NUFFIELD DEPARTMENT OF
PRIMARY CARE
HEALTH SCIENCES

QRESEARCH

Radcliffe Observatory Quarter, Woodstock Road, Oxford. OX2 6GG

Tel: +44(0)1865 289300 • qresearch@phc.ox.ac.uk

23 June 2022

Dr Pui San Tan
University of Oxford

Dear Dr Tan

Reference	OX153
Title	Development and validation of a decision risk tool (equation) [QPancreasD)] for early diagnosis of pancreatic ductal adenocarcinoma (PDAC) among new-onset diabetes in UK primary care
Chief Investigator(s)	Professor Julia Hippisley-Cox/ Dr Pui San Tan
Date Amendment requested	22.12.2022
Date Amendment approved	23.06.2022
Approval granted until	23.06.2024
Special conditions of approval	N/A

Thank you for submitting your amendment application to QResearch.

Scientific Opinion

The members of the committee taking part in the review gave a favourable scientific opinion of the application for an amendment on the basis of the amendment application form and lay summary submitted to QResearch. The committee is satisfied that the research is appropriately designed in relation to its objectives and is likely to add something useful to existing knowledge in accordance with section 4.1.2 of the ethical conditions for QResearch. https://www.qresearch.org/media/1155/257790-18em0400-qresearch-conditioins-of-ethical-approval-27122018.pdf

QResearch is a Research Ethics Approved Research Database and therefore this also constitutes research ethics approval. Please quote REC reference (18/EM/0400) in any publication arising from this work.

Data Access

Subject to funding and relevant Data Sharing Agreements, the dataset requested will be provided to you.

Approved documents

The documents reviewed were

Document	version	Date

Amendment Application Form		22.12.2021

Reviewing Members of the Committee

The members of the scientific committee who took part in the review are listed on the attached sheet.

Annual Review and research outputs

We will contact you annually for an update on progress with your research.

Amendments and changes in responsibility

Any changes to the protocol or responsibility of the applicants for the conduct of the research should be notified to QResearch@phc.ox.ac.uk

Dr Paula Dhiman
Chair of QResearch Scientific Committee

Enclosures:
List of names and professions of members who took part in the review

Copy to: Prof Julia Hippisley-Cox, Chief Investigator of QResearch.

Reviewing Members of the Committee

Name	Profession
Dr Paula Dhiman	Senior Researcher in Medical Statistics

QResearch Application Amendment Form for Scientific Committee

Project Information

1. **Project Reference Number:** OX153
2. **Project Title:** Development and validation of a decision risk tool (equation) [QPancreasD)] for early diagnosis of pancreatic ductal adenocarcinoma (PDAC) among new-onset diabetes in UK primary care
3. **Chief Investigator:** Professor Julia Hippisley-Cox & Dr Pui San Tan

Amendment Information

4. **Name of the person requesting the amendment:** Dr. Pui San Tan
5. **Institutional email address of the person requesting the amendment:** pui.tan@phc.ox.ac.uk
6. **Date amendment requested:** 22/12/2021
7. **Amendment number:** 1
8. **Are there any changes to the hypotheses being tested?:** No
9. **Are there any changes to the design of the study?:** No
10. **Are there any changes to the definition of the study population?:** Yes

Yes - previously only excluded those with type 2 diabetes diagnosis prior to study entry

Exclusion - to add also exclusion to those with records of diabetes medication use prior to type 2 diabetes diagnosis as they could be suggestive of prevalent diabetes.

11. **Are there any changes to the exposures, comparators, or outcomes?:** Yes

Yes.
Outcome – change from PDAC (pancreatic ductal adenocarcinoma) to pancreatic cancer as a whole as risk of pancreatic cancer in type 2 diabetes is not just limited to the specific subtype i.e. PDAC, and histology records are not complete to be able to distinguish between different histologies very well

Covariates (changes)
Demographics – age (year of birth) change to age at first diagnosis of diabetes, drop GP practice and region as they will not be useful variables for risk prediction

Conditions (based on updated literature review)
- drop hepatitis B – no evidence from literature
- add family history of diabetes
- drop hypercholesterolaemia

- autoimmune diseases – only keep coeliac disease and HIV/AIDS

Tests (based on updated literature review)
- glucose test only keep HbA1c (drop glucose fasting/random) (HbA1c is more reliable measure of glucose control)
- LFT only keep ALT and bilirubin (drop AST, GGT – not well recorded)
- Drop Weight (BMI is better captured)
- Drop Cholesterol Tested - HDL cholesterol, LDL cholesterol, cholesterol/HDL ratio (no strong evidence in literature)
- Drop Cancer marker – CA-199 pancreas23, CA-125, CEA (very rarely measured)

Drugs – drop diabetes drugs (not likely to be on diabetes drugs prior to type 2 diabetes diagnosis) and immunosuppresants (very rarely used/captured)

12. **Are there any changes to the analysis strategy including not performing a planned analysis?:** Yes

Statistical analysis
Model development – previously we mentioned we will develop the model using
(1) Clinical significance-based approach: fit a full model with all candidate predictors, and retain variables that show clinically meaningful effects i.e. HR <0.9 or >1.1 (with p<0.01)
(2) Data-driven approach: fit a full model by including all potential predictors, followed by backwards elimination approach at specified significance level of p<0.01.

Considering approach (2) which retains predictors using p<0.01 is also built into (1) which also retains variables using p<0.01, we will instead take (1) as the one approach which combines both clinical and statistical significance, and approach (2) will be dropped as updated below:
We will fit the model using a combination of clinical and statistical significance approach. Binary and categorical variables will be kept in the model if they show clinical significance i.e. HR <0.9 or >1.1 and statistical significance (p<0.01). For continuous predictors, we will model these using fractional polynomial (FP) terms to account for non-linear associations and include based on statistical significance (p<0.01).

Cox-proportional hazards model will be used to model predictor variables and outcome. Non-linear relationships of continuous variables with outcome will be considered and modelled with fractional polynomials. Further, established risk factors based on prior knowledge (e.g. age, sex) will be retained in the model irrespective of clinical or statistical significance. Pre-specified interactions between relevant variables will also be examined, for example between age and family history of gastrointestinal cancer. Only in the case of non-proportional hazards, we will explore time-varying risks using flexible parametric models to explore potential changing risks over time of developing pancreatic cancer from type 2 diabetes diagnosis.

Formulation of risk equations
We will use regression coefficients obtained from the model as weights to be combined with baseline survivor function for 2 years to compute absolute risk equations. Baseline survivor function will be estimated by centring all continuous variables at mean and setting binary variables at zero.
We will not adjust for optimism as overfitting is less likely to be an issue in a large cohort in QResearch and might be a challenge to incorporate within an IECV framework.

We will also perform sensitivity analyses to compare robustness of results between

complete-case and imputed analyses.

Machine learning
We will utilise the pseudovalues-based approach as below:
Jack-knife pseudo-values of the Kaplan-Meier failure function at 2 years follow-up will be estimated for the entire cohort using the 'stpsurv' command in Stata. These will be used as continuous outcome variables to fit a neural network model and an XGBoost model to the entire cohort data, which will then undergo evaluation with IECV. For the neural network (NN), continuous variables will be min-max scaled (to between 0-1), and for both models, categorical variables with more than 2 levels will be converted to dummies. To handle missing data, the imputations obtained earlier will be stacked to form a single 'long' dataset, with each observation assigned a weight of 1/number of imputations. This enables all imputations to be used in the model fitting by weighting the contribution of each individual's 'variant' to the loss function. This 'stacked and weighted' approach circumvents the issue that these models cannot pool parameters/standard errors in the same fashion as a regression model. The neural network will have a single output node with a linear activation function, ReLU activation functions on nodes in hidden layers, use the Adam optimiser, and use the root mean squared error (RMSE) between predicted and observed pseudo-values as a loss function. The XGBoost model will have a regression objective, with the same loss function as the NN.

13. **Are there any changes to other methods such as to control for missing data, confounding, or sensitivity analyses?:** Yes

Corresponding comparison of models based on changes above:

Comparison of models (earlier version)
We will compare performances of the following models in terms of calibration and discrimination using measures as described above
i. risk equation using regression approach (variable selection: clinical-significance-based)
ii. risk equation using regression approach (variable selection: data-driven-based)
iii. risk equation using machine-learning approach (ANN and XGBoost) (variable selection: best performing approach between clinical-significance vs data-driven as accessed in (i) and (ii))
iv. existing risk equation (QCancer)
v. simple decision support risk equation using various cut-offs of BMI and age (comparing sensitivity and specificity) (this potentially could be implemented easily as opposed to more complex models above)

Model comparisons (updated version)
In addition, we will compare performance of the following models/interventions in terms of calibration, discrimination and decision curve analysis (where applicable). Decision curve analysis will be performed to evaluate clinical benefit of following interventions across different risk threshold probabilities compared to "screen all" and "screen none" approaches:
i. risk equation using regression approach
ii. risk equation using machine-learning approach (ANN) (variable selection: best performing as determined in ML1-3 above)
iii. risk equation using machine-learning approach (XGBoost) (variable selection: best performing as determined in ML1-3 above)
iii. existing risk equation (QCancer)
iv. simple decision support risk equation using NICE guidelines; aged 60 and over presenting with both weight loss and new-onset diabetes (comparing sensitivity and specificity - this potentially could be implemented easily as opposed to more complex models above).

Note: ML1) Include the variables selected in the regression modelling approach, ML2) A full model with all candidate predictors to estimate a ‘benchmark’ for a maximally complex model, and ML3) variable selection based on a threshold of relative variable importance after fitting the previous ‘maximally complex’ models.

14. **Are there any changes to the linkages to other databases, including additional linkages, or not using linked data which were part of the approved protocol?:** No
15. **Are there any changes to the research team including the Chief Investigator, co-applicants, or researchers?:** No
16. **Are there any changes to the UK university at which any of the research team including the Chief Investigator, co-applicants, or researchers are based?:** No
17. **Are there any other changes not listed above?:** No
18. **References:** N/A