Comparison of Bayesian approaches for developing prediction models in rare disease: application to the identification of patients with Maturity-Onset Diabetes of the Young

Pedro Cardoso[1], Timothy J. McDonald[1], Kashyap A. Patel[1], Ewan R. Pearson[2], Andrew T. Hattersley[1], Beverley M. Shields[1*], Trevelyan J. McKinley[1*†]

*Joint senior

[†]Corresponding Author, t.mckinley@exeter.ac.uk

Affiliations:

[1]University of Exeter Medical School. Address: Clinical and Biomedical Sciences, RILD Building, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK

[2]University of Dundee. Address: Division of Population Health & Genomics, Ninewells Hospital and Medical School, University of Dundee, Dundee DD1 9SY, UK

Abstract

Background: Clinical prediction models can help identify high-risk patients and facilitate timely interventions. However, developing such models for rare diseases presents challenges due to the scarcity of affected patients for developing and calibrating models. Methods that pool information from multiple sources can help with these challenges.

Methods: We compared three approaches for developing clinical prediction models for population-screening based on an example of discriminating a rare form of diabetes (Maturity-Onset Diabetes of the Young - MODY) in insulin-treated patients from the more common Type 1 diabetes (T1D). Two datasets were used: a case-control dataset (278 T1D, 177 MODY) and a population-representative dataset (1418 patients, 96 MODY tested with biomarker testing, 7 MODY positive). To build a population-level prediction model, we compared three methods for recalibrating models developed in case-control data. These were: prevalence adjustment ("offset"), shrinkage recalibration in the population-level dataset ("recalibration"), and a refitting of the model to the population-level dataset ("re-estimation"). We then developed a Bayesian hierarchical mixture model combining shrinkage recalibration with additional informative biomarker information only available in the population-representative dataset. We developed prior information from the literature and other data sources to deal with missing biomarker and outcome information, and to ensure clinical validity of predictions for certain biomarker combinations.

Results: The offset, re-estimation, and recalibration methods showed good calibration in the population-representative dataset. Among these, the offset and recalibration methods displayed the lowest predictive uncertainty, due to borrowing information from the fitted case-control model. We demonstrate the potential of a mixture model for incorporating informative biomarkers, which significantly enhanced the model's predictive accuracy, reduced uncertainty, and showed higher stability in all ranges of predictive outcome probabilities.

Conclusion: We have compared several approaches that could be used to develop prediction models for rare diseases. Our findings highlight the recalibration mixture model as the optimal strategy if a population-level dataset is available. This approach offers the flexibility to incorporate additional

predictors and informed prior probabilities, contributing to enhanced prediction accuracy for rare diseases, and allows predictions without these additional tests, providing additional information on whether a patient should undergo further biomarker testing before genetic testing.

Keywords: MODY, Bayesian modelling, rare diseases, prior elicitation, recalibration

1. Background

Clinical prediction models can be useful in rare diseases to aid earlier diagnosis and more appropriate management, but developing these models can be challenging as suitable data sources for model development may be difficult to acquire. The prevalence of a rare disease in a population-of-interest can be informed by population cohorts [1], but low numbers of cases in these datasets limit the ability to identify patterns and robustly quantify observed relationships that may be predictive of the disease [2]. Case-control studies [3] can be used to identify disease risk factors, and the rise of rare disease registries [4] is making it easier to recruit larger case numbers when modelling rare diseases. The latter enrich the study population with more disease cases than a random sample from the population, facilitating more robust estimates of associations between patient features and the disease using measures such as odds ratios. Clinically, however, disease risk probabilities are more natural metrics than odds ratios for diagnosis or screening purposes, but estimated risk probabilities from case-control data will be overestimated as they are not based on random samples from the general population [5, 6]. A key challenge is therefore how to produce well-calibrated estimates of individual disease risk probabilities for rare diseases.

Recalibration methods have been developed that can help better estimate individual risk probabilities for rare diseases by borrowing information from one population and recalibrating their outputs to be valid in another population [7, 8, 9, 10, 11, 12]. These approaches could include adjustments of the likelihood ratio based on the sensitivity and specificity of the test at various thresholds [13], offset updating to adjust the overall model probabilities (e.g. for updating to a more appropriate population prevalence) [11, 10], as well as more complex techniques using shrinkage methods to provide adjustments to both the intercept and model coefficients [12, 7] or using previous studies to inform the prior belief of model parameters in Bayesian modelling [14]. These methods combine multiple sources of information and/or clinical expertise to provide more robust disease risk predictions [15].

Example application

Our motivating disease system in this manuscript is a rare young-onset genetic form of diabetes called Maturity-Onset Diabetes of the Young (MODY) [16], which is estimated to account for 1–2% of all diabetes cases [17, 18]. MODY is challenging to identify and is estimated to be misdiagnosed in up to 77% of cases [19]. Diagnostic genetic testing is expensive; however, it is crucial to properly diagnose as these patients do not require treatment with insulin injections [20], unlike the most common young-onset form of diabetes, Type 1 diabetes (T1D).

Decisions regarding which patients to refer for diagnostic MODY testing can be aided with statistical models that use patient characteristics to predict the probability of having MODY. One such set of models is routinely used in clinical practice via an online calculator [13] (found at: https://www.diabetesgenes.org/exeter-diabetes-app/ModyCalculator), and has been shown to improve positive test rates of new MODY cases [19]. These models produce initial probability predictions for new patients using models developed on case-control data, and then recalibrate these probabilities to apply to the general population using conversion tables derived from the sensitivities and specificities at different probability thresholds [13]. There are several consequences of this approach for prevalence adjustment: i) the recalibrated probabilities end up being grouped; ii) individuals cannot have a recalibrated probability that is lower than the estimated prevalence in the general population; and iii) the recalibrated probabilities can be sensitive to the choice of grouping used.

As such, in this manuscript, we explore a range of other approaches to try to deal with some of these challenges, including: an *offset* approach that requires an estimate of the prevalence of MODY in individuals diagnosed with T1D in the population-of-interest [10], and a shrinkage *recalibration* approach that fits a model to a dataset consisting of a random sample from the general population; however, this model uses linear predictors obtained from the original case-control model as covariates. This allows the model in the general population to borrow information from the case-control model, which can help to stabilise predictions for rare outcomes [7]. We compare these to a *re-estimation* approach that simply refits the model developed on the case-control data directly to the population-representative dataset. All of these approaches require additional information from other sources to that used in the development of the original model: either an estimate of the prevalence in the general population or a secondary data set, which in our case consists of a random sample of individuals from the population-of-interest (in our case, from the UNITED study [21]).

In addition, we explore how the results of additional biomarker screening tests could be built into the models to further improve predictions. Two clinical biomarker tests are available that can be useful for discriminating MODY from T1D: C-peptide (a marker of a person's endogenous insulin secretion [22]) and islet autoantibodies (markers of the autoimmune process in T1D [23]), which have not been included in any of the previously developed statistical models. Negative C-peptide results mean that a person is insulin deficient and will need daily insulin injections, characteristic of T1D and not seen in non-insulin-dependent forms of diabetes such as MODY [24]. Positive islet autoantibodies are seen in <1% of MODY patients as MODY is not an autoimmune form of diabetes [25], and this is consistent with the fact that the threshold for positivity for autoantibody tests is set at the 99th percentile of the non-diabetic population [23, 26]. Whilst these two biomarkers are not 100% predictive of MODY, as other diagnoses are possible, including antibody-negative T1D or young-onset Type 2 diabetes (T2D), they can be used to rule out individuals in the population who are highly unlikely to have it as their test results are consistent with T1D.

Since genetic testing is expensive, these highly clinically relevant relationships manifested directly in the study design for UNITED, whereby individuals who were either C-peptide negative or antibody positive were not genetically tested for MODY, under the prior belief that their probability of having MODY was very low. This means we have some missing response information (Fig. 1) but with strong prior information attached to the missing data. Furthermore, since C-peptide and islet autoantibody testing are not routinely conducted in the UK, it is also desirable to build a framework that allows for biologically and clinically valid predictions to be made for individuals whether they do or do not have this additional test information.

To deal with these last challenges, we also develop a Bayesian hierarchical mixture modelling approach which combines the recalibration methods described above with the additional biomarker test information and utilises prior information from other data sources to ensure that the recalibrated probabilities are clinically plausible given specific test results. Using a Bayesian framework can also allow a robust way to deal with missing diagnostic test information and propagate uncertainties through the hierarchical model in a robust manner. We compare these different methods and perform internal validation using a bootstrapping approach, including assessing the calibration and stability of the model predictions in the UNITED dataset.

2. Methods

Setting

The diagnosis of MODY requires expensive genetic testing. Currently, patients are referred for diagnostic genetic testing on an ad-hoc basis when the clinician considers a MODY diagnosis. In line with guidelines (ISPAD [27] and NHS genomic testing criteria [28]), criteria for referring can include:

- Clinical presentation and features of the patient
- Results of biomarker testing (C-peptide and islet autoantibodies)
- The use of prediction models in the form of the MODY calculator (can be found at: https://www.diabetesgenes.org/exeter-diabetes-app/ModyCalculator)

Study population

Case-control dataset (Fig. 1a)

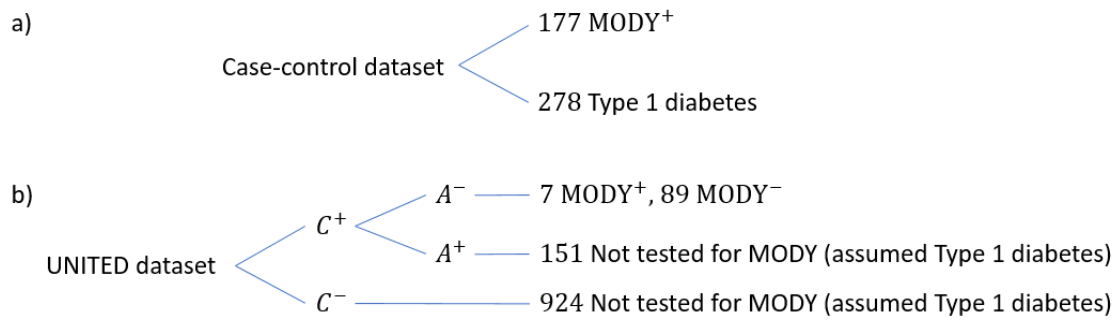
This dataset was used to develop the original MODY prediction model [13]. All participants were diagnosed with diabetes between the ages of 1 and 35. T1D was defined as occurring in patients treated with insulin within 6 months of diagnosis [13]. The dataset includes 278 patients with T1D and 177 probands with a genetic diagnosis of MODY obtained from referrals to the Molecular Genetics Laboratory at the Royal Devon and Exeter NHS Foundation Trust, UK. The dataset is comprised of the following variables: sex, age-at-diagnosis, age-at-recruitment, BMI, parents affected with diabetes and HbA_{1c} (%).

Population-representative dataset (UNITED – Fig. 1b)

The UNITED study [21] was a population-representative cohort that recruited 62% of all patients with diabetes diagnosed between the ages of 1 and 30 in two regions of the UK (Exeter and Tayside) (n=1418). Due to the expense of genetic testing, a screening strategy with two biomarkers was used to narrow down the cohort eligible for MODY testing (Fig. 1b). The two biomarkers were C-peptide (a measure of endogenous insulin secretion) and islet autoantibodies (a marker of the autoimmune process in T1D). As described previously, MODY is characterised by non-insulin dependence, so these patients produce significant amounts of their own endogenous insulin, and they do not have the autoimmune process associated with T1D, whereas being C-peptide negative (i.e. insulin deficient) or having positive islet autoantibodies is characteristic of T1D.

For this model, consistent with the original model [13], we analysed all patients insulin-treated within 6 months of diagnosis, corresponding to 1171 patients, of which 96 were tested for MODY (given that they were C-peptide positive and antibody negative) and 7 MODY cases were diagnosed (Fig. 1b). The dataset is comprised of the following variables: sex, age-at-diagnosis, age-at-recruitment, BMI, parents affected with diabetes and HbA_{1c} (%), with additional C-peptide and islet autoantibodies test results.

Fig. 1: Structure of a) case-control and b) UNITED (population) datasets. *corresponds to a positive test when genetically tested for MODY and corresponds to a negative test when genetically tested for MODY. C^+ = C-peptide positive, C^- = C-peptide negative, A^+ = Antibody positive, A^- = Antibody negative*



Approaches for recalibration

The analysis in this paper was split into three different scenarios to enable population-appropriate probabilities to be calculated with and without the additional biomarker information:

Scenario a) Assume all those not MODY tested are : For this analysis, we used all patients in the population-representative dataset (UNITED). This included 7 MODY positive patients and 1,164 MODY negative (of which 1,075 were not tested for MODY but are assumed to be MODY negative).

Scenario b) Only analyse patients which tested C-peptide positive and autoantibody negative: For this analysis, we focus only on those tested for the presence of a MODY gene. This included 96 patients, of which 7 are MODY positive and 89 MODY negative.

Scenario c) Analyse all patients: For this analysis, we included 96 tested patients (7 MODY positive and 89 MODY negative) and 1,075 non-tested patients. The biomarker information of those not MODY tested was used to adjust probabilities (151 C-peptide positive and autoantibody positive, 924 C-peptide negative).

We explored six approaches for producing predictions using different degrees of data availability, which fall into three groups:

1. Approaches that only utilise case-control data and adjust to a known population prevalence: *Original* and *Offset*;
2. Approaches that utilise a case-control dataset and additional calibration dataset (e.g. UNITED): *Re-estimation* and *Recalibration*;
3. Approaches that utilise additional data on informative diagnostic tests and provide biologically plausible constraints: *mixture model approaches* (for both *Re-estimation* and *Recalibration*). This mixture model splits individuals into two groups according to their diagnostic test information (here a C-peptide negative *or* antibody positive group: ; and a C-peptide positive *and* antibody negative group:). We use an informative prior distribution to constrain the probability of having MODY in the group and uses one of the other recalibration methods in the group.

We note that although the approaches in 1 above only build models for the case-control population, they require additional information about the population-of-interest to do the recalibration, notably the prevalence of the disease. As such, all these methods require additional

data rather than just a case-control dataset, although this additional information is often easier to come by for approaches in 1 than the others. For consistency throughout the manuscript, we use a Bayesian framework to fit all models. The models were fitted using the R package *nimble* (version 1.0.1).

1. Training dataset only approaches

Original approach

This method was implemented by Shields *et al.* (2012) [13] during the development of the original MODY prediction model. The approach fitted a model to a case-control dataset using the patients' characteristics and used the relationship:

$$\frac{P(M^+|R^+)}{P(M^-|R^+)} = \frac{P(R^+|M^+)}{P(R^+|M^-)} \times \frac{P(M^+)}{P(M^-)}$$

or

$$\text{post-recalibration odds} = \frac{\text{sensitivity}}{1 - \text{specificity}} \times \text{pre-test odds},$$

where M^+ is the event that the patient has MODY, and R^+ is whether a hypothetical "test" is positive (and similarly for M^- and R^-). In this case the comparator "test" for MODY is derived by applying a threshold, p^*, to the predicted probabilities from the case-control model, p_i^C, such that an individual is classed as positive if $p_i^C > p^*$ and negative otherwise. Estimates of the sensitivity and specificity of these classifications at a range of thresholds were calculated using the case-control data (since consistent estimates of the sensitivities and specificities can be obtained from case-control data given a specific threshold). The original model probabilities adjusted to a prior prevalence of MODY of 0.7%, which assumed no knowledge of biomarker test results. This approach could be adapted to build in knowledge regarding prior probability from scenario a) or scenario b). In scenario a), we estimated the pre-test probability to be 0.6% (informed by UNITED). For scenario b), we estimated the pre-test probability to be 7.3% (informed by UNITED). Table 1, therefore, maps a predicted p_i from the case-control model to a *recalibrated probability* of having MODY in the general population using these methods.

In the training data (C – case-control dataset), we model M_j^C as:

$$M_j^C | X_j^C \sim \text{Bernoulli}\left(p_j^C\right)$$

where the log odds is given by:

$$\log\left(\frac{p_j^C}{1 - p_j^C}\right) = \beta_0^C + \beta_1^C X_{j1}^C + \cdots + \beta_p^C X_{jp}^C$$

with X_{jv}^C $(v = 1, \ldots, p)$ a set of p covariates for individual j. We put independent vague $\text{Normal}(\mu = 0, \text{sd} = 10)$ priors on the regression parameters.

The posterior for this model then takes the form:

$$\pi(\theta | M^C, X^C) \propto \pi(M^C | X^C, \theta)\pi(\theta)$$

where $\theta = (\beta_v^C; v = 0, \ldots, p)$, with $\pi(\cdot)$ denoting the relevant probability (density) mass functions derived above for the model and joint prior distribution.

Albert Offset approach

This approach was proposed by Albert (1982) [10] and similarly to before, leverages the relationship:

$$\frac{P(M^+|X)}{P(M^-|X)} = \frac{P(X|M^+)}{P(X|M^-)} \times \frac{P(M^+)}{P(M^-)},$$

where X is a set of explanatory variables. In words:

$$\text{post-recalibration odds} = \text{ likelihood ratio} \times \text{pre-test odds}.$$

As in the *Original* method, in the training data (C – case-control dataset) we model M_j^C as:

$$M_i^C | X_j^C, \theta \sim \text{Bernoulli}\left(p_j^C\right)$$

where, as before, the log odds is given directly by:

$$\log\left(\frac{p_j^C}{1-p_j^C}\right) = \beta_0^C + \beta_1^C X_{j1}^C + \cdots + \beta_p^C X_{iv}^C$$

with X_{jv}^C $(v = 1, \ldots, p)$ a set of p covariates for individual j. The idea is that if we know the *pre-test odds*, then we can write:

$$\log\left(\frac{p_j^C}{1-p_j^C}\right) = \eta_0^C + \beta_1^C X_{j1}^C + \cdots + \beta_p^C X_{jp}^C + \log(\text{pre-test odds training}),$$

hence the original $\beta_0^C = \eta_0^C + \log\,(\text{pre-test odds training})$. Therefore, under the assumption that the likelihood ratio for any given set of covariates is the same in the training and calibration dataset, then for a new individual i in the calibration data set (U – UNITED), with covariates X_i^U, we can recalibrate as:

$$\begin{aligned}\log(\text{post-recalibration odds GP}) &= \beta_0^C + \beta_1^C X_{i1}^U + \cdots + \beta_p^C X_{ip}^U \\ &\quad - \log\,(\text{pre-test odds training}) \\ &\quad + \log(\text{pre-test odds general}).\end{aligned}$$

This approach gives individual-level recalibration probabilities that do not rely on thresholding. The *Albert Offset* approach requires a training dataset for fitting the original model and an estimate of the *pre-test odds* in both the training data and the population-of-interest. For this cohort you could adapt the offset based on the prevalence of MODY of 0.6% based on scenario a), or 7.3% based on scenario b). For illustrative purposes, we also explore an example where the likelihood ratio assumption is not maintained between datasets. We put independent vague $\text{Normal}(\mu = 0, \text{sd} = 10)$ priors on the regression parameters. Since we are only fitting the case-control model here, then the posterior takes the same form as for the *Original* method.

2. Population-representative dataset approaches

Re-estimation approach

This approach simply fits a new model directly to the population-representative dataset (UNITED), ignoring the case-control dataset entirely. Given sufficient numbers of cases and controls in a given dataset, this model fitted using, e.g. maximum likelihood, will give asymptotically unbiased estimates for the odds ratios and probabilities. However, for rare diseases, one would have to have very large sample sizes to get sufficient numbers of cases to develop a completely new model. As a comparison here, we use the model structure developed in the case-control dataset, but then refit the model to the population-representative dataset (UNITED). Hence, we can model M_i^U as

$$M_i^U | X_i^U, \theta \sim \text{Bernoulli}\left(p_i^U\right)$$

for individual i, where the log odds is:

$$\log\left(\frac{p_i^U}{1-p_i^U}\right) = \beta_0^U + \beta_1^U X_{i1}^U + \cdots + \beta_p^U X_{ip}^U,$$

with X_{iv}^U $(v = 1, \ldots, p)$ a set of p covariates for individual i, and we place independent vague $\mathrm{Normal}(\mu = 0, \mathrm{sd} = 10)$ priors on the regression parameters.

The posterior for this model then takes the form:

$$\pi(\theta|M^U, X^U) \propto \pi(M^U|X^U, \theta)\pi(\theta)$$

where $\theta = (\beta_v^U;\ v = 1, \ldots, p)$, with $\pi(\cdot)$ denoting the relevant probability (density) mass functions derived above for the model and joint prior distribution.

Recalibration approach

In the context of the models developed here, the *Recalibration* approach [7] uses a model fitted to the case-control dataset to generate predictions of the linear predictor for each individual present in the population-representative data set (UNITED). In the training data, for individual j we model:

$$M_j^C | X_j^C, \theta \sim \mathrm{Bernoulli}\left(p_j^C\right)$$

where

$$\log\left(\frac{p_j^C}{1-p_j^C}\right) = \beta_0^C + \beta_1^C X_{j1}^C + \cdots + \beta_p^C X_{jp}^C,$$

with X_{jv}^C $(v = 1, \ldots, p)$ a set of p covariates for individual j as before. Then, for each individual i in the calibration dataset (UNITED), we calculate the linear predictors $z_i = \widehat{\beta_0^C} + \widehat{\beta_1^C} X_{i1}^U + \cdots + \widehat{\beta_p^C} X_{ip}^U$, where $\widehat{\beta_v^C}$ is a point estimate of the vth regression parameter from the case-control model, and use these z_i terms as covariates in a second shrinkage model:

$$\log\left(\frac{p_i^U}{1-\ p_i^U}\right) = \gamma_0 + \gamma_1 z_i.$$

This approach [7] can have the effect of scaling the odds ratios and intercept terms where necessary, and a side-effect of this approach is that if no recalibration is required, then $\gamma_0 = 0$ and $\gamma_1 = 1$. Again, these approaches could be built using scenarios a) and b), dependent on the assumptions we are willing to make with UNITED. The method used by Steyerberg *et al.* (2004) [7] uses the point predictions for z_i based on the maximum likelihood estimates from the case-control data, which ignores the uncertainty in the estimations of z_i. Instead, we develop a joint Bayesian hierarchical model where we simultaneously fit both models and propagate the uncertainties directly from the case-control model to the recalibration model [29]. We put independent vague $\mathrm{Normal}(\mu = 0, \mathrm{sd} = 10)$ prior distributions on the regression parameters, with a $\mathrm{Normal}(\mu = 0, \mathrm{sd} = 10)$ prior for γ_0 and a $\mathrm{Normal}(\mu = 1, \mathrm{sd} = 10)$ prior for γ_1.

The posterior for this joint model then takes the form:

$$\pi(\theta, M^U, X^U, M^C, X^C) \propto \pi(M^U|X^U, \theta)\pi(M\text{^}C|X^C, \theta^C)\pi(\theta)$$

where $\theta = (\theta^U, \theta^C)$ corresponds to the full vector of parameters, with $\theta^U = (\gamma_0, \gamma_1)$ and $\theta^C = (\beta_v^C; v = 1, \ldots, p)$, with $\pi(\cdot)$ denoting the relevant probability (density) mass functions derived above for the different component models and joint prior distribution.

3. Mixture model approach

One area of development in this manuscript is how to incorporate biomarker test information into the model when the biomarker tests can place very strong constraints on the post-recalibration probabilities depending on their specific values. For example, a simple way to include a binary test result would be to add another covariate into the linear predictor in one of the previous methods. In the analysis here the biomarker data for only exists in the calibration data [UNITED], but not the training data [case-control]), and so this approach would only use information from the calibration data to estimate the parameters relating to the biomarkers. Since there are few cases in the calibration data, this would necessarily result in large standard errors for the estimated effects and could lead to biologically implausible estimates. For example, an individual who is C-peptide negative *or* antibody positive can be considered to have a very low probability of having MODY, justified through prior data and biological plausibility (for example, C-peptide negativity means that an individual is producing negligible amounts of their own insulin, which defines T1D). In clinical practice, an individual with these biomarker results would be treated as having T1D, which is equivalent to assuming that the probability of having MODY given these results is exactly zero. However, this approach does not allow for the rare (but possible) event that an individual has a positive genetic MODY test but is antibody positive or C-peptide negative (which would ideally also allow for imperfect sensitivities and specificities of the biomarker tests).

Using the mixture model approach in scenario c), it is possible to incorporate a non-zero prior probability of having MODY in these cases, where we use independent data sets to inform the prior distribution for this probability. We note that the mixture model allows for different prior constraints to be used for different subsets of the data: here the prior probability of having MODY is very low for C-peptide negative or antibody positive individuals [30, 31], but is not similarly constrained for C-peptide positive and antibody negative individuals. Similar ideas could be used for other diseases where the prior information may not be as strong.

For the UNITED data, we let

$$M_i^U = \begin{cases} 1 & \text{if individual } i \text{ has MODY} \\ 0 & \text{otherwise,} \end{cases}$$

with

$$A_i^U = \begin{cases} 1 & \text{if individual } i \text{ is } A^+ \\ 0 & \text{otherwise,} \end{cases}$$

and

$$C_i^U = \begin{cases} 1 & \text{if individual } i \text{ is } C^+ \\ 0 & \text{otherwise.} \end{cases}$$

We then set:

$$T_i^U = \begin{cases} 1 & \text{if individual } i \text{ is } [C^- \cup A^+] \\ 0 & \text{otherwise.} \end{cases}$$

Letting X_i^U be a vector of additional covariates for individual i, we can model M_i^U as

$$M_i^U | X_i^U, T_i^U \sim \text{Bernoulli}\left(p_{M^+|T_i,X_i}\right)$$

where

$$p_{M^+|T_i,X_i} = \begin{cases} p_{M^+|C^- \cup A^+} & \text{if } T_i^U = 1, \\ p_{M^+|C^+ \cap A^-,X_i} & \text{if } T_i^U = 0. \end{cases}$$

We model $p_{M^+|C^- \cup A^+}$ using a $\text{Beta}(\alpha = 2.2, \beta = 7361.3)$ prior probability distribution. (see the Prior Elicitation section for a justification of this choice). We then model $p_{M^+|C^+ \cap A^-, X_i}$ differently, depending on whether we use the *Re-estimation* or *Recalibration* approaches (see below).

Re-estimation approach

For the *Re-estimation* approach we model

$$\log\left(\frac{P_{M^+|C^+ \cap A^-, X_i}}{1 - P_{M^+|C^+ \cap A^-, X_i}}\right) = \beta_0^U + \beta_1^U X_{i1}^U + \cdots + \beta_p^U X_{ip}^{\mathrm{U}}$$

and to finalise, we put independent vague $\text{Normal}(\mu = 0, \text{sd} = 10)$ priors on the regression parameters.

Recalibration approach

For the *Recalibration* approach we also utilise the case-control data. If we let M_j^C be the MODY status for individual j in the case-control dataset, with vector of covariates X_j^C, then we model

$$M_j^C | X_j^C \sim \text{Bernoulli}(p_j^C)$$

where

$$\log\left(\frac{p_j^c}{1 - p_j^C}\right) = \beta_0^C + \beta_1^C X_{j1}^C + \cdots + \beta_p^C X_{jp}^C.$$

Then, for individual i in the UNITED dataset (with $C_i^+ \cap A_i^-$), we model

$$\log\left(\frac{p_{M^+|C^+ \cap A^-, X_i}}{1 - P_{M^+|C^+ \cap A^-, X_i}}\right) = \gamma_0 + \gamma_1 z_i$$

where

$$z_i = \beta_0^C + \beta_1^C X_{i1}^U + \cdots + \beta_p^C X_{ip}^U.$$

Test results

To allow for predictions in the absence of biomarker test results (which are not routinely collected in clinical practice), we model

$$T_i^U \sim \text{Bernoulli}(q_i)$$

$$\log\left(\frac{q_i}{1 - q_i}\right) = \beta_0^* + \beta_1^* X_{i1}^{*U} + \cdots + \beta_p^* X_{ip}^{*U}$$

with X^{*U} comprised of the variables BMI, age-of-diagnosis, age-of-recruitment and family history (here we use restricted cubic splines with 3 knots to model the continuous variables). In this case the predicted probability of MODY for an individual with unknown test results will be a weighted average of the $p_{M^+|C^- \cup A^+}$ and $p_{M^+|C^+ \cap A^-, X_i}$, weighted by the probability of being $C^- \cup A^+$ based on suitable individual-level characteristics. We place independent vague $\text{Normal}(\mu = 0, \text{sd} = 10)$ prior distributions on the regression parameters, with a $\text{Normal}(\mu = 0, \text{sd} = 10)$ prior on γ_0 and a $\text{Normal}(\mu = 1, \text{sd} = 10)$ prior on γ_1.

For the *Re-estimation mixture*, the posterior then takes the form:

$$\pi(\theta|M^U, X^U, X^{*U}, T^U) \propto \pi\left(M^U \middle| X^U, T^U, \theta^U, p_{M^\wedge+|C^-\cup A^+}\right)\pi(T^U, X^{*U}, \theta^T)\pi(\theta),$$

where $\theta = \left(\theta^U, \theta^T, p_{M^\wedge+|C^-\cup A^+}\right)$ corresponds to the full vector of parameters, with $\theta^U = (\beta_v^U; v = 1, \dots, p)$ and $\theta^T = (\beta_v^*; v = 1, \dots, r)$ with $\pi(\cdot)$ denoting the relevant probability (density) mass functions derived above for the different component models and joint prior distribution.

For the *Recalibration mixture*, the posterior then takes the form:

$$\pi(\theta|M^U, X^U, T^U, M^C, X^C) \propto \pi\left(M^U \middle| T^U, X^U, \theta^U, p_{M^\wedge+|C^-\cup A^+}\right)\pi(T^U|X^{*U}, \theta^T)\pi(M^C|X^C, \theta^C)\pi(\theta)$$

where $\theta = (\theta^U, \theta^T, p_{M^\wedge+|C^-\cup A^+}, \theta^C)$ corresponds to the full vector of parameters, with $\theta^U = (\gamma_0, \gamma_1)$, $\theta^C = (\theta_v^C; v = 1, \dots, p)$ and $\theta^T = (\beta_v^*; v = 1, \dots, r)$, with $\pi(\cdot)$ denoting the relevant probability (density) mass functions derived above for the different component models and joint prior distributions.

Assessment/calibration

In scenario a), we validate fitted probabilities for all patients in UNITED (setting those with missing MODY testing to MODY^-). In scenarios b) and c), we only validate fitted probabilities on $C^- \cup A^+$ patients as these were the only patients referred for genetic testing of MODY genes. The area under the receiver operating characteristic (AUROC) curve was used as a measure of overall discrimination performance. Calibration curves were plotted to visualise how well the predicted probabilities are calibrated against the observed data. For the calibration curves, predicted probabilities are grouped by quintiles and plotted against the observed probability of positive individuals within each quintile. To assess convergence, we monitored the available parameters for evidence of convergence (trace plots available on request) and Gelman-Rubin $\hat{R}$ values [32].

Prior elicitation

A major advantage of the *mixture model* approach is formally building in appropriate low prior probabilities for the individuals with C-peptide negative or antibody positive test results rather than just considering all as MODY^-. The prior probability of $M^+|C^- \cup A^+$ can be modelled as:

$$P(M^+|C^- \cup A^+) \propto P(C^- \cup A^+|M^+)P(M^+), \quad (1)$$

from a straightforward application of Bayes' Theorem. We can estimate $P(C^- \cup A^+|M^+)$ using data from the "One tube study", which tested C-peptide and islet autoantibodies in sequential referrals to the UK MODY diagnostic genetic testing service within the Exeter Genomics Laboratory at the Royal Devon and Exeter National Health Service (NHS) Foundation Trust, which provides monogenic diabetes testing for England, Wales, and Northern Ireland. In this cohort, in 282 individuals who tested positive for MODY, only 6 were $C^- \cup A^+$(personal communication T McDonald). Under a flat U(0, 1) prior distribution for $P(C^- \cup A^+|M^+)$, we can fit a binomial model to these data giving a posterior distribution:

$$P(C^- \cup A^+|M^+) \sim \mathrm{Beta}(\alpha = 7, \beta = 277) \quad (2)$$

Furthermore, we can estimate $P(M^+)$ (the prevalence of MODY in young-onset insulin-treated patients) from an independent UK population screening study [30], which found 2 individuals with MODY out of 237 individuals diagnosed with T1D. Again, under a flat U(0, 1) prior distribution for $P(M^+)$, we can fit a binomial model to these data giving a posterior distribution:

$$P(M^+) \sim \mathrm{Beta}(\alpha = 3, \beta = 246) \quad (3)$$

These two distributions in (2) and (3) can then be used to generate an estimate for (1) using Monte Carlo simulation. Here we take independent random samples from (2) and (3), and thus derive random samples for $P(M^+|C^- \cup A^+)$ by taking their product. We then fit a Beta distribution to these random samples using maximum likelihood, and we can use this fitted beta distribution as a tractable prior distribution for (1) in our MODY model, giving:

$$P(M^+|C^- \cup A^+) \sim \mathrm{Beta}(\alpha = 2.2, \beta = 7361.3) \quad (4)$$

We note that $E[P(M^+|C^- \cup A^+)]$ is very small here and is in line with the choice not to genetically test for MODY for $C^- \cup A^+$ individuals during the UNITED study, however using this Beta prior is more conservative than assuming a prior probability of exactly zero. Furthermore, the distributions in (2) and (3) were generated using flat uniform priors and so (4) should be conservative in the sense of having the largest variance compared to other choices of prior distributions. A plot of the histogram of the random samples versus the fitted Beta distribution is shown in sFig. 1.

Stability analysis

We compared model stability for the *Re-estimation* and *Recalibration* mixture approaches to investigate the impact of low-case numbers. The study investigated the instability of the model by fitting the same model to 1,000 bootstrapped datasets [33]. The instability is shown by deriving a prediction instability plot of bootstrap model predictions (y-axis) versus original model predictions (x-axis). Stability is demonstrated when most values land close to the diagonal line (y=x).

3. Results

Comparing datasets

In the case control dataset 177 out of 455 patients had MODY, leading to an enriched proportion with MODY of 40%. In contrast, in the recalibration population (UNITED) cohort, 7 out of 1171 patients (0.6%) had MODY, much more consistent with the prevalence of MODY in the population. The characteristics of patients in the two populations were broadly similar (sFig. 2).

Models from the 6 different approaches

For all approaches in this study converged quickly, so we ran two chains of 20,000 iterations with the first 10,000 discarded for burn-in for all models. The *Original* approach achieved an $\hat{R} < 1.02$ for all parameters and the choice of prevalence affected the conversion probabilities. Table 1 shows the different *post-recalibration probabilities* of having MODY using the different prevalences for both scenarios a) and b).

Table 1: Probability conversion table for MODY using the Original method with an adjusted pre-test probability. *(A) based on the original approach, (B) UNITED using those genetically tested for MODY. An individual with a predicted $p_i = 0.85$ from approach (A) model (85%), would get mapped to a post-recalibration probability of 12.6%.*

(A) Original model adjusting probabilities for prevalence (pre-test probability = 0.6%)		(B) Original model adjusting probabilities based on UNITED genetically tested patients [scenario b)] (pre-test probability=7.3%)	
Pre-test probability (%)	Post-recalibration probability (%)	Pre-test probability (%)	Post-recalibration probability (%)
0	0.6	0	7.3
10	1.6	10	17.9
20	2.3	20	23.2
30	3.4	30	31.6
40	4.2	40	36.6
50	5.5	50	43.1
60	6.2	60	46.5
70	7.1	70	49.9
80	10.9	80	61.6
90	45.4	90	91.6

Table 2 describes the model parameter estimates for the *Albert Offset* and *Re-estimation* approaches in scenarios a) and b). The *Albert Offset* approach achieved an $\hat{R} < 1.02$ and the *Re-estimation* approach achieved an $\hat{R} < 1.01$ for all parameters.

Table 2: Model parameter estimates for Albert Offset and Re-estimation approaches for scenarios a) and b). *Scenario a): assume all those not MODY tested are* MODY^- *(based on strong clinical knowledge (n=1,171). Scenario b): only analyse patients which tested C-peptide positive and autoantibody negative (n=96). The numbers in the brackets correspond to the 95% credible intervals.*

	Shields *et al.* [13]	Scenario a)		Scenario b)	
Model parameters	Original Model	Albert Offset	Re-estimation	Albert Offset	Re-estimation
Intercept	1.91 (-0.11;4.57)	-2.75 (-4.76;-0.08)	-11.63 (-4.38;-23.47)	-0.19 (-2.21;2.47)	-10.51 (-23.80;-1.36)
At least one parent affected with diabetes	3.22 (2.57;3.96)	3.22 (2.57;3.96)	3.57 (1.53;6.28)	3.22 (2.57;3.96)	4.18 (1.52;7.34)
Age at recruitment (years)	-0.09 (-0.11;-0.06)	-0.09 (-0.11;-0.06)	-0.03 (-0.12;0.06)	-0.09 (-0.11;-0.06)	0.04 (-0.07;0.14)
Hba_{1c} (%)	-0.68 (-0.93;-0.50)	-0.68 (-0.93;-0.50)	-0.44 (-0.97;0.06)	-0.68 (-0.93;-0.50)	-0.79 (-1.58;-0.19)
Age at diagnosis (years)	1.37 (0.66;2.09)	1.37 (0.66;2.09)	0.07 (-0.04;0.19)	1.37 (0.66;2.09)	-0.09 (-0.30;-0.10)
Sex (baseline Male)	1.91 (-0.11;4.57)	1.91 (-0.11;4.57)	4.19 (1.33;8.38)	1.91 (-0.11;4.57)	6.56 (2.05;13.87)

The *Recalibration* approach achieved an $\hat{R} < 1.03$ for all parameters. In scenario a), estimates were $\gamma_0 = -4.4$ (95%CI -5.4;-3.6) and $\gamma_1 = 0.95$ (95%CI 0.5;1.5) and in scenario b), the estimates were $\gamma_0 = -2.2$ (95%CI -3.3;-1.4) and $\gamma_1 = 0.8$ (95%CI 0.3;1.6). In scenario c), the *Recalibration mixture* approach achieved an $\hat{R} < 1.06$ for all parameters, with an estimated $\gamma_0 = -2.3$ (95%CI -3.3;-1.4) and $\gamma_1 = 0.9$ (95%CI 0.3;1.6). The *Re-estimation mixture* approach achieved an $\hat{R} < 1.16$. The model that estimates T achieved an AUROC of 0.76 (95%CI 0.74;0.77) in both mixture approaches, with model parameters described in sTable 1.

Discrimination and calibration of the models developed using the 6 different approaches

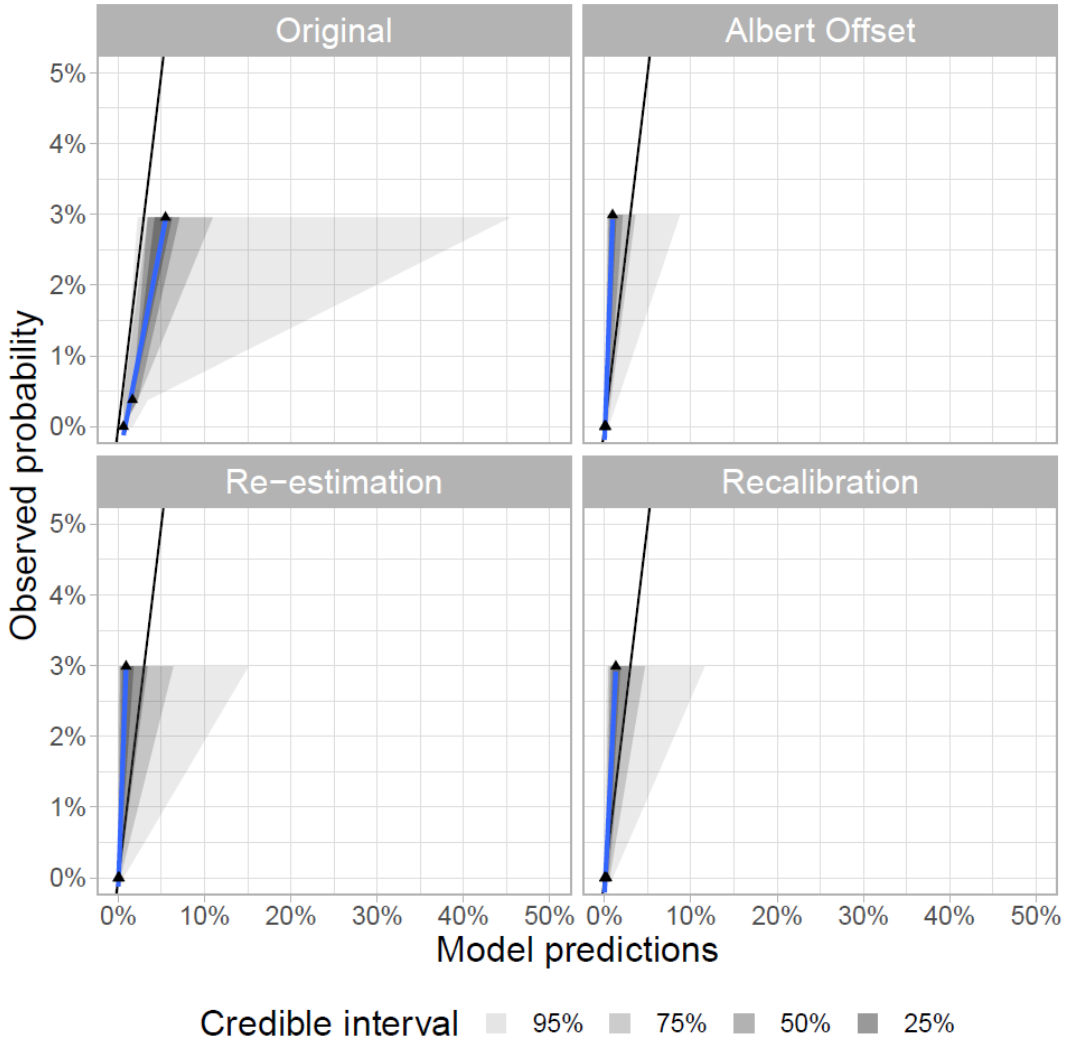
All approaches led to good model discrimination, with (as may be expected) the *Re-estimation* approaches having the highest AUROC (Table 3).

Table 3: Area under the receiver operating characteristics (AUROC) for all approaches in both scenarios. *In scenario a): assume all not MODY tested are* MODY^{-} *(based on strong clinical knowledge (n=1,171). Scenario b): only analyse patients which tested C-peptide positive and autoantibody negative (n=96). In scenario c): analyse all patients (n=1,171, validated in n=96). CI: credible interval*

	Approach	Mean	CI: 2.5%	CI: 97.5%
Scenario a)	Original	0.93	0.90	0.94
	Albert Offset	0.93	0.90	0.94
	Re-estimation	0.94	0.90	0.96
	Recalibration	0.93	0.92	0.94
Scenario b)	Original	0.86	0.81	0.89
	Albert Offset	0.86	0.82	0.89
	Re-estimation	0.92	0.89	0.94
	Recalibration	0.87	0.84	0.90
Scenario c)	Re-estimation Mixture	0.92	0.87	0.94
	Recalibration Mixture	0.87	0.84	0.90

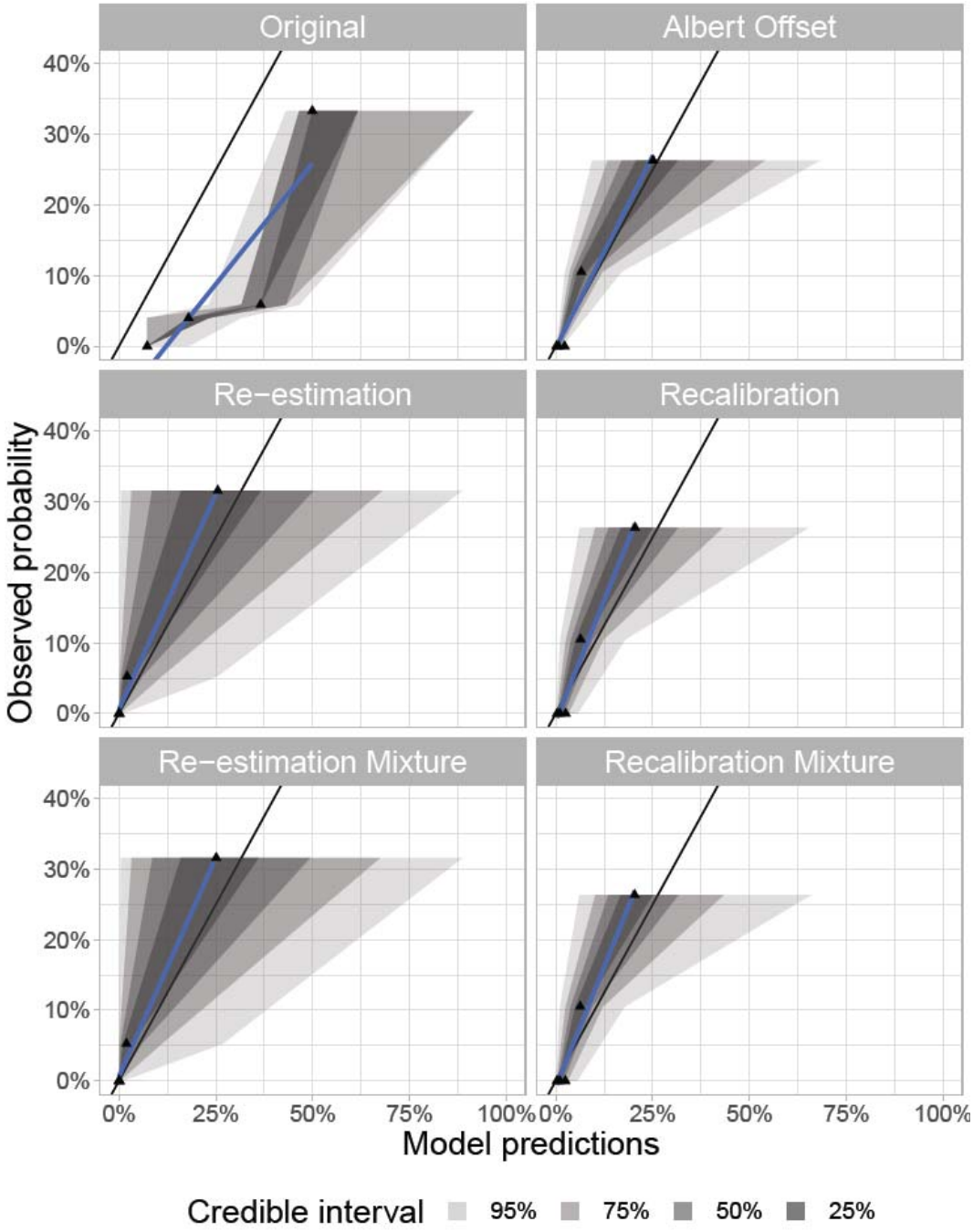
In scenario a), for the approaches that used only the case-control dataset and the patient characteristics, the *Original* approach overestimated the observed probability of MODY in the UNITED population and had large uncertainty at higher percentages. In contrast, the *Albert Offset* approach slightly underestimated the observed probability of MODY in the UNITED population. Looking at the approaches that used the population-representative dataset (UNITED), both the *Re-estimation* and *Recalibration* approaches slightly underestimated the observed probability of MODY with slightly more uncertainty in the predictions from the *Re-estimation* approach (Fig. 2).

Fig. 2: Calibration of scenario a) in UNITED. *Scenario a): assume all not MODY tested are* MODY^{-} *(based on strong clinical knowledge (n=1,171).*



In scenario b), for the approaches that used the case-control dataset and the patient characteristics, the *Original* approach overestimated the observed probability of MODY in the UNITED population, whereas the *Albert Offset* approach was able to calibrate well. Looking at the approaches that used an additional calibration dataset (UNITED), both the *Re-estimation* and *Recalibration* approaches calibrated well, but the *Re-estimation* approach demonstrated more uncertainty in the probability predictions (Fig. 3). In scenario c), the *Re-estimation* and *Recalibration mixture* approaches demonstrated similar performance to the equivalent models that did not use a *mixture model* approach, with similar levels of uncertainty in probability predictions (Fig. 3). In this case, the *Albert Offset* method worked well, but it relies on the assumption that the likelihood ratio is the same in the two populations. For illustrative purposes, we also provide an example setting where the likelihood ratio is different between the training and calibration datasets (violating the assumption). In this latter example, the *Albert Offset* approach fails to calibrate well. In contrast, the *Recalibration* approach can scale the odds ratios and calibrates well (sFig. 3), and so if a recalibration dataset is available, we would prefer this method in general.

Fig. 3: Calibration of scenario b) and c) in UNITED. *Scenario b): only analyse patients which tested C-peptide positive and autoantibody negative (n=96) – Original, Albert Offset, Re-estimation and Recalibration approaches. In scenario c): analyse all patients (n=1,171, validated in n=96) – Re-estimation mixture and Recalibration mixture approaches.*

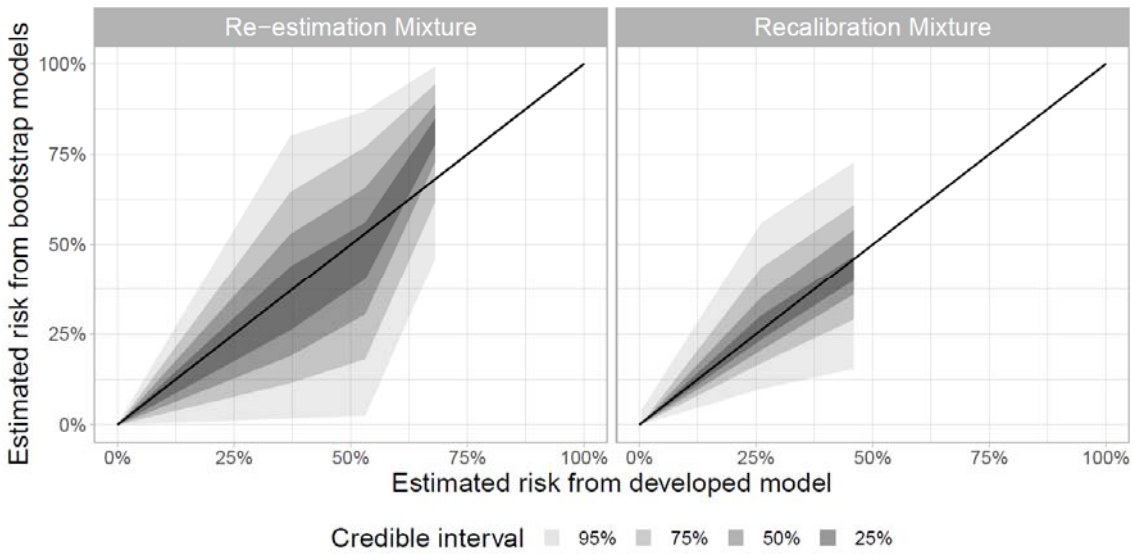


Stability plots

The bootstrap stability test was made for both mixture approaches. Both mixture approaches were ran 20,000 iterations with the first 10,000 discarded for burn-in, with an average $\hat{R} = 1.02$ (95% 1.0;2.1) for the *Re-estimation mixture* approach and an average $\hat{R} = 1.01$ (95% 1.0;1.6) for the *Recalibration mixture* approach (higher $\hat{R}$ values occurred in bootstrapped datasets with less than 2 positive MODY cases). Both recalibration approaches showed some variability in the estimated

probabilities, with the *Re-estimation mixture* approach demonstrating higher uncertainty across all estimated probability levels (Fig. 4). However, we can see that because the *Recalibration mixture* approach borrows weight from the case-control data, the estimates were more stable than the *Re-estimation mixture* method. We also noted that by using the hierarchical modelling approach, the *Recalibration mixture* model uncertainty included the uncertainty in the case-control predictions. Thus, these uncertainty estimates are larger than a model where this additional predictive uncertainty is ignored.

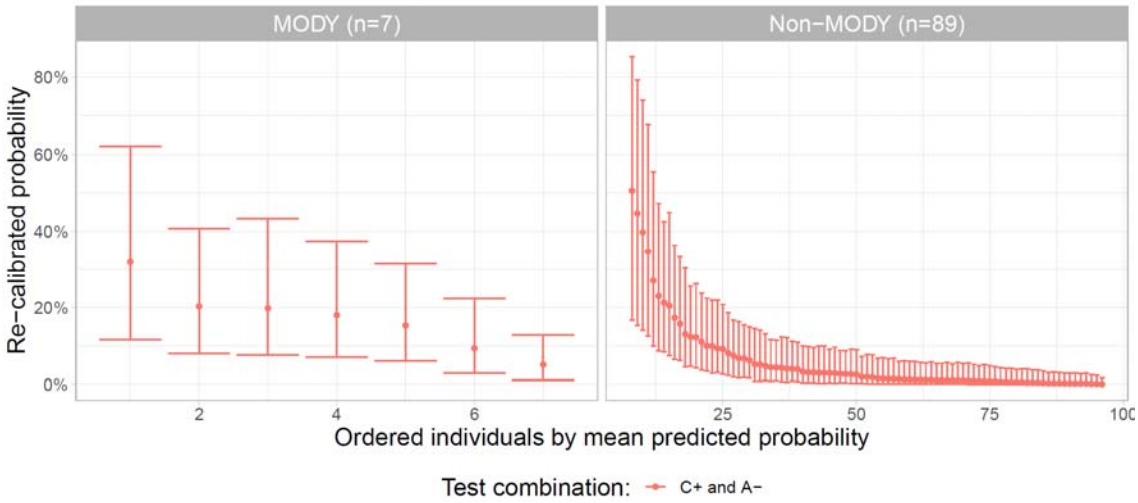
Fig. 4: Stability plots for Re-estimation and Recalibration mixture approaches. *Estimations of MODY probability from bootstrapped models are plotted against estimated MODY probabilities from the developed model.*



Final recalibrated probabilities

In terms of average probability of MODY in the general population, the *Original* approach estimates a MODY prevalence of 16%, whereas the *Recalibration mixture* approach estimates a MODY prevalence of 0.6%, which was similar to the actual prevalence of 0.6% in this dataset, suggesting it provides better calibration overall. The approach chosen for our final models was the *Recalibration mixture* approach, which incorporated the most information with the lowest uncertainty in probability predictions. The mixture model ensures that those with biomarkers consistent with T1D (the $C^- \cup A^+$ individuals) are predicted to have a very low probability of MODY, consistent with independent prior information. Fig. 5 shows the predicted probability of MODY in the remaining $C^+ \cap A^-$ individuals. Considering only 0.6% of the cohort had MODY, the model produced a wide range of probabilities. Most non-MODY cases were predicted to have a low probability of MODY, with 97.2% (1,132/1,164) of individuals having an upper 95% CI probability of MODY under 10%. In contrast, 7 out of the 7 MODY cases had an upper 95% CI probability >10%. This would mean that if using a >10% threshold to initiate MODY testing for the population, 39 patients would be tested, giving a positive predictive value of 17.9% (Fig. 5), which is equivalent to the *Original* approach.

Fig. 5: Estimated probabilities of MODY from the Recalibration mixture model. *A threshold of 10% to initiate MODY testing in the population would result in 39 patients being tested, with a positive predictive value of 17.9%. Mean probability plotted with 95% credible intervals.*



4. Discussion

This paper explored recalibration methods for adapting a statistical model from case-control data to the general population for rare disease prediction. We show that the *Original* method tends to overestimate the probabilities in the general population, but that the *Albert Offset* [10], *Re-estimation* and *Recalibration* [7] approaches achieve good calibration of MODY probability predictions in both the model of the overall population and also in the model examining only the subset who were $C^+ \cap A^-$ (those genetically tested for MODY). The *Albert Offset* [10] and *Recalibration* [7] approaches achieved the smallest uncertainty around the observed probability of MODY. Furthermore, we showed that a mixture model can be developed to incorporate additional informative biomarkers (C-peptide and antibodies tests) that provide strong clinical information in predicting the probability of MODY in insulin-treated patients (noting that in our case, the biomarkers were only available in the population-level dataset, and not in the original case-control data). The *Recalibration mixture* model showed stability in our study and is the only approach that appropriately constrained the probability of MODY in $C^- \cup A^+$ individuals to be consistent with the strong prior information available in this setting. The *mixture* model approaches also allowed predictions in the absence of these additional tests. When developing prediction models for rare disease in practice, different approaches will be plausible in different scenarios based on the available data sources. Table 4 provides an overview on the advantages and disadvantages of all modelling approaches explored in the manuscript.

Table 4: Summary of key characteristics of the several modelling approaches.

Modelling Approaches	Advantages	Disadvantages
Original	- Model can be built in one training dataset (e.g. case-control) but probabilities can be updated based on information of prior probability from other studies.	- Individuals cannot have a recalibrated probability lower than the estimated prevalence in the general population. - Probabilities are aggregated into groups.

		- The recalibrated probabilities can be sensitive to the choice of grouping used.
Albert Offset	- Model can be built in one training dataset (e.g. case-control) and probabilities can be updated based on information of prior probability from other studies.	- Assumes the likelihood ratio for any given set of covariates is the same in the training data and calibration/target population.
Re-estimation	- Model can be built on one training dataset (population-representative) so no recalibration necessary.	- Requires a large sample size from the general population if model development required. - A lot of uncertainty in a rare disease setting due to very low number of positive cases.
Recalibration	- Scales odds ratios allowing for easy recalibration so that the model can be used in different settings (e.g.: general population or lab referral usage).	- Requires two datasets: a training dataset (e.g. case-control) and recalibration dataset (e.g. general population).
Re-estimation Mixture	- The model can be specified to include additional information on biomarker testing. - Model can be built on one training dataset (population-representative) so no recalibration necessary.	- Requires a large sample size from the general population. - A lot of uncertainty in a rare disease setting due to very low number of positive cases.
Recalibration Mixture	- The model can be specified to include additional information on biomarker testing. - Scales odds ratios allowing for easy recalibration so that the model can be used in different settings (e.g.: general population or lab referral usage).	- Requires two datasets: a training dataset (e.g. case-control) and recalibration dataset (e.g.: general population).

When only a training dataset (case-control dataset in our setting) is available and the aim is to adjust probabilities based on population prevalence, then the *Albert Offset* approach was the preferred method as it estimated the probability of MODY well in both our scenarios, with reasonable uncertainty in the predictions. In contrast, the *Original* approach [13] relies on thresholding probabilities and overestimated the probability of MODY in both scenarios. The *Albert Offset* approach has also been compared to other recalibration methods in other studies. Chan *et al*. (2008) had similar findings and deemed the *Albert Offset* the best approach [11]. In contrast, Grill *et al*. (2016) described this approach as the worse performing one in their study [12]. These differences may relate to the strong assumption of the *Albert Offset* approach that assumes the covariate distribution is the same in the training dataset as in the population for which the probabilities are adjusted [11, 10], and as we showed in sFig. 3, for datasets where the covariate distribution is different, the *Albert Offset* approach can perform poorly. When recalibrating models for different settings, the likelihood ratio assumption could become harder to justify depending on the specific setting, and particular caution would be required in populations where the clinical characteristics of

the patients differ substantially from the case-control dataset used for original model development. Establishing whether the similarities between covariate distributions is sufficient for pre-assessing the performance of the *Albert Offset* method is of interest for future research.

When only a population-representative dataset is available, the *Re-estimation* approach would be necessary. The *Re-estimation* approach calibrates well in the general population dataset for both scenarios but demonstrates a high level of uncertainty surrounding the model predictions. In contrast, Grill *et al.* (2016) describe the *Re-estimation* approach as having equivalent performance to the *Albert Offset*, with both performing worse than all other approaches [12]. The high uncertainty in our analysis can be attributed to the fact that there are only 7 positive MODY cases in the general population dataset (prevalence of 0.6%) and that the distribution of predicted probabilities is skewed towards zero. This highlights the problem with fitting models for rare diseases to population data [34], where the low prevalence means very large sample sizes would be required to reduce the uncertainty around the predictions, and it would be important to assess the adequacy of the sample size prior to model fitting [35]. When both training and population-representative datasets are available (as in our study), we showed that the *Recalibration* approach demonstrates good calibration in the population-representative dataset for both scenarios. This approach combines the information captured from the training data with information from the calibration dataset [7, 36], producing relatively low uncertainty in the model predictions compared to other approaches explored in this paper.

We also explored a scenario where additional biomarker testing was available but performed only on a limited subset of patients. Screening using biomarkers is common in clinical practice and often used in rare diseases where universal testing is not cost-effective or could be invasive (e.g. screening for chromosomal defects in pregnancy [37]). We developed a Bayesian hierarchical mixture model to follow the referral process involved in MODY testing and, therefore, utilise the additional biomarker tests for further refinement in the prediction of MODY probabilities. As a Bayesian model, we can incorporate additional information from other studies into the prior distributions for certain parameters, something previously explored by Boonstra *et al.* [14] in a different setting where additional information is only present for a subset of individuals. This approach has a further advantage in that predictions can still be made for patients with missing additional biomarkers, which are modelled using patient characteristics. This is important for our setting which instead of ignoring the biomarker results altogether, the model has used this information to improve predictions, so that even when biomarker information is missing, the MODY probabilities are a weighted sum across the latent biomarker test results, where the weights are informed by a model relating potential biomarker test outcomes conditional on a set of clinical features. We also combined the mixture model with the *Re-estimation* and the *Recalibration* approaches for just $C^+ \cap A^-$ individuals. Both approaches showed uncertainty levels in the probability predictions consistent with the previously observed uncertainty estimates. Furthermore, both approaches were tested for stability using bootstrapped versions of the population-representative dataset [33], demonstrating that the *Recalibration mixture* approach was more stable with the predicted probabilities of MODY than the *Re-estimation mixture* approach.

Other approaches for recalibration have been considered in previous work. Chan *et al.* (2008) [11] compared three methods to update pre-test probability with information on a new test: the Albert (*Albert Offset* in our study), Spiegelhalter and Knill-Jones (SKJ) [38] and Knottnerus [39] approaches. The SKJ represents an alternative to the *Albert Offset* approach, with similar performance in their paper. The Knottnerus approach was more suited to cases with sequential biomarker testing, which was not appropriate for our work since we did not have data on some

combinations of tests, instead we grouped biomarkers together into a composite measure T. The Knottnerus approach could be compared to the mixture model approaches (allowing for non-independence between both biomarkers), and examining these approaches when considering sequential testing of more than one biomarker could be considered in future work. Grill *et al*. (2016) compared several methods for incorporating new information into existing risk prediction models, logistic-new (equivalent to *Re-estimation*), LR-joint, LR-offset (*Albert Offset*), and LR-shrink (equivalent to SKJ from reference [11, 38]). In contrast to our study, their original models were built in population data, and the new datasets with additional features in were either cohort or case control data [12]. In the context of rare diseases, case-control data is likely to provide the best dataset for initial model development since this gives the most power for estimating model parameters and the population representative model can then borrow information from the case-control model. In cases where the additional data are only available in the case-control setting and original models were built on population data, however, the joint model approach (*Recalibration*) could be adaptable to this scenario.

The model that we recommend for the available MODY data is the *Recalibration mixture* approach. A major strength of this procedure is that it allows predictions for patients with missing biomarker testing [40] and this weighted prediction of MODY probability can be used to inform whether a patient should be referred for further testing. This model allows for the incorporation of strong prior information regarding the probability of having MODY for $C^- \cup A^+$ individuals, propagates uncertainties regarding the missing data in the UNITED study, and borrows weight from the case-control model through the recalibration procedure [7], thus improving the stability of predictions [33]. This model provides sensible predictions for the probability of MODY for patients with/without additional testing for C-peptide and antibodies. Patients with missing MODY testing (i.e. $C^- \cup A^+$) could have been set as a negative result test for all approaches due to the strong clinical knowledge of these tests being consistent with a Type 1 diagnosis. However, this may not be the case for other settings, where patients with missing outcomes could be believed to have a higher probability of the outcome, and therefore, assuming that the outcome is negative may be less justifiable.

There are some limitations to the *Recalibration Mixture* approach. We currently use biomarker tests as binary (positive/negative) results; in practice, biomarkers may be on a continuous scale. As such, the model could be adapted to include the biomarker results as additional covariates, which could be numerically integrated out if predicting to an individual that was missing this information in practice [40]. We are also limited by the small sample sizes in rare diseases [41], and even with our final model utilising two datasets, model predictions still have some uncertainty. However, we still saw good separation between MODY and T1D, and even accounting for the uncertainty, probability thresholds could be defined that rule out clear non-MODY cases and can be used to determine positive test rates at different probabilities in practice. These thresholds would balance the amount of testing to be carried out against the potential for missing genuine MODY cases, depending on how conservative the choice of threshold. The model has yet to be validated in a hold-out dataset, but the stability plots using bootstrapped datasets provide some insight into the stability of model predictions [33]. Although the 95% credible interval of bootstrapped probabilities is relatively wide at higher values, the 50% credible interval is narrow for all probabilities around the equal line.

This paper provides a comparison of several recalibration approaches. The development of our recalibration approach uses established methodologies and we have shown how it could apply to identifying patients at high probability of MODY to allow more targeted diagnostic testing, but

these ideas could be applied in other diseases. In practice, other settings could benefit from a similar Bayesian hierarchical model structure where informative biomarkers or additional testing information are available but only in a subset of patients due to its invasive nature or high cost of testing. With this structure, the model can be used to consider whether additional testing should be carried out when the individual already has a low probability (not on the cusp of referral), something explored previously in treatment selection for T2D [40]. Furthermore, this modelling structure could be particularly useful in other rare diseases with low sample sizes since it borrows weight from multiple datasets through recalibration, improving predictions.

5. Conclusion

We have compared several approaches that could be used to develop prediction models for rare diseases. We found the *Recalibration mixture* model approach to be the best approach combining both case-control and population-representative data sources, allowing the incorporation of additional data on biomarkers and appropriate prior probabilities.

6. Ethics approval and consent to participate

For the UNITED study, ethics approval was granted by the NRES Committee South West–Central Bristol (REC no. 10/H0106/03). For the case control data, this study was approved by the Genetic Beta Cell Research Bank, Exeter, UK (ethical approval was provided by the North Wales Research Ethics Committee, UK (IRAS project ID 231760)) and the Peninsula Research Bank (Devon & Torbay Research Ethics Committee, ref: 2002/7/118).

7. Consent for publication

Not applicable.

8. Availability of data and materials

Data are available through application to the Genetic Beta Cell Research Bank https://www.diabetesgenes.org/current-research/genetic-beta-cell-research-bank/ and the Peninsula Research Bank https://exetercrfnihr.org/about/exeter-10000-prb/. Example R code for fitting the approaches used in this study is available on GitHub within the repository "Exeter-Diabetes/Pedro-MODY_recal_approaches".

9. Competing interests

All authors declare no competing interests.

10. Funding

PC and TJM (McKinley) are funded by Research England's Expanding Excellence in England (E3) fund. KAP is a Wellcome Trust fellow (219606/Z/19/Z). This work was further supported by Diabetes UK (reference 21/0006328), the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC) and the National Institute for Health and Care Research Exeter Clinical Research Facility. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Wellcome Trust or the Department of Health and Social Care.

11. Authors' contributions

PC, BMS and TJM (McKinley) conceived and designed the study. PC, BMS and TJM (McKinley) analysed the data and developed the code. KAP and TJM (McDonald) provided the cohort data partly

used to estimate the prior distribution of a positive MODY test result, into those with C-peptide negative or autoantibody positive test results. ATH and ERP led the UNITED population study. All authors contributed to the writing of the article, provided support for the analysis and interpretation of results, critically revised the article, and approved the final article.

12. Acknowledgements

For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

13. References

[1] S. R. Johnson, J. J. Ellis, P. J. Leo, L. K. Anderson, U. Ganti, J. E. Harris, J. A. Curran, A. M. McInerney-Leo, N. Paramalingam, X. Song, L. S. Conwell, M. Harris, T. W. Jones, M. A. Brown, E. A. Davis and E. L. Duncan, "Comprehensive genetic screening: the prevalence of maturity-onset diabetes of the young gene variants in a population-based childhood diabetes cohort," *Pediatric Diabetes,* vol. 20, no. 1, pp. 57-64, 2018.

[2] A. A. Mitani and S. Haneuse, "Small data challenges of studying rare diseases," *Diabetes and Endocrinology,* vol. 3, no. 3, p. e201965, 2020.

[3] K. F. Schulz and D. A. Grimes, "Case-control studies: research in reverse," *Epidemiology,* vol. 359, no. 9304, pp. 431-434, 2002.

[4] S. Kölker, F. Gleich, U. Mütze and T. Opladen, "Rare disease registries are key to evidence-basec personalized medicine: highlighting the european experience," *Frontiers in Endocrinology,* vol. 13, p. 832063, 2022.

[5] S. Greenland, "Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies," *American Journal of Epidemiology,* vol. 160, no. 4, pp. 301-305, 2004.

[6] K. J. Rothman and S. Greenland, Modern Epidemiology, Philadelphia: Lippincott-Raven, 1998.

[7] E. W. Steyerberg, G. J. Borsboom, H. C. van Houwelingen, M. J. Eijkemans and J. D. F. Habbema, "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage," *Statistics in Medicine,* vol. 23, no. 16, pp. 2567-2586, 2004.

[8] E. Steyerberg, Clinical Prediction Models: a practical approach to development, validation and updating, Springer International P, 2019.

[9] P. Schuetz, M. T. Koller, M. Christ-Crain, E. W. Steyerberg, D. Stolz, C. A. Müller, H. C. Bucher, R. M. Bingisser, M. Tamm and B. Müller, "Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings," *Epidemiology and Infection,* vol. 136, no. 12, pp. 1628-1637, 2008.

[10] A. Albert, "On the use and computation of likelihood ratios in clinical chemistry," *Clinical Chemistry,* vol. 28, no. 5, pp. 1113-1119, 1982.

[11] S. F. Chan, J. J. Deeks, P. Macaskill and L. Irwig, "Three methods to construct predictive models using logistic regression and likelihood ratios to facilitate adjustment for pretest probability give similar results," *Journal of Clinical Epidemiology,* vol. 61, no. 1, pp. 52-63, 2008.

[12] S. Grill, D. P. Ankerst, M. H. Gail, N. Chatterjee and R. M. Pfeiffer, "Comparison of approaches for incorporating new information into existing risk prediction models," *Statistics in Medicine,* vol. 36, no. 7, pp. 1134-1156, 2016.

[13] B. M. Shields, T. J. McDonald, M. J. Campbell, C. Hyde and A. T. Hattersley, "The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes," *Diabetologia,* vol. 55, pp. 1265-1272, 2012.

[14] P. S. Boonstra and R. P. Barbaro, "Incorporating historical models with adaptive Bayesian updates," *Biostatistics,* vol. 21, no. 2, pp. e47-e64, 2020.

[15] W. Cheng, J. M. Taylor, T. Gu, S. A. Tomlins and B. Mukherjee, "Informing a risk prediction model for binary outcomes with external coefficient information," *Journal of the Royal Statistical Society. Series C, Applied statistics,* vol. 68, no. 1, pp. 121-139, 2019.

[16] K. Colclough and K. Patel, "How do I diagnose maturity onset diabetes of the young in my patients?," *Clinical Endocrinology,* vol. 97, no. 4, pp. 436-447, 2022.

[17] D. Gardner and E.-S. Tai, "Clinical features and treatment of maturity onset diabetes of the young (MODY)," *Diabetes, Metabolic Syndrome and Obesity,* vol. 2012, no. 5, pp. 101-108, 2012.

[18] R. Naylor, A. Johnson, D. Gaudio, M. Adam, J. Feldman, G. Mirzaa, R. Pagon, S. Wallace, L. Bean, K. Gripp and A. Amemiya, Maturity-onset diabetes of the young overview, University of Washington, Seattle, 1993-2023.

[19] L. Pang, K. C. Colclough, M. H. Shepherd, J. McLean, E. R. Pearson, S. Ellard, A. T. Hattersley and B. M. Shields, "Improvements in awareness and testing have led to a threefold increase over 10 years in the identification of monogenic diabetes in the U.K.," *Diabetes Care,* vol. 45, no. 3, pp. 642-649, 2022.

[20] M. Shepherd, B. Shields, M. Hudson, E. Pearson, C. Hyde, S. Ellard, A. Hattersley and K. Patel, "A UK nationwide prospective study of treatment change in MODY: genetic subtype and clinical characteristics predict optimal glycaemic control after discontinuing insulin and metformin," *Diabetologia,* vol. 61, no. 12, pp. 2520-2527, 2018.

[21] B. Shields, M. Shepherd, M. Hudson, T. McDonald, K. Colclough, J. Peters, B. Knight, C. Hyde, S. Ellard, E. Pearson, A. Hattersley and UNITED study team, "Population-based assessment of a biomarker-based screening pathway to aid diagnosis of monogenic diabetes in young-onset patients," *Diabetes Care,* vol. 40, no. 8, pp. 1017-1025, 2017.

[22] S. Y. Agladioglu, E. Sagsak and Z. Aycan, "Urinary C-peptide/creatinine ratio can distinguish maturity-onset diabetes of the young from type 1 diabetes in children and adolescents: a single-center experience," *Hormone Research in Paediatrics,* vol. 84, no. 1, pp. 54-61, 2015.

[23] T. J. McDonald, K. Colclough, R. Brown, B. Shields, M. Shepherd, P. Bingley, A. Williams, A. T. Hattersley and S. Ellard, "Islet autoantibodies can discriminate maturity-onset diabetes of the

young (MODY) from Type 1 diabetes," *Diabetic Medicine,* vol. 28, no. 9, pp. 1028-1033, 2011.

[24] R. E. Besser, B. M. Shields, S. E. Hammersley, K. Colclough, T. J. McDonald, Z. Gray, J. J. Heywood, T. G. Barrett and A. T. Hattersley, "Home urine C-peptide creatinine ratio (UCPCR) testing can identify type 2 and MODY in pediatric diabetes," *Pediatric Diabetes,* vol. 14, no. 3, pp. 181-188, 2013.

[25] T. J. McDonald, K. Colclough, R. Brown, B. Shields, M. Shepherd, P. Bingley, A. Williams, A. T. Hattersley and S. Ellard, "Islet autoantibodies can discriminate maturity-onset diabetes of the young (MODY) from type 1 diabetes," *Diabetic Medicine,* vol. 28, no. 9, pp. 1028-1033, 2011.

[26] A. Carlsson, M. Shepherd, S. Ellard, M. Weedon, Å. Lernmark, G. Forsander, K. Colclough, Q. Brahimi, C. Valtonen-Andre, S. Ivarsson, H. Larsson, U. Samuelsson, E. Örtqvist, L. Groop, J. Ludvigsson, C. Marcus and A. Hattersley, "Absence of islet autoantibodies and modestly raised glucose values at diabetes diagnosis should lead to testing for MODY: lessons from a 5-year pediatric swedish national cohort study," *Diabetes Care,* vol. 43, no. 1, pp. 82-89, 2020.

[27] S. A. Greeley, M. Polak, P. R. Njølstad, F. Barbetti, R. Williams, L. Castano, K. Raile, D. V. Chi, A. Habeb, A. T. Hattersley and E. Codner, "ISPAD clinical practice consensus guidelines 2022: the diagnosis and management of monogenic diabetes in children and adolescents," *Pediatric Diabetes,* vol. 23, no. 8, pp. 1188-1211, 2022.

[28] National Health Service, "National Genomic Test Directory: testing criteria for rare and inherited disease.," [Online]. Available: https://www.england.nhs.uk/wp-content/uploads/2018/08/rare-and-inherited-disease-eligibility-criteria-v2.pdf. [Accessed 06 08 2023].

[29] A. Gelman, J. Carlin, H. Stern and D. Rubin, Bayesian data analysis, New York: Chapman and Hall/CRC, 1995.

[30] G. Thanabalasingham, A. Pal, M. P. Selwood, C. Dudley, K. Fisher, P. J. Bingley, S. Ellard, A. J. Farmer, M. I. McCarthy and K. R. Owen, "Systematic assessment of etiology in adults with a clinical diagnosis of young-onset type 2 diabetes is a successful strategy for identifying maturity-onset diabetes of the young," *Diabetes Care,* vol. 35, no. 6, pp. 1206-1212, 2012.

[31] R. E. Besser, M. H. Shepherd, T. J. McDonald, B. M. Shields, B. A. Knight, S. Ellard and A. T. Hattersley, "Urinary C-peptide creatinine ration is a practical outpatient tool for identifying hepatocyte nuclear factor 1-α/hepatocyte nuclear factor 4-α maturity-onset diabetes of the young from long-duration type 1 diabetes," *Diabetes Care,* vol. 34, no. 2, pp. 286-291, 2011.

[32] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science,* vol. 7, no. 4, pp. 457-472, 1992.

[33] R. D. Riley and G. S. Collins, "Stability of clinical prediction models developed using statistical or machine learning methods," *Biometrical Journal,* p. 2200302, 2023.

[34] R. Griggs, M. Batshaw, M. Dunkle, R. Gopal-Srivastava, E. Kaye, J. Krischer, T. Nguyen, K. Paulus and P. Merkel, "Clinical research for rare disease: opportunities, challenges, and solutions," *Molecular Genetics and Metabolism,* vol. 96, no. 1, pp. 20-26, 2009.

[35] A. Mitani and S. Haneuse, "Small data challenges of studying rare diseases," *JAMA Network*

Open, vol. 3, no. 3, p. e201965, 2020.

[36] K. Moons, A. Kengne, D. Grobbee, P. Royston, Y. Vergouwe, D. Altman and M. Woodward, "Risk prediction models: II. external validation, model updating, and impact assessment," *Heart,* vol. 98, pp. 691-698, 2012.

[37] D. Wright, K. Kagan, F. Molina, A. Gazzoni and K. Nicolaides, "A mixture model of nuchal translucency thickness in screening for chromosomal defects," *Ultrasound in Obstetrics & Gynecology,* vol. 31, no. 4, pp. 376-383, 2008.

[38] D. Spiegelhalter and R. Knill-Jones, "Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology," *Journal of the Royal Statistical Society. Series A,* vol. 147, no. 1, pp. 35-77, 1984.

[39] J. Knottnerus, "Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary varibles," *Medical Decision Making,* vol. 12, no. 2, pp. 93-108, 1992.

[40] P. Cardoso, J. M. Dennis, J. Bowden, B. M. Shields and T. J. McKinley, "Dirichlet process mixture models to impute missing predictor data in counterfactual prediction models: an application to predict optimal type 2 diabetes therapy," *BMC Medical Informatics and Decision Making,* vol. 24, no. 12, 2024.

[41] R. D. Riley, K. I. Snell, D. L. Burke, F. E. Harrel Jr., K. G. Moons and G. S. Collins, "Minimum samples size for developing a multivariate prediction model: part II - binary and time-to-event outcomes," *Statistics in Medicine,* vol. 38, pp. 1276-1296, 2019.