Project DELTA - integrateD diagnostic solution for EarLy deTection of oesophageal cAncer

Work Package 1:

Data Mining and Algorithm Development

Research Protocol

Research reference numbers	
Protocol version and date	Version 1.5 11th February 2021.
REC reference:	18/EM/0400
IRAS Number:	257790
QResearch reference:	OX39
Sponsor:	University of Oxford
Funder	INNOVATE UK
Funder reference:	41162
Related protocols	DELTA Protocol for WP 2-4, Fitzgerald etc al.

Table of Contents

1 Project Team

Principal investigator and WP1 led

Julia Hippisley-Cox, Professor of Clinical Epidemiology and General Practice, University of Oxford.

Co-investigators

Rebekah Fitzgerald, Professor of cancer prevention, director of the MRC cancer unit at the University of Cambridge, Honorary Consultant in Gastroenterology and General Medicine at Addenbrooke's Hospital, Cambridge.
Winnie Mei, Epidemiologist, University of Oxford
Pui San Tan, Epidemiologist and Pharmacist University of Oxford
Carol Coupland, Honorary Professor of Medical Statistics in Primary Care, University of Oxford.
Peter Sasieni, Professor of Cancer Prevention and Academic Director of the King's Clinical Trials Unit and Cancer Research UK & King's College London Cancer Prevention Trials Unit

Collaborators

Alison Hall, Foundation for Genomics and Population Health (PHG Foundation)

2 Programme summary

2.1 Background

Oesophageal cancer is the sixth most common cause for cancer related deaths with over 450,000 new cases and 400,000 resulting deaths per year globally. Most cases in the UK are adenocarcinoma with some of the poorest outcomes from this cancer type in Europe -- mainly

due to late diagnosis. The main risk factor is chronic reflux disease and due to the high prevalence and non-specific nature of these symptoms most patients are managed with acid-reflux (PPI) medication without referral for endoscopy. For those patients that are referred the endoscopy is normal in over 70% of cases, and there is not enough capacity for endoscopy especially considering colon cancer screening.

2.2 Innovation

The Cytosponge-TFF3 test is a disruptive technology developed by Cambridge University that could revolutionise the clinical care pathway for reflux disease, which is a risk factor for adenocarcinoma. This device has been shown to be safe and acceptable to patients in studies involving >4,000 individuals across 3 continents and a randomised study of over 13,000 eligible individuals on PPI medication to identify Barrett's oesophagus has just been successfully completed[1]. This innovation could focus procedures on those at greatest risk of cancer - especially relevant since we have effective, NICE approved endoscopic interventions for early oesophageal cancer. Through improved diagnosis we can also reduce the over-use of proton pump inhibitor (PPI) medication which is expensive with long-term side-effects. Health economic modelling studies have shown that this is a cost-effective solution falling below the NICE threshold. It would also be possible to implement Cytosponge for individuals at risk for squamous cell carcinoma of the oesophagus. The device has been licensed to Covidien GI Solutions, now Medtronic by the MRC. A new Early Detection company called Cyted has been spun out from the University of Cambridge and is providing quality assured processing of Cytosponge with AI solutions for economic high throughput (Company Number: 11478299).

2.3 Main Aims

2.3.1 Work package 1

Mine electronic health records and endoscopy databases at a national level to improve identification of individuals at increased risk of oesophageal cancer

2.3.2 work package 2

Build a transferrable operating model for a nurse-lead Cytosponge clinic

2.3.3 Work package 3

Implement Artificial Intelligence algorithms for high throughput computational pathology for the Cytosponge-TFF3 test and endoscopic biopsies. Samples/data collected from newly recruited patients will feed this activity.

2.3.4 Work package 4

Health economic and implementation research to assess effectiveness of the novel pathway including user preferences for patients and clinicians. Data collected from newly recruited patients will feed this activity.

2.4 Vision

To re-design and evaluate the clinical pathway such that we systematically identify those at risk, perform a simple test to inform who needs endoscopy and in so doing rationalise the use of long-term PPI medication. Due to the cost-effectiveness of Cytosponge-TFF3 compared with endoscopy these changes will likely result in an economic benefit to the NHS, a social benefit for early detection of a lethal cancer and a reduction in over-use of PPI medication.

3 Summary Work Package 1

Objective To derive and evaluate a risk prediction tool to predict oesophageal cancer (all types and according to squamous or adenocarcinoma subtype) which can be applied in primary care to identify high risk patients suitable for assessment with the Cytosponge in primary care.

Design Prospective population based open cohort study using routinely collected data from 1500 GP practices in England in the QResearch database (derived from EMIS electronic medical records) between 01.01.2000 and 31.12.2020 (or latest data available). We will use the QResearch database to develop the risk equations. Models will be developed on a derivation dataset from a subset of approximately 1100 practices and validated in the remaining 400 practices. The model will be externally evaluated in other relevant datasets as they become available.

Subjects We will study adults aged 40 years and over who are free of oesophageal cancer at study entry. Patients with a new onset of alarm symptoms (e.g. dysphagia, weight loss etc) in the preceding 3 months will also be excluded.

Methods Cox time-to-event models will be used in the derivation data to derive separate risk equations in males and females for estimating risk of oesophageal cancer. Predictors considered will include age, ethnicity, deprivation, smoking status, alcohol intake, body mass index, pre-existing medical co-morbidities, and concurrent medication. Measures of performance (prediction errors, calibration and discrimination) will be determined in the validation data for men and women separately and by ten-year age group.

Outcome Our primary outcome is oesophageal cancer defined as recorded on GP record or linked hospital, cancer registry or death certificate.

4 Background

4.1 Clinical need

5.1.1 Epidemiology of oesophageal cancers

Oesophageal cancer is the 8th most common cancer and one of the deadliest cancers in the world[2]. Although oesophageal squamous cell carcinoma (OSCC) is the predominant histological type worldwide, oesophageal adenocarcinoma (OAC) is more common in developed countries such as United States, Australia, United Kingdom, and Western Europe[2].

Incidence of the cancer type OAC has increased 6-fold since the 1990s and carries a dismal prognosis. The UK has some of the worst outcomes from this disease in Europe. Clinical guidelines have focussed on minimising endoscopy referrals unless patients have "alarm symptoms" suggestive of cancer. Nevertheless, General Practice referral rates vary widely, and low endoscopy referral rates have been linked with poor outcomes from oesophageal cancer.

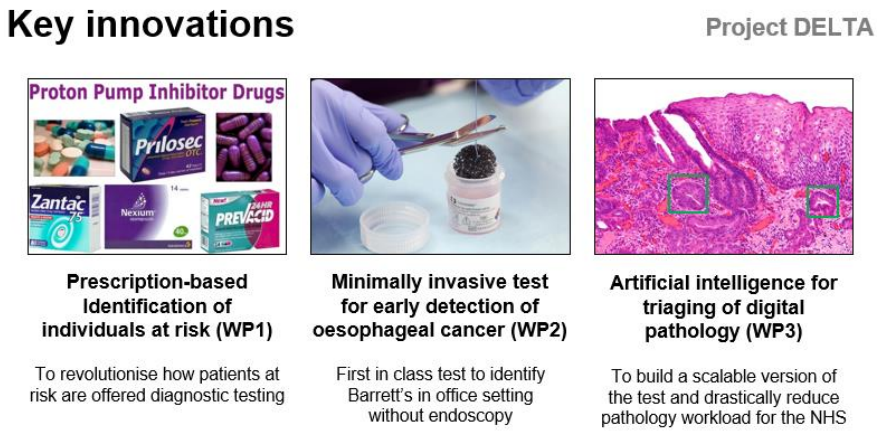
A major risk factor for this cancer is chronic heartburn caused by reflux. Three to six percent of individuals with reflux predominant symptoms may have the precursor lesion called Barrett's oesophagus, but only around 20% of patients with Barrett's are diagnosed. It is estimated that the burden of OAC could be reduced by up to 50% as a result of increasing the proportion of individuals with reflux symptoms who are investigated. This is a formidable task since heartburn symptoms affect between 5%-20% of the population and account for up to 10% of GP consultations. GPs therefore focus on controlling reflux symptoms with acid-suppressant medication, particularly proton pump inhibitor therapy (PPI). PPIs are highly effective, but patients often continue taking them life-long and there are increasing concerns about long-term side effects including osteoporosis, pneumonia[3 4] and recently allergy[3] .

4.2 Introduction to DELTA

In 2008 the Chief Medical Officer, Sir Liam Donaldson, raised oesophageal cancer as a public health concern and identified an urgent need to develop a need for a safe, minimally invasive, affordable test applicable to the office setting to diagnose Barrett's oesophagus. Cambridge University have developed a new minimally invasive test for patients with reflux that can be performed in the GP surgery. This test is called Cytosponge™ -TFF3 which has been tested in over 4,000 individuals across 3 continents.

Our vision is to re-design the clinical pathway for reflux. In primary care a new algorithm applied to NHS prescribing databases will flag symptomatic individuals at risk for oesophageal cancer to their GP. Individuals most at risk will be offered a Cytosponge™ test in a nurse led clinic. The sample will be sent to a centralised laboratory for processing. The H&E and TFF3 stained cells collected by the Cytosponge™ will be assessed using Artificial Intelligence to increase the throughput and reduce the cost. For cases with atypia detected by the

pathologist a p53 stain will be added. Individuals at high risk for cancer will be referred for endoscopy and PPI use will be rationalised.

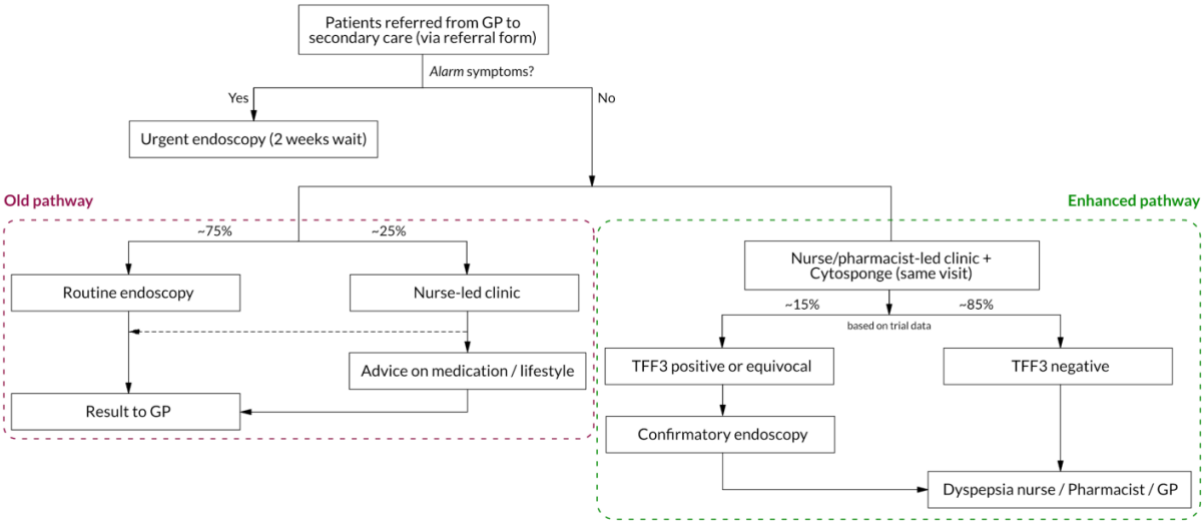


These changes could result in more efficient triage to endoscopy, an economic benefit to the NHS, a social benefit for early detection of a lethal cancer and a reduction in over-use of PPI medication.

4.3 Enhancing the primary care clinical pathway

Currently the clinical pathway for reflux in primary care relies on excluding alarm symptoms (e.g. dysphagia, weight loss) and managing reflux symptoms with acid-suppressant medication. The NICE guidelines recommend endoscopy via the 2 Week Wait pathway if alarm symptoms are present, and routine referral if the symptoms are persistent and occur in the context of other risks factors such as family history (CG27). If a patient is referred to secondary care via the routine pathway the vast majority of patients will be triaged straight to endoscopy. Clinic appointments in secondary care may be led by a dyspepsia nurse or a Gastroenterologist depending on the local policy and resources.

The current, conventional care pathway has a strong focus on triggering endoscopies which results in high costs for the NHS. Additionally, around 70% of these endoscopies are normal with no clinically significant finding (Endoscopy Service Report). The proposed, enhanced care pathway in this project implements a nurse/pharmacist-led clinic for patients with reflux predominant symptoms whereby a Cytosponge procedure is performed in a one-stop visit. This serves as an intermediate step in which the Cytosponge results, together with the high specificity of the test, are used to avoid unnecessary endoscopies.



The Cytosponge™ can be readily placed into this pathway (see Fig). We propose that primary care individuals at high risk including those undergoing repeat prescriptions for reflux symptoms are considered for a Cytosponge™ test, with more consideration given to their requirement for long term PPI medication as part of this process. The patients could be flagged electronically to a GP and the test performed by the practice nurse. Alternatively, with the expanding role of the community Pharmacist they would be ideally placed to perform this role. Thus, only patients deemed to be at risk for Barrett's oesophagus (or another oesophageal condition, or gastric IM) ascertained by the Cytosponge™ test would be referred for endoscopy.

4.4 Introduction to cancer risk prediction tools

Over the last 10 years, Hippisley-Cox et al have developed, validated and implemented a novel set of risk prediction algorithms collectively known as the QCancer algorithms which predict the risk of different types of cancer using readily available information from routinely available electronic health records[5-12].

The first set of algorithms are designed to improve early diagnosis of an existing cancer on the basis of a combination of symptoms and readily available risk factors in order to identify those patients needing urgent investigation and referral[5].

The second set of algorithms are designed to estimate longer term cancer risks in order to identify high risk asymptomatic patients with combinations of risk factors who might benefit from systematic screening or interventions to reduce their risk[13].

This project provides the opportunity to introduce improvements to both sets of QCancer algorithms which are now possible due to the increased size of the QResearch database (which will enable us to distinguish between gastric and oesophageal cancer). It also provides an opportunity to explore risks associated with long-term medication use. For example, we can explore risks associated with proton pump inhibitors which are used to treat chronic

reflux symptoms which could also potentially mask development of oesophageal cancer; use of NSAIDs which tend to increase reflux symptoms and use of medications such as aspirin[14 15] or statins[15] which may lower risk of oesophageal cancer on their own or in combination with PPI[16] or cyclo-oxygenase inhibitors[15 17].

5 Objectives

The primary objective is to update and validate the QCancer algorithm to identify patients at highest risk of oesophageal cancer who may be suitable for assessment using the Cytosponge device.

Once developed and validated, we envisage that the updated algorithm could be used:

1. Within a consultation between the patient and a clinician with the intention of sharing the information with the patient to assess management options including assessment with Cytosponge
2. To electronically risk-stratify populations by applying the algorithm to all patients to then target and recall patients for a medication review and consideration of Cytosponge sponge assessment based on their levels of risk.
3. To inform mathematical modelling of the potential impact of changing the clinical pathway to include risk assessment +/- Cytosponge.
4. Adapted for use by the general public to improve communication and understanding of risk through implementation into web-based tools
5. Use by researchers to help generate new knowledge or insights.

6 Methods

6.1 Study design

We will undertake a cohort study in a large population of primary care patients using the latest version of the QResearch® database (currently version 45).

6.1.1 Data Sources

QResearch is a high-quality research database established in 2002 which has been used extensively for the development of risk prediction tools which are widely used across the NHS[18-23] as well as a wide range of high impact epidemiological research[24-26]. QResearch is a large, representative, validated GP practice research database nationally[27]. The database is linked at an individual patient level to hospital admissions data (including intensive care unit data), cancer registrations and mortality records obtained from the Office for National Statistics. The records are linked using a project-specific pseudonymized NHS number. The

recording of NHS numbers is valid and complete for 99.8% of QResearch patients, 99.9% for ONS mortality records and 98% for hospital admissions records[28 29].

This project will use all four linked data sources (GP, hospital, mortality and cancer registry)

6.2 Study Population

6.2.1 Practice inclusion

We will include all practices in England who had been using their EMIS computer system for at least a year. We will randomly allocate three quarters of QResearch practices to the derivation dataset and the remaining quarter to a test (validation) dataset.

6.2.2 Patient inclusion

We will identify an open cohort of individuals aged 25-90 years who are registered with practices that have been contributing to QResearch for over 12 months on or after 1st January 2000 (study start date).

6.2.3 Patient exclusions

Exclusions will broadly match those for WP2 which aims to implement the risk algorithm for recalling patients for the Cytosponge and hence the following will be excluded

- Patients with a new onset of alarm symptoms (including haematemesis, unexplained weight loss, dysphagia) in the last 3 months before study entry since these are likely to require urgent referral on the 2-Week Wait pathway for endoscopy.
- Recorded diagnosis of a current or previous oro-pharynx, oesophageal or gastro-oesophageal tumour
- Received prior surgical intervention to the oesophagus
- Recorded oesophageal varices, cirrhosis of the liver

Patients with recorded Barrett's oesophagus will be included and the presence of Barrett's oesophagus will be evaluated as a risk factor.

6.3 Study period

Patients will enter the cohort on the latest of the study start date or the date on which they become 25 or 12 months after registering with their GP practice. Patients will be followed up until they develop oesophageal cancer, die, leave the practice or the study end date or a maximum follow up period of 15 years.

6.4 Primary Outcome

Our primary outcome of interest is the diagnosis of oesophageal cancer during follow up (all

types and subdivided into adenocarcinoma or squamous cell cancer). We will ascertain the diagnoses on the basis of a diagnosis recorded in any of the four linked data sources (1) patients GP record (2) on their linked mortality record (3) hospital record or (4) cancer registry record. We will use the earliest recorded date of oesophageal cancer diagnosis on any of the four data sources as the index date.

6.5 Secondary outcome

Our secondary outcome is the diagnosis of Barrett's Oesophagus.

6.6 Predictor variables

We will examine the following candidate predictor variables based on 12 risk factors already in the QCancer tool[13] (marked with an asterix) as well as risk factors identified in the literature. The predictor list can be amended by future updates as new knowledge on emerging risk factors becomes available.

6.6.1 Demographic variables

1. Age (continuous variable)*.
2. Townsend deprivation score*. This is an area-level continuous score based on the patients' postcode[30]. Originally developed by Townsend[30], it includes unemployment (as a percentage of those aged 16 and over who are economically active); non-car ownership (as a percentage of all households); non-home ownership (as a percentage of all households) and household overcrowding. These variables are measured for a given area of approximately 120 households, via the 2011 census, and combined to give a "Townsend score" for that area. A greater Townsend score implies a greater level of deprivation.
3. Ethnicity (9 categories)

6.6.2 concurrent medication (at study entry)

1. Proton pump inhibitors (which are used to treat reflux symptoms)
2. H2 blockers (which are used to treat reflux symptoms)
3. NSAIDs (which can worsen reflux symptoms)
4. Aspirin (which may lower OAC risk)[14 16]
5. Statins (which may lower OAC risk)[17 31]
6. COX inhibitors (which may lower OAC risk)[17]
7. Metformin (which may lower OAC risk)[15]
8. HRT (in female only, may lower OAC risk)[15]
9. Other diabetic drugs

6.6.3 Lifestyle and family history

4. Smoking status - non-smoker, ex-smoker, light smoker (1-9/day), moderate (10-19/day) or heavy(20+/day)*.
5. Body mass index (continuous variable, z-scores will be used)*.
6. Alcohol use -non-drinker; light drinker (<1 unit/day); moderate (3-6 units/day); heavy (6+ units/day)*.
7. Family history of bowel or gastric or colorectal cancer.

6.6.4 Co-morbidities and investigations

8. gastric cancer
9. Barrett's oesophagus*
10. Peptic ulcer disease*
11. Gastro-oesophageal reflux disease (including heart burn)
12. Type 1 and type 2 diabetes*
13. previous blood cancer*
14. previous breast cancer*
15. previous oral cancer*
16. previous pancreatic cancer*
17. Current H.Pylori infection (which may lower OAC risk)
18. pernicious anaemia
19. hiatus hernia
20. anaemia (including Haemoglobin values)
21. Full blood count
22. CT scan abdomen within the previous 5 years
23. CT scan pelvis within the previous 5 years
24. Barium meal/swallow within the previous 5 years
25. Endoscopy within the previous 5 years

All predictor variables will be based on the latest coded information recorded in the GP record prior to entry to the cohort.

6.7 Descriptive analysis

We will produce the following descriptive analyses for comparison with the literature and national statistics from the CRUK website.

- crude incidence of oesophageal cancer by age, sex, ethnicity, deprivation and calendar time.
- age standardised incidence of oesophageal cancer overall and by type.
- characteristics of cases diagnosed with oesophageal cancer - age at diagnosis, stage, grade, histology, treatment (surgery, radiotherapy, chemotherapy, other).

6.8 Development of the models

6.8.1 Model development overview

We will use the following steps:

1. Development of prognostic models for each outcome within the derivation data.
2. Evaluation of predictive performance in the validation data.

Separate models will be developed and evaluated for males and females.

6.8.2 Development of the models using the derivation data

For all analyses, the time origin is entry to the study cohort and the risk period of interest is from the time origin up to the first date of diagnosis of oesophageal cancer. We will develop and evaluate the risk prediction equations using established methods[19 32-35] We will use second degree fractional polynomials (i.e. with up to two powers)[31] to model non-linear relationships for continuous variables (age, body mass index and Townsend score). Models will include interactions between age and predictor variables focussing on predictor variables which apply across the age range where numbers allow.

6.8.3 Handling of missing data

For all predictor variables, we will use the most recently available value at the time origin. For indicators of co-morbidities and medication use, the absence of information being recorded is assumed to mean absence of the factor in question. There may be missing data in some variables due to never being recorded: ethnicity, Townsend score, body mass index, smoking status and alcohol intake. We will use multiple imputation with chained equations to replace missing values for these variables [36-39].

Prior to the imputation, a complete-case analysis will be fitted using a model containing only the continuous covariates within the derivation data to derive the fractional polynomial order and corresponding powers. Then a multiple imputation model using chained equations will be fitted in the derivation data and will include all predictor variables along with age interaction terms, the Nelson–Aalen estimators of the baseline cumulative hazard, and the outcome indicators (namely, oesophageal cancer). Separate imputation models will be fitted for men and women. We will carry out 5 imputations as this has a relatively high efficiency[29] and is a pragmatic approach accounting for the size of the datasets and capacity of the available servers and software.

Each analysis model will be fitted in each imputed data set. We will use Rubin's rules to combine the model parameter estimates across the imputed datasets[40].

6.8.4 Variable selection

We will fit models that include all predictor variables initially and retain variables if they have a hazard ratio of < 0.90 or > 1.10 (for binary variables) and are statistically significant at the 0.01 level. For previous diagnoses of other cancers, we will retain variables which were significant at the 0.05 level since some of the cancers are rare. In order to simplify the models, we will focus on variables for the most common conditions and medications and combine similar variables with comparable hazard ratios where possible. If some predictor variables result in very sparse cells (i.e. with not enough participants or events to obtain point estimates and standard errors), we will combine some of these if clinically similar in nature.

For PPI and H2 blockers medication usage, we will determine the association between risk of oesophageal cancer and type of medication (omeprazole etc), dose and duration of exposure (e.g. < 6 months; 6-11 months; 12-23 months; 24-47 months; 48 months or more). We will examine both usage at baseline and also as a time varying exposure during study follow-up.

6.8.5 Risk equations

We will use the regression coefficients for each variable from the final model as weights which we will combine with the baseline survivor function evaluated for each year up to 10 years to derive risk equations over a period of 10 years of follow-up[41]. This will enable us to derive risk estimates for each year of follow-up, with a specific focus on 10-year risk estimates. We will estimate the baseline survivor function based on zero values of centred continuous variables, with all binary predictor values set to zero.

6.8.6 Validation of the models

In the validation data, we will fit an imputation model to enable imputation of missing values for ethnicity, body mass index, alcohol and smoking status. We will carry out 5 imputations. We will apply the risk equations for males and females obtained from the derivation data to the validation data and calculate measures of performance.

As in previous studies[42], we will calculate R^2 values (explained variation where higher values indicate a greater proportion of variation in survival time explained by the model [43]), D statistics[44] (a measure of discrimination which quantifies the separation in survival between patients with different levels of predicted risk where higher values indicate better discrimination), Brier scores, and Harrell's C statistics at 1, 2, 5 and 10 years and combine these across datasets using Rubin's rules. Harrell's C statistic[45] is a measure of discrimination (separation) which quantifies the extent to which those with earlier events have higher risk scores. Higher values of Harrell's C indicate better performance of the model for predicting

the relevant outcome. A value of 1 indicates that the model has perfect discrimination. A value of 0.5 indicates that the model discrimination is no better than chance.

We will calculate 95% confidence intervals for the performance statistics to allow comparisons with alternative models for the same outcome and across different subgroups.[46]

We will assess calibration of the risk scores by comparing the mean predicted risks with the observed risks by tenth of predicted risk. The observed risks will be obtained using Kaplan-Meier estimates evaluated at 10 years, obtained for men and women.

We will also evaluate these performance measures in 6 pre-specified age groups (25-49; 50-59; 60-69; 70-79; 80+), and in people on long term PPIs.

We will also compare the risk prediction model with a simple algorithm based only on criteria such as age and number of electronic prescriptions of PPI/H2 blockers and consider the use of decision curve analysies.

6.8.7 Updating of the model using new data

We will update the models to ensure the model remains up to date. The baseline survivor function may change after the widespread introduction of the Cytosponge (for example), so will be updated in future models where possible[32]. Even though it may not be possible to fully account for changes in baseline survival over time the risk scores will give a rank ordering of patients that can be used for risk stratification/identification of high-risk groups.

6.8.8 Development of risk categories

Since there is no currently accepted threshold for classifying high risk of oesophageal cancer, we will examine the distribution of predicted risks and calculate a series of centile values. For each centile threshold, we will calculate the sensitivity of the risk scores.

6.8.9 Sample size

Sample size calculations for a risk prediction model aim to ensure precise estimation of model parameters whilst minimising potential overfitting. We have used the criteria of Riley et al.[47] o derive a minimum sample size of 312,616 men corresponding to 543,952 person-years of follow-up. The number of outcome events needed are 349 assuming up to 70 predictors, an event rate of 0.00064, mean follow up of 1.74 years; timepoint 10 years, a R^2 value of 0.002013. Similarly, a minimum sample size of 665,050 women corresponding to 1,157,1871 person-years of follow-up. The number of outcome events needed are 359 assuming up to 70 predictors, an event rate of 0.00031; mean follow up of 1.74 years; a R^2 value of 0.0009468. Hence a minimum sample of just over 1 million men and women would be needed in the derivation dataset.

Collins, et al[48] suggests externally validating a prognostic model requires at least 100 events and ideally, at least 200.

With over 35 million patients and at least 18,000 incident cases of oesophageal cancer on QResearch, we can confirm we have more than ample data both in the training and validation datasets. We will use all the relevant patients on the database to maximise the power and generalisability of the results. We will use STATA (version 16) for analyses. We will adhere to the TRIPOD statement for reporting[49].

6.8.10 Public and patient involvement.

Patients will be involved in setting the research question, the outcome measures, the design, implementation and dissemination of the study findings. Patient representatives will also advise on dissemination including the use of culturally appropriate lay summaries describing the research and its results.

6.9 Methodological considerations

6.9.1 Strengths

The methods to derive and validate these models are broadly the same as for a range of other widely used clinical risk prediction tools derived from the QResearch database [18-22]. The strengths and limitations of the approach have already been discussed in detail [19 22 32 33 50 51]. Key strengths include size, wealth of data on risk factors, good ascertainment of outcomes through multiple record linkage, prospective recording of outcomes, use of an established validated database which has been used to develop many risk prediction tools, and lack of selection, recall and respondent bias and robust analysis. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and prescribed medications [52]. We think our study has good face validity since it will be conducted in the setting where most patients in the UK are assessed, treated and followed up.

6.9.2 Limitations

Limitations of our study include the lack of formal adjudication of diagnoses, potential for misclassification of outcomes depending on testing, information bias, and potential for bias due to missing data. However, our database has linked cancer registry, death registry mortality and hospital admissions data and is therefore likely to have picked up the great majority of cases and death thereby minimising ascertainment bias.

The initial evaluation will be done on a separate set of practices and individuals to those which were used to develop the score although the practices all use the same GP clinical computer system (EMIS – the computer system used by 55% of UK GPs). An independent evaluation will be a more stringent test and should be done (e.g. using data from different clinical systems

or the other countries within the UK), but when such independent studies have examined other risk equations, [50 51 53 54] they have demonstrated similar performance compared with the validation in the QResearch database[18 19 32]. Whilst our study population is from England and is representative of the English population, it will need to be locally evaluated if used outside of England.

6.9.3 Regulatory and ethical challenges

We will mine a variety of databases to identify at-risk individuals (including electronic health records and endoscopy and public databases). Through collaborating with Alison Hall at PHG, we will explore the regulatory and ethical challenges associated with accessing and processing patient data for risk stratification and personalised prevention. This work will include consideration of the policy landscape for utilising data mining using conventional methods or AI/ML for risk stratification and personalised prevention; issues relating to the nature and quality of the data (to the extent that they impact on regulatory factors such as bias and discrimination); the reasonable expectations of key stakeholders for data processing (including patients and health providers); and assessing the potential legal and regulatory challenges associated with compliance with the GDPR, particularly if data mining is solely automated. This will include consideration of the requirements for information provision, transparency, and explanation (GDPR Articles 5, 13-15 and Article 22 of the GDPR).

6.9.4 Implementation intentions (to be finalised)

Any new intellectual property or improvements generated in this project will be owned by the University of Oxford and handled according to the University statutes, the terms of QResearch and the terms of the INNOVATE UK grant. It will be made available through publication in peer reviewed journals and via Oxford University Innovations as appropriate.

At timing of writing and subject to agreement, then it is envisaged that a web-based program will implement the new risk algorithm. In a similar manner for QCovid (www.qcovid.org), a range of alternative communication formats are possible and will be empirically evaluated. These could include a full risk-score, a risk-categorisation, alternative representations of relative and absolute risk for the appropriate risk-category. Multiple versions of the tool will be available to allow the user to directly enter information (for example, via an “app”) as well as versions which allow pre-population of existing data via electronic health care record systems. In all implementations it will be made clear that the risks being communicated are not specifically ‘your’ risks: they are essentially what we have observed based on medical records in a group of people with the same risk factors.

Oxford and Oxford University Innovation Ltd (OUI) can develop and support the implementation and UK deployment of resulting risk prediction tools. This process will be overseen by Oxford University Innovation. OUI will licence and deliver the risk engine packaged in a suitable format to relevant parties and provide support for integration with

their data platforms. The algorithm and associated SDK will be classified as Class 1 medical device. OUI will carry out all the necessary Quality Assurance and Regulatory activities to support the registration with the MHRA, including all code development being carried out under an appropriate Quality Management System. The resulting technical pack created for this will be provided to licensees to support their own MHRA submission to cover the tool as it is integrated into their own platforms and data systems.

OUI will also provide and host a reference website on behalf of Oxford (based on an identical implementation of the risk engine) for collaboration and demonstration purposes and to support validation of local deployments of the algorithm.

Access to the algorithm will be via the reference website protected an academic license. This will allow transparency whilst retaining some necessary controls to ensure that there is just one version of the algorithm and to prevent unregulated clinical use.

7 Other information

Acknowledgements: We acknowledge the contribution of EMIS (Egton Medical Information Systems) practices who contribute to the QResearch database and EMIS, and the Universities of Nottingham and Oxford for expertise in establishing, developing, and supporting the QResearch database. QResearch acknowledges funding from the NIHR funded Biomedical Research Centre. the cancer registration data used in these analyses were supplied by Public Health England. The Hospital Episode Statistics data and Civil registration date used in this analysis are re-used by permission from the NHS Digital who retain the copyright.

Funding: QResearch receives infrastructure support from NIHR Oxford Biomedical Research Centre; the Medical Sciences Division of the University of Oxford; John Fell Oxford University Press Research Fund; the Oxford Wellcome Institutional Strategic Support Fund (204826/Z/16/Z); Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK Oxford Centre. It also receives contributions in kind from EMIS Health (commercial supplier of NHS health computer systems).

Ethical approval: The protocol has been reviewed in accordance with the requirements for the East Midlands Derby research ethic committee (ref 18/EM/0400) and approved on 02/11/2021 (reference OX39).

Declaration: The lead author affirms that the manuscript resulting from this work will be an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted and that any discrepancies from the study as planned have been explained.

Competing Interests: JHC reports grants from National Institute for Health Research Biomedical Research Centre, Oxford, grants from John Fell Oxford University Press Research Fund, grants from Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK Oxford Centre, grants from the Oxford Wellcome Institutional Strategic

Support Fund (204826/Z/16/Z), during the conduct of the study. JHC is an unpaid director of QResearch, a not-for-profit organisation which is a partnership between the University of Oxford and EMIS Health who supply the QResearch database used for this work. JHC is a founder and shareholder of ClinRisk ltd and was its medical director until 31st May 2019. ClinRisk Ltd produces open and closed source software to implement clinical risk algorithms into clinical computer systems including the QCancer tools which will be further developed in this application. RCF developed and patented the Cytosponge-TFF3 test which has been licensed by the Medical Research Council to Covidien (now Medtronic). RCF is a co-founder and consultant for an early diagnostics company Cyted Ltd. PST reports consultation with Astra Zenenca and Duke-NUS. WM has no conflict of interest.

Table 1 Baseline characteristics of patients in the derivation and validation cohorts aged 25-84 years

	Derivation men (%)	Derivation women (%)	Validation men (%)	Validation women (%)
total				
25-49 years				
50-59 years				
60-69 years				
70-79 years				
80+ years				
Mean age (SD)				
mean Townsend score (SD)				
Body mass index recorded				
mean BMI (SD)				
ethnicity recorded				
White				
Indian				
Pakistani				
Bangladeshi				
Other Asian				
Caribbean				
Black African				
Chinese				
Other				
Ethnicity not recorded				
smoking recorded				
non-smoker				
ex-smoker				
light smoker (1-9 cigarettes/day)				
moderate smoker (10-19 cigarettes /day)				
heavy smoker (20+ cigarettes /day)				
Smoking not recorded				
alcohol recorded				
Non-drinker				
Trivial drinker (<1 unit/day)				
Light drinker (1-2 units/day)				
Moderate or heavy drinker (3+units/day)				

Alcohol not recorded				
Family history of bowel cancer				
Prior diagnosis of cancer				
Prior bowel cancer				
Prior pancreatic cancer				
Prior lung cancer				
Prior gastric cancer				
Prior renal cancer				
Prior blood cancer				
Prior oral cancer				
Prior brain cancer				
Prior breast cancer				
Prior uterine cancer				
Prior ovarian cancer				
Prior cervical cancer				
Prior prostate cancer				
Co-morbidities				
Type 1 diabetes				
Type 2 diabetes				
Barrett's oesophagus				
Peptic ulcer disease				
Ulcerative colitis				
chronic pancreatitis				
Manic depression/schizophrenia				
etc				
Prescribed medication				

Table 2: Numbers of incident cases, age standardised incidence rates per 10,000 person years in the derivation cohort in men aged 25-84 years.

	Cases on GP record			**Cases on either GP or linked mortality record**		**Cases on either GP, linked mortality or hospital record**		**cases on either GP, linked hospital, mortality or cancer record**	
Cancer Type	**Cases**	**row % of total**	**Age standardised rate per 10000**	**Cases**	**Age standardised rate per 10000**	**Cases**	**Age standardised rate per 10000**	**Total Cases**	**Age standardised rate per 10000**
Men									
Women									
total									

*patients with existing diagnoses of each cancer at baseline were dropped from the relevant cohort.
Rates were age standardised to the overall QResearch population in 5-year age bands.

8 Version History

Version	date	Issued by	Notes
1.0	07.09.2020	JHC	First issue to WM and PST
1.1	17.10.2020	JHC	Updated to include comments from WM and PST. added sample size, amended age range.
1.2	24.10.2020	JHC	Addition of sample size
1.3	25.11.2021	JHC	Update to include additional confounders/investigation
1.4	06.02.2021	JHC	Updated to correct typos, formatting
1.5	11.02.2021	JHC	Updated with CC comments

9 References

1. Fitzgerald RC, di Pietro M, O'Donovan M, et al. Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial. *The Lancet* 2020;396(10247):333-44. doi: https://doi.org/10.1016/S0140-6736(20)31099-0
2. Pennathur A, Gibson MK, Jobe BA, et al. Oesophageal carcinoma. *Lancet* 2013;381(9864):400-12. doi: 10.1016/s0140-6736(12)60643-6 [published Online First: 2013/02/05]
3. Kinoshita Y, Ishimura N, Ishihara S. Advantages and Disadvantages of Long-term Proton Pump Inhibitor Use. *J Neurogastroenterol Motil* 2018;24(2):182-96. doi: 10.5056/jnm18001
4. Lambert AA, Lam JO, Paik JJ, et al. Risk of community-acquired pneumonia with outpatient proton-pump inhibitor therapy: a systematic review and meta-analysis. *PLoS One* 2015;10(6):e0128004.
5. Hippisley-Cox J, Coupland C. Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* 2011;61(592):e707-14. doi: 10.3399/bjgp11X606609 [published Online First: 2011/11/08]
6. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *The British journal of general practice : the journal of the Royal College of General Practitioners* 2011;61(592):e715-23. doi: 10.3399/bjgp11X606627 [published Online First: 2011/11/08]
7. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ* 2012;344 doi: 10.1136/bmj.d8009
8. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice* 2012;62(594):e29-e37. doi: 10.3399/bjgp12X616346
9. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. *The British journal of general practice : the journal of the Royal College of General Practitioners* 2012;62(594):e38-e45. doi: 10.3399/bjgp12X616355

10. Hippisley-Cox J, Coupland C. Identifying patients with suspected renal tract cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2012;62(597):e251-60. doi: 10.3399/bjgp12X636074
11. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2013;63(606):11-21. doi: 10.3399/bjgp13X660733
12. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2013;63(606):1-10. doi: 10.3399/bjgp13X660724
13. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015;5(3):e007825. doi: 10.1136/bmjopen-2015-007825
14. Qiao Y, Yang T, Gan Y, et al. Associations between aspirin use and the risk of cancers: a meta-analysis of observational studies. *BMC Cancer* 2018;18(1):288. doi: 10.1186/s12885-018-4156-5 [published Online First: 2018/03/15]
15. Snider EJ, Kaz AM, Inadomi JM, et al. Chemoprevention of esophageal adenocarcinoma. *Gastroenterol Rep (Oxf)* 2020;8(4):253-60. doi: 10.1093/gastro/goaa040
16. Jankowski JAZ, de Caestecker J, Love SB, et al. Esomeprazole and aspirin in Barrett's oesophagus (AspECT): a randomised factorial trial. *Lancet* 2018;392(10145):400-08.
17. Beales IL, Hensley A, Loke Y. Reduced esophageal cancer incidence in statin users, particularly with cyclo-oxygenase inhibition. *World J Gastrointest Pharmacol Ther* 2013;4(3):69-79. doi: 10.4292/wjgpt.v4.i3.69 [published Online First: 2013/08/07]
18. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008:bmj.39609.449676.25. doi: 10.1136/bmj.39609.449676.25
19. Hippisley-Cox J, Coupland C, Robson J, et al. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880-. doi: 10.1136/bmj.b880
20. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012;344(may22 1):e3427-e27. doi: 10.1136/bmj.e3427
21. Hippisley-Cox J, Coupland C. Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney(R) Scores. *BMC Family Practice* 2010;11:49.
22. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. *BMJ* 2011;343:d4656. doi: 10.1136/bmj.d4656 [published Online First: 2011/08/19]
23. Clift AK, Coupland CAC, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020;371:m3731. doi: 10.1136/bmj.m3731 [published Online First: 2020/10/22]
24. Coupland CAC, Hill T, Dening T, et al. Anticholinergic Drug Exposure and the Risk of Dementia: A Nested Case-Control Study. *JAMA Intern Med* 2019 doi: 10.1001/jamainternmed.2019.0677 [published Online First: 2019/06/25]
25. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of hormone replacement therapy and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. *BMJ* 2019;364:k4810. doi: 10.1136/bmj.k4810 [published Online First: 2019/01/11]

26. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of combined oral contraceptives and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. *BMJ* 2015;350:h2135. doi: 10.1136/bmj.h2135
27. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018;8(2) doi: 10.1136/bmjopen-2017-020738
28. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013;3(8):e003482. doi: 10.1136/bmjopen-2013-003482 [published Online First: 2013/08/21]
29. Hippisley-Cox J. Validity and completeness of the NHS Number in primary and secondary care electronic data in England 1991-2013. 2013; 1. Hippisley-Cox J. Validity and completeness of the NHS number in primary and secondary care: electronic data in England 1991-2013 http://eprints.nottingham.ac.uk/3153/1/Validity%26CompletenessNHSNumber.pdf (accessed June 2013).
30. Townsend P, Davidson N. The Black report. London: Penguin 1982.
31. Thomas T, Loke Y, Beales ILP. Systematic Review and Meta-analysis: Use of Statins Is Associated with a Reduced Incidence of Oesophageal Adenocarcinoma. *J Gastrointest Cancer* 2018;49(4):442-54. doi: 10.1007/s12029-017-9983-0
32. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 2009;339:b4229. doi: 10.1136/bmj.b4229
33. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34-39. doi: 10.1136/hrt.2007.134890
34. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099. doi: 10.1136/bmj.j2099
35. Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ* 2017;359:j5019. doi: 10.1136/bmj.j5019
36. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychological Methods* 2002;7:147-77.
37. Group TAM. Academic Medicine: problems and solutions. *BMJ* 1989;298:573-79.
38. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007;60:979.
39. Moons KGM, Donders RART, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006;59:1092.
40. Rubin DB. Multiple Imputation for Non-response in Surveys. New York: John Wiley 1987.
41. Hosmer D, Lemeshow S. Applied Logistic Regressopm. New York: John Wiley & Sons, Inc. 1989.
42. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* 2014;4(8):e005809. doi: 10.1136/bmjopen-2014-005809
43. Royston P. Explained variation for survival models. *Stata J* 2006;6:1-14.

44. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48.
45. Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361 - 87.
46. Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. *Stata Journal* 2010;10(3):339-58.
47. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441. doi: 10.1136/bmj.m441 [published Online First: 2020/03/20]
48. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2015 doi: 10.1002/sim.6787 [published Online First: 2015/11/11]
49. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. *Annals of Internal Medicine* 2015;162(1):55-63. doi: 10.7326/M14-0697
50. Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 2011;342:d3651.
51. Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing Type 2 diabetes. *Diabetic Medicine* 2011;28:599-607. doi: 10.1111/j.1464-5491.2011.03237.x
52. Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. *Health stat* 2004(21):5-14.
53. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012;344:e4181. doi: 10.1136/bmj.e4181
54. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;340:c2442. doi: 10.1136/bmj.c2442

AUTHOR: julia.hippisley-cox@phc.ox.ac.uk
DATE: 04.09.2020
VERSION: 1.0