	2·77–10·2	4·12%	1·13–7·10
Any of the 4Ts	NA	NA	NA	NA	43·2%	36·8–49·5
Sensitivity analysis						
All eligible children with suspected type 1 diabetes						
SuperLearner method	72·4%	70·2–74·6	62·0%	59·7–64·4	52·3%	49·8–54·7
Logistic-regression method	60·9%	58·5–63·3	53·1%	50·6–55·5	47·1%	44·6–49·5
Chance method	29·9%	27·7–32·2	19·3%	17·4–21·3	13·6%	11·9–15·3
Any of the 4Ts	NA	NA	NA	NA	40·0%	37·5–42·4

Subgroup evaluations for children diagnosed with diabetes ketoacidosis, and at different ages, are also included, together with a sensitivity analysis that includes those with suspected type 1 diabetes even though their estimated date of diagnosis was not deemed sufficiently reliable. 4Ts=toilet, thirsty, thinner, and tired. NA=not applicable.

***Table 2*: More details of the results of the comparative discriminatory performance evaluation of the SuperLearner-based algorithm versus logistic regression, chance, and the 4Ts, at three chosen thresholds**

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

There were 34 754 400 primary care contacts involving 952 402 children in the development dataset, and 43 089 103 contacts involving 1 493 328 children in the validation dataset (figure 1). Of these, 1829 (0·19%) children younger than 15 years in the development dataset and 2538 (0·17%) children younger than 15 years in the validation dataset were diagnosed with type 1 diabetes during the study period. However, in the validation dataset, 1022 (40%) of these younger children were excluded because of ambiguity (appendix p 4) concerning the date of diagnosis. These 1022 children, and the 25 001 contacts they would have contributed were reintroduced in the sensitivity analyses (appendix p 13). Therefore, 1516 (0·10%) children with type 1 diabetes were included in the main validation analyses. 452 (25%) of 1829 children in the development dataset and 390 (26%) of 1516 children in the validation dataset had diabetic ketoacidosis at diagnosis.

At each possible threshold for each algorithm, the estimated benefit and 95% CI were calculated (figure 2).

	Mean number of days by which diagnosis would have been anticipated						75th centile of the distribution of number of days			
	10% threshold		5% threshold		3·1% threshold		10% threshold		5% threshold	
	Estimated mean	95% CI	Estimated mean	95% CI	Estimated mean	95% CI	Estimated 75th centile	95% CI	Estimated 75th centile	95% CI
Main analysis										
All eligible children with reliable diagnosis date for type 1 diabetes										
SuperLearner method	9·34	7·77–10·9	6·36	5·00–7·72	4·63	3·46–5·80	4·00	2·25–5·75	2·00	0·481–3·52
Logistic-regression method	11·1	9·4–12·9	7·96	6·38–9·55	5·71	4·29–7·13	6·25	4·32–8·18	2·00	0·198–3·80
Chance method	20·9	16·9–25·0	21·6	16·1–27·1	22·1	15·1–29·1	32	27·3–36·7	40·5	33·1–47·9
Any of the 4Ts	NA	NA	NA	NA	4·42	3·36–5·48	NA	NA	NA	NA
Subgroup analyses										
Only children diagnosed with diabetic ketoacidosis										
SuperLearner method	11·1	7·18–15·1	6·92	3·42–10·4	5·35	2·48–8·23	NA	NA	NA	NA
Logistic-regression method	14·3	9·70–18·9	8·10	4·48–11·7	5·18	2·29–8·07	NA	NA	NA	NA
Chance method	24·9	16·7–33·1	26·6	14·8–38·4	23·4	10·2–36·6	NA	NA	NA	NA
Any of the 4Ts	NA	NA	NA	NA	5·79	2·95–8·64	NA	NA	NA	NA
Children diagnosed when younger than 7 years										
SuperLearner method	4·26	2·24–6·28	2·37	1·00–3·73	1·94	0·70–3·17	NA	NA	NA	NA
Logistic-regression method	1·83	0·51–3·16	1·98	0·55–3·42	2·02	0·51–3·52	NA	NA	NA	NA
Chance method	20·9	14·7–27·1	23·6	15·1–32·1	21·7	9·71–33·8	NA	NA	NA	NA
Any of the 4Ts	NA	NA	NA	NA	3·06	1·52–4·60	NA	NA	NA	NA
Children diagnosed between age 7 years and 12 years										
SuperLearner method	10·8	8·57–13·0	7·43	5·50–9·37	5·69	3·95–7·44	NA	NA	NA	NA
Logistic-regression method	15·1	12·7–17·6	10·4	8·11–12·7	7·19	5·12–9·26	NA	NA	NA	NA
Chance method	18·1	13·0–23·1	16·9	9·15–24·7	18·0	7·32–28·6	NA	NA	NA	NA
Any of the 4Ts	NA	NA	NA	NA	4·98	3·44–6·52	NA	NA	NA	NA
Children diagnosed when aged 13 years and older										
SuperLearner method	11·5	7·35–15·7	8·31	4·57–12·1	5·04	2·16–7·92	NA	NA	NA	NA
Logistic-regression method	7·47	3·88–11·1	6·70	3·22–10·2	5·25	2·06–8·44	NA	NA	NA	NA
Chance method	16·8	6·54–27·0	29·0	8·58–49·4	26·0	1·57–50·4	NA	NA	NA	NA
Any of the 4Ts	NA	NA	NA	NA	5·18	2·21–8·15	NA	NA	NA	NA

Subgroup evaluations for children diagnosed in diabetes ketoacidosis, and at different ages, are also included. 4Ts=toilet, thirsty, thinner, and tired. NA=not applicable.

Table 3: **Results of the comparative performance evaluation of the SuperLearner-based algorithm versus logistic regression, chance, and the 4Ts, at three chosen thresholds, in terms of the estimated distribution of the number of days by which diagnosis might be anticipated**

The effort was estimated with such precision (because of the large number of actionable contacts) that the 95% CI would have width 0 when reported to the same number of decimal places as the estimate, and thus uncertainty in the estimated effort is henceforth ignored (figure 2).

Results for the three chosen levels are shown in more detail in table 2. For example, for a threshold leading to alerts in 10% of all actionable contacts, an estimated 71·6% (95% CI 68·8–74·4) of the children with type 1 diabetes would receive at least one alert by the algorithm in the 90 days before and including their true date of diagnosis (table 3). This proportion decreased to an estimated benefit of 64·2% (61·2–67·2) for a 5% effort, and 59·8% (56·7–62·9) for a 3·1% effort, which is the proportion of all eligible actionable contacts in which at least one of the 4Ts symptoms was coded. A simple system of alerting when at least one of the 4Ts symptoms is coded would, by comparison, only lead to an estimated benefit of 40·0% (37·6–42·5). These results are quite robust to changes in the choice of window length (appendix pp 11–12).

We now turn our attention to the mean and 75th centile of the estimated number of days by which diagnosis would be anticipated among correctly alerted children (figure 3; table 3; appendix p 9).

The three chosen levels are shown in more detail in table 3. For example, for a threshold leading to alerts in 10% of contacts, children correctly alerted by the SuperLearner-based algorithm in the 90 days before diagnosis would, on average, have been alerted an estimated 9·34 days (95% CI 7·77–10·9) before their actual date of diagnosis. The estimated distribution of anticipated days is zero inflated and right skewed, thus results based on quantiles were also considered. The estimated median number of days by which diagnosis was anticipated was 0 days for all three of the considered thresholds, for both the SuperLearner-based algorithm

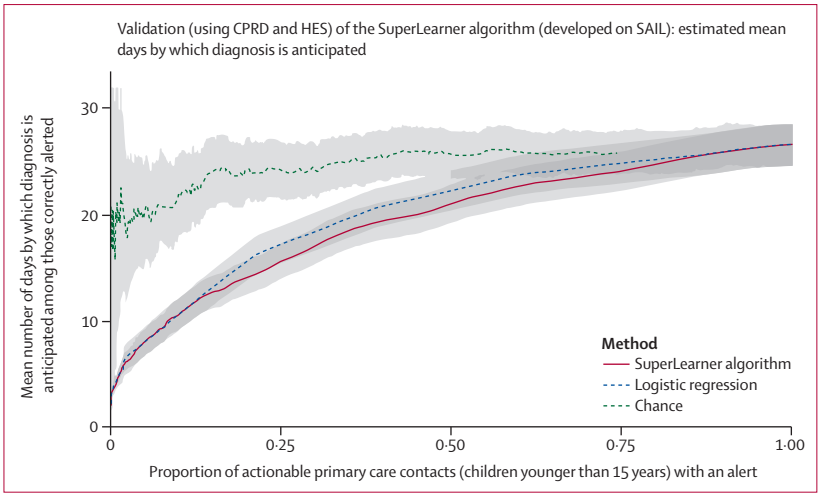

Figure 3: **Estimated mean anticipatory benefit of the final fitted algorithm based on the SuperLearner against the effort**

The mean anticipatory benefit against effort for SuperLearner is compared with logistic regression and chance. Mean anticipatory benefit indicates the mean number of days by which a child's diagnosis would be anticipated using the algorithm, among those who receive a correct alert. Effort indicates the proportion of all included actionable primary care contacts in which an alert is raised. CRPD=Clinical Practice Research Datalink. HES=Hospital Episode Statistics. SAIL=Secured Anonymised Information Linkage.

and logistic regression. However, the estimated 75th centile of the distribution was consistently higher than 0. At the 10% effort threshold, 25% of all correctly alerted children were estimated to be alerted at least 4 days (95% CI 2·25–5·75) before their actual date of diagnosis by the SuperLearner-based algorithm. These results are highly sensitive to the choice of window length (appendix pp 11–12).

The same estimated performance metrics, for children diagnosed with diabetic ketoacidosis, and separately by three age groups at diagnosis, are reported (tables 2, 3). Performance was generally worse for children diagnosed with diabetic ketoacidosis and those diagnosed at younger ages.

Results on the proportions alerted were not very sensitive to the choice of benefit window length, whereas the results on the extent to which algorithms anticipated diagnosis were highly sensitive to this choice (appendix pp 11–12).

The decision to exclude children whose suspected date of type 1 diabetes diagnosis was not deemed sufficiently secure did not strongly influence the results (appendix p 13). Sensitivity analyses relating to variable importance are reported in the appendix (p 13).

Discussion

We developed and validated a predictive algorithm for the identification of children with type 1 diabetes presenting to primary care. Our results suggest that a predictive tool based on this algorithm, with the potential to raise an alert at the time of a consultation, might be a viable intervention to identify children earlier than is currently the case with type 1 diabetes, with the potential to reduce those presenting with diabetic ketoacidosis. When using two thresholds of 10% and 5% to trigger an alert, our predictive model was able to identify 71·6% (95% CI 68·8–74·4) of those who would develop type 1 diabetes within the next 90 days with an estimated reduction in mean time to diagnosis of 9·34 days (95% CI 7·77–10·9) for a 10% threshold, and our model was able to identify 64·2% (61·2–67·2) of those who would develop type 1 diabetes within the next 90 days with an estimated reduction in mean time to diagnosis of 6·36 days (5·00–7·72) for a 5% threshold. In addition, our model identified 53·4% (47·0–59·8) of those who presented with diabetic ketoacidosis at onset (using a 10% threshold). Algorithm performance was lowest in those presenting with diabetic ketoacidosis and young children, in whom progression is most rapid and diagnosis in clinical practice most challenging.

A key consideration for the clinical utility of any predictive tool derived from this algorithm is the acceptability and feasibility in primary care. A systematic meta-review of computerised diagnostic decision-support systems concluded that for them to be effective, they needed to be automated, linked to electronic health records, and to give alerts at a suitable point in the diagnostic decision process.[22] However, risk of burnout of primary care providers due to electronic health-related alert workload has been described,[23] and although recognising the importance of clinical reminders, primary care providers have also reported ignoring them.[24] Therefore, the acceptable threshold at which this predictive tool would trigger an alert needs to be evaluated in a realistic feasibility study in primary care. As with all diagnostic tests, by varying the threshold, the effort and benefit either both increase or both decrease, so that the choice of threshold is a trade-off. For a relatively low-prevalence outcome such as type 1 diabetes, the effort needed to achieve a particular benefit is inevitably higher than for more prevalent outcomes predicted using similar tools (appendix pp 6–7).

Any reduction in the time taken to diagnosis would likely reduce the rate of diabetic ketoacidosis at onset, although the relationship between how much earlier the diagnosis is made and clinical outcomes is unclear. It is likely that a fingerprick blood-glucose test, in individuals in whom an alert is raised, might be an appropriate test, but this approach will require evaluation. A future study would also test our assumption, that the case would have been detected if alerted by the algorithm.

This study has some limitations. An accurate date of diagnosis is crucial for the development and validation of a predictive model based on primary care consultations before diagnosis. In our validation dataset, this date had to be inferred by triangulating information from primary and secondary care, which might have led to some inaccuracies. However, there was little difference in the results when the omitted people with suspected type 1 diabetes were reintroduced in a sensitivity analysis.

Another limitation relates to both the quality of the data collected or extracted, or both (eg, evaluating an alert based on the four classic [4Ts] symptoms depends on the presence of codes for these symptoms, rather than free text). A strength of the SAIL primary care dataset is its ability to track children if they move to a different general practitioner practice within Wales.[25] This strength is not present with CPRD GOLD data, which could have led to duplications of the primary care consultations of children as they move across practices. We were able, however, to link the general practitioner records of any child who had an entry in HES-APC during the study period. Thus, this potential for fragmented general practitioner records would not have led to any double counting of the success of our algorithm. Ethnicity data were not available to us in the validation dataset, because they were not included in the data specification request, meaning that we were unable to check for algorithmic bias by ethnicity. Finally, although we have validated our algorithm in a different dataset, its performance could change over time, if data-entry systems change, coding habits evolve, or the wider health-care environment such as antibody screening changes.[26]

In our primary analyses, we specified a maximum time interval of 90 days before the actual date of diagnosis beyond which an alert raised by our algorithm would be too early to count as a success. Our results on the time by which diagnosis would be anticipated were sensitive to this choice, which is relevant for the interpretation of figure 3. The investigation of the timing of diagnosis addresses a conditional question—given that the algorithm is successful in raising an alert, how many days before diagnosis would an alert be raised, within the 90-day window? That the chance algorithm appears to anticipate diagnosis by a longer period than the other algorithms (and logistic regression anticipates diagnosis by a slightly longer period than the SuperLearner) must be balanced against the lower proportion of correct diagnoses that the chance algorithm would achieve (figure 2). We would also expect that, on the rare occasions that a purely random system alerts correctly within the permitted window, it would tend to do so earlier than an algorithm that takes the relevant available information into account (cautioning against over-interpretation of the results presented in figure 3, table 3, and appendix p 9).

In conclusion, at a range of alert thresholds, our algorithm identified, in an independent dataset, a substantial number of children who went on to develop type 1 diabetes, while also demonstrating the ability of the algorithm to reduce the number of days to diagnosis for some children. This reduction would likely decrease the proportion of children presenting with diabetic ketoacidosis at diagnosis, with a subsequent reduction in deaths and serious complications. The acceptability and feasibility of any tool developed from this algorithm will need to be evaluated in primary care, and a health economic evaluation of the costs and benefits will need to be undertaken.

Contributors

All authors were involved in the conception, design, and funding application for the study. HJ and RD had access to the raw data, which were managed by HJ, and verified by RD and HJ. RD developed the machine-learning algorithms and did the statistical analyses that validated their performance. All authors were involved in the interpretation of the outputs. JT and RD drafted the manuscript, which was critically revised and reviewed by all authors. All authors had full access to all the data and accept responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The routinely collected electronic health-care data used in this study are available through application to the Secured Anonymised Information Linkage databank and UK National Health Service Digital, in accordance with their conditions. R code for the analyses are available from https://github.com/RhianDaniel/TEDstudy. Read codes are provided in the appendix (pp 14–42).

Acknowledgments

This project was funded by Diabetes UK (19/0005998). This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff on behalf of the Cardiff Supercomputing Facility and the High Performance Computing Wales and Supercomputing Wales projects. We acknowledge the support of the High Performance Computing Wales (Supercomputing Wales) project, which is part-funded by the European Regional Development Fund via the Welsh Government. The authors are grateful for the support of members of the Brecon Group in the provision of data, and acknowledge Heather O'Connell and John Harvey for their help maintaining and constructing the Brecon Group Database. We acknowledge Dan Thayer at SAIL databank, Swansea University for making the data available, and advice on how to extract the algorithm from SAIL, and the infrastructure support provided to SAIL by Health and Care Research Wales. We also acknowledge Tolulope Sajobi for his advice and support during the application process and study. This study was approved by the School of Medicine, Cardiff University, Research Ethics Committee (19/107).

References

1 Royal College of Paediatrics and Child Health. National Paediatric Diabetes Audit Report 2020-21; Care processes and outcomes. 2022. https://www.rcpch.ac.uk/resources/npda-annual-reports (accessed Feb 23, 2023).

2 National Institute for Health and Care Excellence. Diabetes (type 1 and type 2) in children and young people: diagnosis and management. NICE guideline (NG18). 2015. https://www.nice.org.uk/guidance/ng18 (accessed Feb 5, 2024).

3 Townson J, Cannings-John R, Francis N, Thayer D, Gregory JW. Presentation to primary care during the prodrome of type 1 diabetes in childhood: a case–control study using record data linkage. *Pediatr Diabetes* 2019; **20:** 12829.

4 Diabetes UK. 4Ts campaign. https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-1/symptoms (accessed Feb 23, 2023).

5 Birkebaek NH, Kamrath C, Grimsmann JM, et al. Impact of the COVID-19 pandemic on long-term trends in the prevalence of diabetic ketoacidosis at diagnosis of paediatric type 1 diabetes: an international multicentre study based on data from 13 national diabetes registries. *Lancet Diabetes Endocrinol* 2022; **10:** 786–94.

6 Narendran P. Screening for type 1 diabetes: are we nearly there yet? *Diabetologia* 2019; **62:** 24–27.

7 Lansdown AJ, Barton J, Warner J, et al. Prevalence of ketoacidosis at diagnosis of childhood onset type 1 diabetes in Wales from 1991 to 2009 and effect of a publicity campaign. *Diabetic Medicine* 2012; **29:** 1506–09.

8 Wolfsdorf JI, Glaser N, Agus M, et al. ISPAD Clinical Practice Consensus Guidelines 2018: diabetic ketoacidosis and the hyperglycemic hyperosmolar state. *Pediatr Diabetes* 2018; **19:** 155–77.

9 Duca LM, Reboussin BA, Pihoker C, et al. Diabetic ketoacidosis at diagnosis of type 1 diabetes and glycemic control over time: the SEARCH for diabetes in youth study. *Pediatr Diabetes* 2019; **20:** 172–79.

10 Dhatariya KK, Glaser NS, Codner E, Umpierrez GE. Diabetic ketoacidosis. *Nat Rev Dis Primers* 2020; **6:** 40.

11 Whittemore R, Jaser S, Chao A, Jang M, Grey M. Psychological experience of parents of children with type 1 diabetes: a systematic mixed-studies review. *Diabetes Educ* 2012; **38:** 562–79.

12 Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. *N Engl J Med* 2016; **375:** 1216–19.

13 Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *J Am Med Informatic Assoc* 2014; **21:** 315–25.

14 Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open* 2013; **3:** e002457.

15 Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009; **9:** 3.

16 Ford DV, Jones KH, Verplancke J-P, et al. The SAIL databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009; **9:** 157.

17 Harvey JN, Hibbs R, Maguire MJ, O'Connell H, Gregory JW. The changing incidence of childhood-onset type 1 diabetes in Wales: effect of gender and season at diagnosis and birth. *Diabetes Res Clin Pract* 2021; **175:** 108739.

18 van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol* 2007; published online Sept 16. https://doi.org/10.2202/1544-6115.1309.

19 Hestness J, Narang S, Ardalani N, et al. Deep learning scaling is predictable, empirically. *arXiv* 2017; published online Dec 1. https://doi.org/1712.00409 (preprint).

20 Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44:** 827–36.

21 Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46:** 1093.

22 Nurek M, Kostopoulou O, Delaney BC, Esmail A. Reducing diagnostic errors in primary care. A systematic meta-review of computerized diagnostic decision support systems by the LINNEAUS collaboration on patient safety in primary care. *Eur J Gen Pract* 2015; **21:** 8–13.

23 Gregory ME, Russo E, Singh H. Electronic health record alert-related workload as a predictor of burnout in primary care providers. *Appl Clin Inform* 2017; **8:** 686–97.

24 Cecil E, Dewa LH, Ma R, Majeed A, Aylin P. General practitioner and nurse practitioner attitudes towards electronic reminders in primary care: a qualitative analysis. *BMJ Open* 2021; **11:** e045050.

25 Akbari A, Lyons R, Bandyopadhyay A, et al. Analysis of factors associated with changing general practice in the first 14 years of life in Wales using linked cohort and primary care records: implications for using primary care databanks for life course research. *Int J Popul Data Sci* 2018; **3:** 818.

26 Besser REJ, Ng SM, Gregory JW, Dayan CM, Randell T, Barrett T. General population screening for childhood type 1 diabetes: is it time for a UK strategy? *Arch Dis Child* 2022; **107:** 790–95.