Julia Hippisley-Cox and Carol Coupland

Identifying patients with suspected pancreatic cancer in primary care:

derivation and validation of an algorithm

Abstract

Background
Pancreatic cancer has the worst survival for any cancer and is often diagnosed late when the cancer is advanced. Chances of survival are more likely if patients can be diagnosed earlier.

Aim
To derive and validate an algorithm to estimate absolute risk of having pancreatic cancer in patients with and without symptoms in primary care.

Design and setting
Cohort study using data from 375 UK QResearch® general practices for development and 189 for validation.

Method
Included patients were aged 30–84 years, free at baseline from a diagnosis of pancreatic cancer and had not had dysphagia, abdominal pain, abdominal distension, appetite loss, or weight loss recorded in the preceding 12 months. The primary outcome was incident diagnosis of pancreatic cancer recorded in the following 2 years. Risk factors examined included: age, body mass index, smoking status, alcohol, deprivation, diabetes, pancreatitis, previous diagnosis of cancer apart from pancreatic cancer, dysphagia, abdominal pain, abdominal distension, appetite loss, weight loss, diarrhoea, constipation, tiredness, itching, and anaemia. Cox proportional hazards models were used to develop separate risk equations in males and females. Measures of calibration and discrimination assessed performance in the validation cohort.

Results
There were a total of 1415 incident cases of pancreatic cancer from 4.1 million person-years in the derivation cohort. Independent predictors in both males and females were age, smoking, type 2 diabetes, chronic pancreatitis, abdominal pain, appetite loss, and weight loss. Abdominal distension was a predictor for females only; dysphagia and constipation were predictors for males only. On validation, the algorithms explained 59% of the variation in females and 62% in males. The receiver operating characteristic statistics were 0.84 (females) and 0.87 (males). The D statistic was 2.44 (females) and 2.61 (males). The 10% of patients with the highest predicted risks contained 62% of all pancreatic cancers diagnosed over the following 2 years.

Conclusion
The algorithm has good discrimination and calibration and could potentially be used to help identify those at highest risk of pancreatic cancer to facilitate early referral and investigation.

Keywords
diagnosis; pancreatic cancer; primary care; qresearch; risk prediction; symptoms.

INTRODUCTION

Pancreatic cancer is the 11th most common cancer in the UK.[1] Diagnoses are often made late when the cancer is advanced.[2] Less than 20% of patients are suitable for surgery and 84% of patients are likely to have died within a year of diagnosis,[1] giving the worst survival rate for any cancer.[2] However, the chance of survival is more likely if patients present at an early stage.[1]

There are only a few established risk factors for pancreatic cancer such as age,[2] smoking,[2,3] genetic factors,[2] chronic pancreatitis,[1] and alcohol.[1,4] Diabetes may be a risk factor for pancreatic cancer or an early manifestation of a growing tumour.[5,6] As there are few established risk factors and currently no reliable screening test, it is unlikely that there will be a national screening programme for pancreatic cancer; as such, it is likely that most pancreatic cancers will be diagnosed in patients who are symptomatic and presenting to primary care. The challenge is ensuring earlier diagnosis to help improve treatment options (for example, possibility of surgery) and prognosis. Earlier diagnosis could be helped by increased public awareness of symptoms that might indicate pancreatic cancer, such as weight loss, loss of appetite, and abdominal pain.[3,6,7] Diagnosis could also be improved by more prompt investigation of patients who are symptomatic and present to their GP.[6] In the UK, GPs will soon have better direct access to diagnostic investigations such as ultrasound, computerised tomography (CT) scanning, and magnetic resonance imaging (MRI) but they need better assessment tools to quantify a patient's risk of different types of cancer and thereby ensure the right patients are sent for the right investigations. This would also make efficient use of scare resources.[6]

Surprisingly, pancreatic cancer is not mentioned in current guidelines from the National Institute for Health and Clinical Excellence (NICE) on the referral of patients with suspected cancer.[8] This might be because relatively little is known about the aetiology of pancreatic cancer or its attendant presenting features. Therefore, the pattern of presenting symptoms for patients with pancreatic cancer was investigated with a view to developing an algorithm to quantify the risk of a patient having pancreatic cancer; this algorithm would incorporate both symptoms and baseline risk factors such as age, chronic pancreatitis, and smoking.

QResearch® primary care database was used to develop the risk prediction model as it contains robust data on many of the relevant exposures and outcomes. It is also representative of the population where such a model is likely to be used and has been used successfully to develop and validate a range of prediction models for use in primary care.[9–12] Once validated, the models could be integrated into clinical computer systems to help systematically identify those at high risk and alert clinicians to those who might benefit most from further assessment or interventions.[9–12] The algorithm could also be made available on

J Hippisley-Cox, MD, FRCGP, MRCP, professor of clinical epidemiology and general practice; **C Coupland**, PhD, associate professor in medical statistics, Division of Primary Care, University of Nottingham.

Address for correspondence
Julia Hippisley-Cox, Division of Primary Care, 13th floor, Tower Building, University Park, Nottingham, NG2 7RD.

E-mail: julia.hippisley-cox@nottingham.ac.uk

Submitted: 21 July 2011; **Editor's response:** 4 August 2011; **final acceptance:** 30 August 2011.

This is the full-length article (published online 27 Dec 2011) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2012; DOI: 10.3399/bjgp12X616355**

How this fits in

Pancreatic cancer is often diagnosed late when the cancer is advanced and the chance of survival is, therefore, poor. There is no reliable screening test so most diagnoses are likely to be made in patients who are symptomatic; as such, in order to make diagnoses earlier, there needs to be an increased awareness of symptoms among patients and earlier investigation of patients who are symptomatic by GPs. This study has developed a new algorithm which predicts the chances of having pancreatic cancer based on a combination of symptoms and baseline risk factors such as age, chronic pancreatitis, smoking, and diabetes. The algorithm performed well in an independent sample, both in terms of discrimination and calibration. The sensitivity was high — for example, if a threshold equivalent to the top 10% of patients with the highest risk is selected, this will account for 62% of all cases of pancreatic cancers occurring within the subsequent 2 years.

the internet as a simple calculator for use by the general population to help support the National Early Diagnosis and Awareness Initiative,[7] which aims to raise public awareness of the signs and symptoms of cancer, and encourage those who may have symptoms to seek advice earlier.

METHOD

Study design and data source

A prospective cohort study was carried out in a large population of primary care patients from an open cohort study using the QResearch database (version 30). All practices in England and Wales who had been using their EMIS (Egton Medical Information Systems) computer system for at least a year were included. Two-thirds of practices were randomly allocated to the derivation dataset and the remaining one-third to a validation dataset. An open cohort of patients aged 30–84 years was identified, drawn from patients registered with practices between 1 January 2000 and 30 September 2010. The following were excluded: patients without a postcode-related Townsend score, those with a history of pancreatic cancer at baseline, and those with a recorded red flag symptom[13] in the 12 months prior to the study entry date. For this study, a red-flag symptom was defined as one that might alarm the patient and indicate the presence of pancreatic cancer, that is, symptoms of dysphagia, loss of appetite, weight loss, abdominal pain, or abdominal distension. Jaundice was not included as this is relatively rare, usually considered a sign, and would have its own pathway for investigation.

Patients entered the cohort on the latest of the study start date (1 January 2000) and 12 months after the patient registered with the practice; this ensured that all patients had a minimum of 12 months registration prior to study entry. For patients with incident dysphagia, appetite loss, weight loss, abdominal pain, or abdominal distension, the entry date was the date of first recorded onset within the study period.

Clinical outcome definition

The study outcome was pancreatic cancer, which was defined as incident diagnosis of pancreatic cancer during the 2 years following the entry date recorded on either the patient's GP record using the relevant UK diagnostic Read Codes, or their linked cause of death record, according to the Office for National Statistics (ONS), using the relevant diagnostic codes from the International Classification of Diseases-9 (ICD-9) (157) or ICD-10 (C25).

A 2-year period was used as this represents the period of time during which existing cancers are likely to become clinically manifest.[13] Patients were censored at the earliest date of either diagnosis of pancreatic cancer, date of death, date of leaving the practice, or 2 years after their study entry date.

Predictor variables

Established predictor variables were examined, focusing on those that are likely to be recorded in the patient's electronic record and that the patient is likely to know.[6] Symptoms that might herald a diagnosis of pancreatic cancer were also included.[1,6] Separate analyses were carried out in males and females. The following predictor variables were examined, using information recorded prior to study entry:

- currently consulting GP with first onset of dysphagia (yes/no);
- currently consulting GP with first onset of loss of appetite (yes/no);
- currently consulting GP with first onset of weight loss symptom (yes/no);
- currently consulting GP with first onset of abdominal pain (yes/no);
- currently consulting GP with first onset of abdominal distension (yes/no);
- recently consulted a GP with first onset of any of:

- constipation in past 12 months (yes/no);
- diarrhoea in past 12 months (yes/no);
- tiredness in past 12 months (yes/no);
- itching in past 12 months (yes/no);

- age at baseline (continuous, ranging from 30 to 84) years;
- body mass index (continuous);
- smoking status (non smoker; ex; light (1–9 cigarettes/day); moderate (10–19 cigarettes/day); heavy smoker (≥20 cigarettes/day);[2]
- alcohol status (non-drinker; trivial (<1 unit/day); light (1–2 units/day); moderate/heavy (≥3 units/day));[4]

Table 1. Baseline characteristics of patients in the derivation and validation cohorts. Figures are *n* (%) unless otherwise specified

Characteristic	Derivation cohort, *n* = 2 364 571	Validation cohort, *n* = 1 243 740
Females	1 178 682 (49.8)	619 388 (49.8)
Males	1 185 889 (50.2)	624 352 (50.2)
Mean age, years (SD)	50.1 (15.0)	50.1 (14.9)
Mean Townsend score (SD)[a]	−0.3 (3.4)	−0.2 (3.6)
BMI recorded prior to study entry	1 877 243 (79.4)	1 009 931 (81.2)
Mean BMI (SD)	26.4 (4.6)	26.4 (4.7)
Smoking status		
Non-smoker	1 200 385 (50.8)	627 868 (50.5)
Ex-smoker	426 697 (18.0)	228 970 (18.4)
Current smoker, amount not recorded	71 668 (3.0)	39 438 (3.2)
Light smoker (<10/day)	149 044 (6.3)	80 402 (6.5)
Moderate smoker (10–19/day)	180 887 (7.6)	96 443 (7.8)
Heavy smoker (≥20/day)	135 113 (5.7)	74 140 (6.0)
Smoking status not recorded	200 777 (8.5)	96 479 (7.8)
Alcohol intake		
None	512 816 (21.7)	276 449 (22.2)
Trivial (<1 unit/day)	660 737 (27.9)	358 233 (28.8)
Light (1–2 units/day)	495 561 (21.0)	258 963 (20.8)
Moderate or heavy (≥3 units/day)	177 129 (7.5)	93 705 (7.5)
Alcohol intake not recorded	518 328 (21.9)	256 390 (20.6)
Medical history		
Type 1 diabetes	7269 (0.3)	3986 (0.3)
Type 2 diabetes	78 687 (3.3)	41 869 (3.4)
Prior acute pancreatitis	5029 (0.2)	2707 (0.2)
Prior chronic pancreatitis	2208 (0.1)	1206 (0.1)
Prior cancer apart from pancreatic cancer	54 018 (2.3)	28 578 (2.3)
Symptoms		
Current dysphagia	15 648 (0.7)	8507 (0.7)
Current abdominal pain	232 586 (9.8)	129 924 (10.4)
Current abdominal distension	7985 (0.3)	4929 (0.4)
Current appetite loss	10 351 (0.4)	5567 (0.4)
Current weight loss	26 239 (1.1)	14 686 (1.2)
Constipation in preceding year	15 094 (0.6)	8476 (0.7)
Diarrhoea in preceding year	22 377 (0.9)	12 233 (1.0)
Tiredness in preceding year	22 674 (1.0)	12 688 (1.0)
Itching in preceding year	2615 (0.1)	1454 (0.1)
Haemoglobin recorded in preceding year	398 059 (16.8)	214 497 (17.2)
Haemoglobin <11 g/dl in preceding year	29 808 (1.3)	16 172 (1.3)

[a]Townsend score is a deprivation score derived from patients' postcodes, which ranges between −6 (most affluent) and +11 (most deprived). BMI = body mass index; SD = standard deviation. Patients are free of a diagnosis of pancreatic cancer at baseline.

- Townsend deprivation score, derived from patients' postcodes (continuous);
- diabetes (Type1/Type2/no diabetes) at study entry;[5]
- pancreatitis (acute/chronic/none) at study entry;
- previous diagnosis of cancer apart from pancreatic cancer at study entry; and
- anaemia, defined as recorded haemoglobin <11 g/dl in 12 months before study entry or the 60 days after (yes/no)

Derivation and validation of the models

The risk-prediction algorithm was developed and validated using established methods.[9–12,14–16] Multiple imputation was used to replace missing values for body mass index, alcohol intake, and smoking status and these values were used in the main analyses.[17–20] Five imputations were carried out. Cox's proportional hazards models were used to estimate the coefficients for each risk factor for males and females separately, using robust variance estimates to allow for the clustering of patients within general practices. Rubin's rules were used to combine the results across the imputed datasets.[21] Fractional polynomials were used to model non-linear risk relationships with continuous variables.[22] A full model was fitted initially and variables were retained if they had a hazard ratio of <0.80 or >1.20 (for binary variables) and were statistically significant at the 0.01 level. Interactions between predictor variables and age were examined and included in the final models if they were statistically significant at the 0.01 level.

The regression coefficients for each variable from the final model were used as weights, which were combined with the baseline survivor function evaluated at 2 years, to derive absolute risk equations for 2 years of follow-up.[23] The baseline survivor function was estimated based on zero values of centred continuous variables, with all binary predictor values set to zero, using the methods implemented in Stata.

Multiple imputation was used in the validation cohort to replace missing values for body mass index, alcohol intake, and smoking. The risk equations for males and females obtained from the derivation cohort were then applied to the validation cohort and measures of discrimination calculated. R^2 (estimated variation explained by the risk equation in time to pancreatic cancer[24]), the D statistic[25] (a measure of discrimination

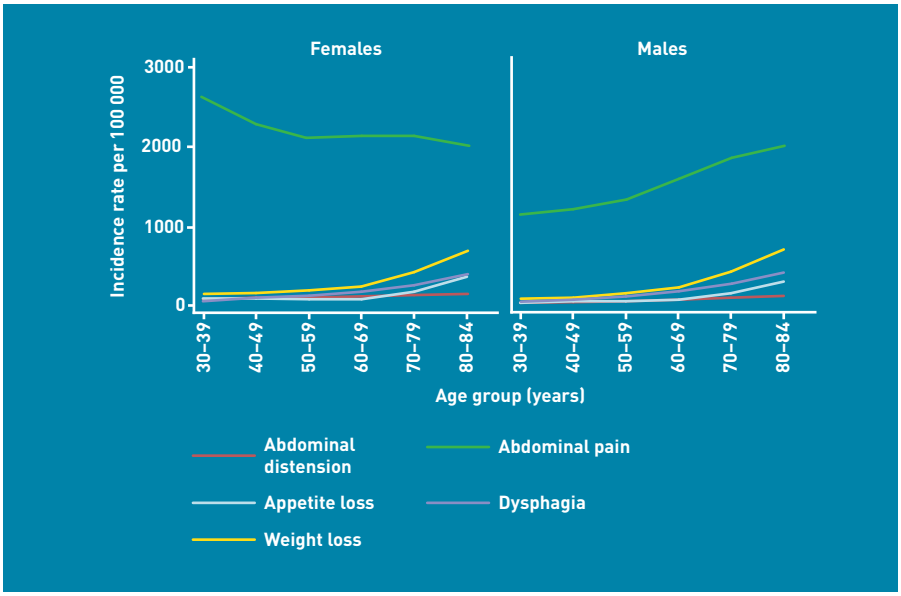

Figure 1. Incidence rates of dsyphagia, abdominal pain, abdominal distension, appetite loss, and weight loss per 100 000 person years in males and females in the derivation cohort.

suitable for use with survival data where higher values indicate better discrimination), and the area under the receiver operating characteristic (ROC) curve (ROC curve statistic) at 2 years were calculated. Calibration was assessed by comparing the mean predicted risks at 2 years with the observed risk by tenth of predicted risk. The observed risk was obtained using the Kaplan–Meier estimate evaluated at 2 years.

The validation cohort was used to define the thresholds for the 0.1%, 0.5%, 1%, 5%, and 10% of patients at highest estimated risk of pancreatic cancer at 2 years. Sensitivity, specificity, and positive and negative predictive values were calculated using these thresholds, restricting the analyses to patients who had the outcome within 2 years or had at least 2 years of follow-up. All available data on the database were used to maximise the power and generalisability of the results. Stata (version 11) was used for all analyses.

Table 2. Adjusted hazard ratios (95% CI) for the final model for pancreatic cancer for males and females in the derivation cohort

	Adjusted hazard ratios for females (95% CI)	Adjusted hazard ratios for males (95% CI)
Non-smoker	1	1
Ex-smoker	0.97 (0.77 to 1.23)	1.37 (1.12 to 1.67)
Light smoker	1.53 (1.04 to 2.25)	1.44 (1.03 to 2.03)
Moderate smoker	2.32 (1.74 to 3.10)	1.63 (1.20 to 2.20)
Heavy smoker	2.39 (1.65 to 3.48)	1.88 (1.36 to 2.61)
Medical history		
Type 2 diabetes	2.07 (1.66 to 2.58)	2.11 (1.76 to 2.52)
Chronic pancreatitis	3.15 (1.17 to 8.46)	3.94 (1.93 to 8.01)
Current symptoms and symptoms in preceding year		
Current appetite loss[a]	3.90 (2.61 to 5.82)	2.46 (1.43 to 4.23)
Current weight loss[a]	3.27 (2.35 to 4.56)	12.5 (7.84 to 19.9)[b]
Current abdominal pain[a]	4.09 (3.46 to 4.84)	5.23 (4.48 to 6.11)
Current abdominal distension[a]	3.04 (1.68 to 5.50)	NS
Current dysphagia[a]	NS	2.56 (1.60 to 4.10)
Constipation in last year[a]	NS	1.91 (1.35 to 2.71)

[a]Compared with person without this characteristic. [b]Interaction term, at mean age in males. The models also included fractional polynomial terms for age, which were age^{-2} and age^{3} for females and age^{-1} for males. The model for males also included an interaction between weight loss and the age term. Hazard ratios adjusted for all other terms in the table and for age. NS = not significant.

RESULTS

Overall study population

Overall, 564 QResearch practices in England and Wales met the inclusion criteria, of which 375 were randomly assigned to the derivation dataset, with the remainder assigned to a validation cohort. A total of 2 538 615 patients aged 30–84 years were identified in the derivation cohort, and 124 458 (4.9%) patients without a recorded Townsend deprivation score were excluded; 161 (0.01%) patients with a history of pancreatic cancer were also excluded, as well as a further 49 425 (1.9%) patients with at least one red-flag symptom recorded in the 12 months prior to entry to the study, leaving 2 364 571 patients for analysis.

A total of 1 342 329 patients aged 30–84 years were identified in the validation cohort; of these, 70 847 (5.3%) patients without a recorded Townsend score were excluded, as well as 96 (0.01%) with a history of pancreatic cancer, and 27 646 (2.1%) with at least one red-flag symptom recorded in the 12 months prior to study entry, leaving 1 243 740 patients for analysis.

The baseline characteristics of each cohort were very similar, as shown in Table 1. As in another study,[10] the patterns of missing data supported the use of multiple imputation to replace missing values for smoking status, alcohol intake, and body mass index (not shown, available from the authors).

Incidence rates for red-flag symptoms

Overall, in the derivation cohort, 15 648 patients with dysphagia were identified, 10 351 with appetite loss, 26 239 with weight loss, 232 586 with abdominal pain, and 7985 with abdominal distension. Figure 1 shows the age–sex incidence rates of each symptom. The incidence rates for abdominal distension, dysphagia, appetite loss, and weight loss were similar in males and females and increased steeply with age. Abdominal pain was more common in females; it tended to decrease with age in females but increase with age in males.

Incidence rates of pancreatic cancer

Overall in the derivation cohort, during the 2-year follow-up period, a total of 1415 incident cases of pancreatic cancer arising from

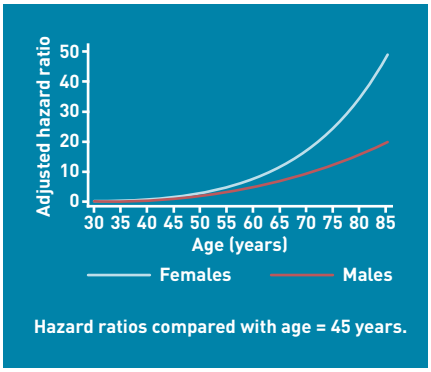

Figure 2. Adjusted hazard ratios for pancreatic cancer by age in males and females.

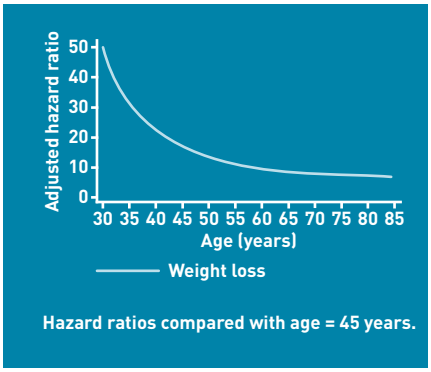

Figure 3. Adjusted hazard ratio for weight loss by age in males.

Table 3. Validation statistics for the risk prediction algorithm in the validation cohort

	Mean (95% CI)
Females	
R^2 statistic[a] (%)	58.7 (55.4 to 61.9)
D statistic[b]	2.44 (2.27 to 2.60)
ROC statistic[c]	0.84 (0.82 to 0.86)
Males	
R^2 statistic[a] (%)	62.0 (59.1 to 64.8)
D statistic[b]	2.61 (2.45 to 2.77)
ROC statistic[c]	0.87 (0.85 to 0.88)

[a]R^2 statistic shows explained variation in time to diagnosis of pancreatic cancer — higher values indicate more variation is explained. [b]D statistic is a measure of discrimination — higher values indicate better discrimination. [c]ROC statistic is a measure of discrimination — higher values indicate better discrimination.

4 149 461 person years of observation were identified, giving a crude rate of 34 cases per 100 000 person years. There were 1080 cases (76.3 % of 1415) identified using the GP record and an additional 335 (23.7%) identified from the linked death record. The incidence rates increased with age and tended to be higher in males than females (data not shown).

In the validation cohort, 781 incident cases of pancreatic cancer were identified arising from 2 184 336 person years of observation, giving a rate of 36 per 100 000 person years. There were 612 cases (78.4%) identified using the GP record and an additional 169 (21.6 %) from the linked death record.

Predictor variables

Table 2 shows the predictor variables selected for the final models for females and males. Independent predictors in both males and females were age, smoking status, type 2 diabetes, chronic pancreatitis, abdominal pain, appetite loss, weight loss. Abdominal distension was a predictor for females only; dysphagia and constipation were predictors for males only. The other variables examined were not independent risk factors so were not included in the final models.

Risk of pancreatic cancer in females was significantly associated with increasing age as shown in Figure 2. Risk also increased with the amount smoked. For example, compared with non-smokers, the risks were increased by 2.4-fold for heavy smokers and 1.5-fold for light smokers (Table 2). The risks were also elevated in females with type 2 diabetes (2.1-fold higher), chronic pancreatitis (3.2-fold higher), abdominal pain (4.1-fold higher), abdominal distension (3.0-fold higher), appetite loss (3.9-fold higher), and weight loss (3.3-fold higher) (Table 2).

The magnitudes of the hazard ratios in males were generally similar to those for females, as shown in Table 2), but the relative increase in risk with increasing age was less steep (Figure 2. The risks were also elevated with appetite loss (2.5-fold higher), abdominal pain (5.2-fold higher), dysphagia (2.6-fold higher), and constipation (1.9-fold higher). There was also a significant interaction between weight loss and age in males, as shown in Figure 3. Unlike in females, abdominal distension was not a significant predictor in males.

Validation

The validation statistics (Table 3) showed that the risk-prediction equations explained 59% of the variation in time to diagnosis in females and 62% of the variation in males. The D statistic was 2.44 for females and 2.61 for males. The area under the ROC statistics were 0.84 for females and 0.87 for males.

Figure 4 shows the mean predicted scores and the observed risks at 2 years within each tenth of predicted risk, in order to assess the calibration of the model in the validation cohort. Overall, the model was well calibrated. There was close correspondence between predicted and observed 2-year risks within each model tenth for males and females, with a small degree of over-prediction in the highest tenth in males.

Individual risk assessment and thresholds

One potential use for this algorithm is within consultations with individual patients particularly if they present with new onset of an alarm symptom such as dysphagia, abdominal pain, weight loss, or appetite loss. Some clinical examples are shown in Box 1. The results could help inform the decision to undertake further investigations such as abdominal ultrasound, MRI, CT scan, or endoscopic retrograde cholangiopancreatography.

The algorithm could also be used for systematic risk stratification for a population of patients aged 30–84 years. Software implementing the algorithm could be integrated in to GP computer systems to calculate the risk of a patient having an existing but as yet undiagnosed pancreatic cancer based on information already recorded in the patient's electronic health record. Patients at highest risk could be identified and recalled for a clinical assessment.

The 90th centile defined a high-risk group with a 2-year risk score of >0.2% (Table 4). There were 487 new cases of pancreatic cancer within this group out of 781 new cases identified in the validation cohort, which accounted for 62.4% of all new cases of pancreatic cancer (sensitivity). The positive predictive value (PPV) with this threshold was 0.6%. Alternatively, using a threshold based on the top 5% of risk (that is, a 2-year risk score >0.3%) had a sensitivity of 45.3% and a PPV of 0.9%. The sensitivity of an approach based on single symptoms ranged from 1.2% for abdominal distension to 39.8% for abdominal pain; 18.6% of cases of pancreatic cancer had a diagnosis of type 2 diabetes at baseline.

DISCUSSION

Summary

This research has developed and validated a new algorithm designed to quantify the

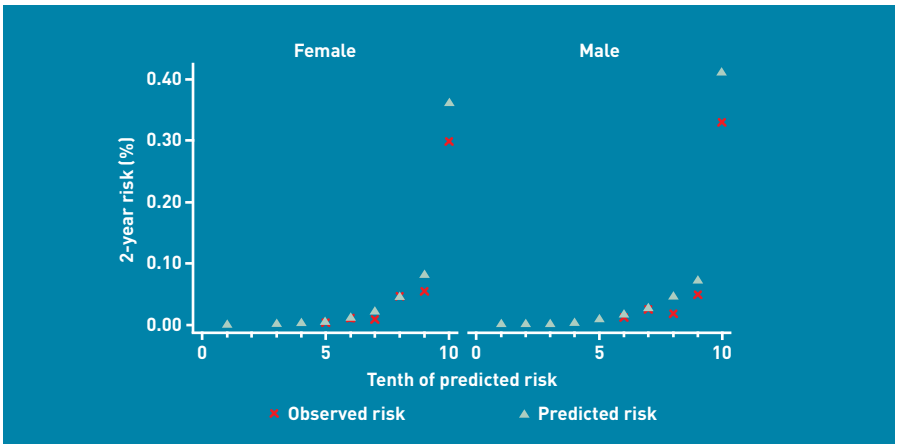

Figure 4. Mean predicted risk and observed risk of pancreatic cancer over 2 years by tenth of predicted risk applying the risk prediction scores to the validation cohort.

absolute risk of having pancreatic cancer, which is either currently present or likely to become manifest within 2 years. To the authors' knowledge, this is the first algorithm of its kind and is based on age, smoking status, alcohol status, chronic pancreatitis, type 2 diabetes, loss of appetite, weight loss, abdominal pain, abdominal distension, dysphagia, and constipation. The algorithm is based on simple clinical variables which can be ascertained in clinical practice. The algorithm performed well in a separate validation sample with good discrimination and calibration.

Strengths and limitations

Key strengths of the study include size, duration of follow-up, representativeness, and lack of selection, recall, and responder bias. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses.[26] The authors consider that the study has good face validity since, it has been conducted in the setting where the majority of patients in the UK are assessed, treated, and followed-up. The algorithm has good face validity as it confirms the significance of established risk factors such as age, smoking, chronic pancreatitis, type 2 diabetes, and associated symptoms such as weight loss, appetite loss, and abdominal pain.

Limitations of the study include lack of formally-adjudicated outcomes, information bias, and missing data. Not all patients with symptoms will attend their GP, and in those who do, not all symptoms will be reported or recorded.

The database has linked cause of death from the UK ONS and the study is therefore likely to have picked up the majority of cases of pancreatic cancer, thereby minimising ascertainment bias. The incidence rate in this study's population is close to European estimates suggesting good ascertainment of cases.[27] It is higher than rates reported in UK cancer registry reports which probably reflects the improved ascertainment resulting from including of diagnoses recorded in either the primary care or the cause of death record. Patients who die of pancreatic cancer will be included on the linked cause of death data. Patients diagnosed with pancreatic cancer in hospital will have the information recorded in hospital discharge letters which are sent to the GP and then entered into the patient's electronic record.

Comparison with existing literature

While the study was reliant on accuracy of information recorded by primary care

Table 4. Comparison of strategies to identify patients at risk of having a diagnosis of pancreatic cancer in the next 2 years based on the validation cohort

Criteria	Risk threshold %	True negative[a]	False negative[b]	False positive[c]	True positive[d]	Sensitivity, %	Specificity, %	Positive predictive value, %	Negative predictive value, %
Type 2 diabetes	n/a	940 800	636	30 125	145	18.6	96.9	0.5	99.9
Chronic pancreatitis	n/a	970 094	777	831	4	0.5	99.9	0.5	99.9
Current symptoms									
Abdominal pain	n/a	877 133	470	93 792	311	39.8	90.3	0.3	99.9
Abdominal distension	n/a	967 478	772	3447	9	1.2	99.6	0.3	99.9
Dysphagia	n/a	965 494	770	5431	11	1.4	99.4	0.2	99.9
Appetite loss	n/a	967 570	754	3355	27	3.5	99.7	0.8	99.9
Weight loss	n/a	961 571	720	9354	61	7.8	99.0	0.6	99.9
Risk threshold									
Top 10% risk	0.2	884 670	294	86 255	487	62.4	91.1	0.6	100.0
Top 5% risk	0.3	931 077	427	39 848	354	45.3	95.9	0.9	100.0
Top 1% risk	0.9	964 373	685	6552	96	12.3	99.3	1.4	99.9
Top 0.5% risk	1.3	967 852	716	3073	65	8.3	99.7	2.1	99.9
Top 0.1% risk	2.8	970 423	758	502	23	2.9	99.9	4.4	99.9

The positive predictive values (PPVs) are an average for patients in each category; PPVs for individuals can be calculated using the web calculator to take their characteristics into account. n/a = not applicable. [a]Criterion not met does not have disease. [b]Criterion not met does have disease. [c]Criterion met does not have disease. [d]Criterion met does have disease.

Box 1. Clinical examples

- A 70-year-old male with type 2 diabetes, who is a heavy smoker and presents with abdominal pain, loss of appetite, and dysphagia has a 10.2% estimated risk of having pancreatic cancer. If he also has a history of chronic pancreatitis, the estimated risk would be 34.4%.
- A 75-year-old female, who is a moderate smoker and has abdominal pain, abdominal distension, and loss of appetite, has an 11.3% estimated risk of having pancreatic cancer. If she also has weight loss, the estimated risk would be 44.1%.
- A 50-year-old male, who is a heavy smoker with type 2 diabetes and no symptoms has a 0.1% estimated risk of developing pancreatic cancer in the next 2 years. If the same male had abdominal pain, appetite loss, dysphagia, and constipation the estimated risk would be 3.9%.

Funding

This study was undertaken by ClinRisk Ltd. There was no external funding.

Ethics committee

All QResearch® studies are independently reviewed in accordance with the QResearch® agreement with Trent Multi-Centre Ethics Committee (UK).

Provenance

Freely submitted; externally peer reviewed.

Web calculator

Here is a simple web calculator to implement the QCancer® (pancreatic) algorithm, which is publically available alongside the paper and open source software http://www.qcancer.org/pancreas.

Competing interests

Julia Hippisley-Cox is professor of clinical epidemiology at the University of Nottingham and co-director of QResearch® — a not-for-profit organisation which is a joint partnership between the University of Nottingham and EMIS (leading commercial supplier of IT for 60% of general practices in the UK). Julia Hippisley-Cox is also a paid director of ClinRisk Ltd, which produces software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help improve patient care. Carol Coupland is associate professor of medical statistics at the University of Nottingham and a paid consultant statistician for ClinRisk Ltd. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations.

Acknowledgements

We acknowledge the contribution of EMIS practices who contribute to QResearch® and EMIS for expertise in establishing, developing and supporting the database. The algorithms presented in this paper will be released as Open Source Software under the GNU lesser GPL v3.

Discuss this article

Contribute and read comments about this article on the Discussion Forum: **http://www.rcgp.org.uk/bjgp-discuss**

physicians, the quality of information is likely to be good since previous studies have validated similar outcomes and exposures using questionnaire data and found levels of completeness and accuracy in similar GP databases to be good.[28,29] For example, one systematic review reported that on average 89% of diagnoses recorded on the GP electronic record are confirmed from other data sources.[28] Currently however, there is limited information on the QResearch database regarding the precise type of cancer which means it was not possible to include the precise type of pancreatic cancer in the outcome (that is, distinguish between endocrine and exocrine tumours), the stage, or the grade. The QResearch database will be linked with information from cancer registries in the near future which is likely to increase the accuracy and completeness of this information.

Implications for research and practice

This study could help raise awareness of symptom complexes predictive of pancreatic cancer especially since current NICE guidelines on suspected cancer fail to mention pancreatic cancer despite its substantial morbidity and mortality. Further research is needed to assess whether use of this symptom based tool can lead to earlier identification of pancreatic cancer at a stage where curative treatment is more likely to be possible.

Although primarily designed to be used by GPs at the point of care to assess risk in symptomatic patients, and inform the decision to investigate or refer, the algorithm could be also be used by members of the public via a simple web calculator which could then prompt symptomatic patients to consult their GP. The algorithm could be integrated into GP clinical computer systems and used to generate a list of high risk patients who could then be recalled and systematically assessed. For example, the algorithm can identify 10% of the population in which approximately 62% of all new pancreatic cancer cases are likely to be diagnosed over the next 2 years. Table 4 shows the possible thresholds along with the sensitivity, specificity, positive and negative predictive values. This is intended to inform subsequent cost-effectiveness modelling and the choice of thresholds which is outside the scope of this paper.

The algorithms have been developed in one cohort and validated in a separate cohort representative of the patients likely to be considered for preventative measures. The algorithm performed well with good discrimination and calibration. Following independent external validation and cost-effectiveness modelling (which is outside the scope of the present study), the algorithm could potentially be used in clinical practice to identify those at highest risk of having pancreatic cancer to facilitate early referral and investigation and so help earlier identification of patients with pancreatic cancer.

While the study was not designed to determine whether type 2 diabetes causes pancreatic cancer it was found that pancreatic cancer is twice as likely to occur in patients with type 2 diabetes than those without it, although the absolute risk across all patients with diabetes is low (0.3%). No increased risk of pancreatic cancer among patients with type 1 diabetes was found, although there were only one-tenth of the number of patients with type 1 diabetes compared with type 2 diabetes so a type 2 error cannot be excluded. Further research into the potential mechanisms underlying this association for type 2 diabetes may be warranted.

Several additional symptoms not traditionally thought to be associated with pancreatic cancer (such as, abdominal distension in females, dysphagia, and constipation in males) were also identified but which remained independently predictive on multivariate analysis. Given the location of the pancreas, it is possible that a tumour could result in pressure on the fundus of the stomach or lower end of the oesophagus resulting in dysphagia. If confirmed by other studies, this study's results suggest that investigation of the pancreas should be considered alongside other upper gastrointestinal investigations for patients presenting with dysphagia. There may be other symptoms which could predict pancreatic cancer that were not included in this study, such as vomiting, nausea, fever, dyspepsia, backache, and depression. These could be explored in future versions of the algorithm to determine whether they improve its performance or results in reclassification around a risk threshold.

REFERENCES

1. UK CR. *Pancreatic cancer statistics — key facts*. London: Cancer Research, 2011.
2. Li D, Xie K, Wolff R, Abbruzzese JL. Pancreatic cancer. *Lancet* 2004; **363(9414):** 1049–1057.
3. Bowles MJ, Benjamin IS. Cancer of the stomach and pancreas. *BMJ* 2001; **323(7326):** 1413–1416.
4. Purohit V, Khalsa J, Serrano J. Mechanisms of alcohol-associated cancers: introduction and summary of the symposium. *Alcohol* 2005; **35(3):** 155–160.
5. Gullo L. Diabetes and the risk of pancreatic cancer. *Ann Oncol* 1999;**10 (Suppl 4):** 79–81.
6. Takhar AS, Palaniappan P, Dhingsa R, Lobo DN. Recent developments in diagnosis of pancreatic cancer. *BMJ* 2004; **329(7467):** 668–673.
7. Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009; **101 (Suppl 2):** S1–4.
8. NICE. *Referral guidelines for suspected cancer*. London: National Institute for Health and Clinical Excellence, 2005.
9. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336(7659):** 1475–1482.
10. 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.
11. 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.
12. 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 Fam Pract* 2010; **11:** 49.
13. Jones R, Latinovic R, Charlton J, Gulliford MC. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. *BMJ* 2007; **334(7602):** 1040.
14. Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing Type 2 diabetes. *Diabet Med* 2011; **28(5):** 599–607.
15. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; **340:** c2442.
16. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009; **339:** b2584.
17. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7(2):** 147–177.
18. No authors listed. Academic medicine: problems and solutions. The Academic Medicine Group. *BMJ* 1989; **298(6673):** 573–579.
19. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Clin Epidemiol* 2007; **60(9):** 979.
20. Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; **59(10):** 1092–1101.
21. Rubin DB. *Multiple imputation for non-response in surveys*. New York, NY: John Wiley, 1987.
22. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999; **28(5):** 964–974.
23. Hosmer D, Lemeshow S. *Applied logistic regression*. New York, NY: John Wiley & Sons, Inc, 1989.
24. Royston P. Explained variation for survival models. *Stata J* 2006; **6:** 1–14.
25. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23(5):** 723–748.
26. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991; **302(6779):** 766–768.
27. Ferlay J, Autier P, Boniol M, *et al*. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol* 2007; **18(3):** 581–592.
28. Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010; **69(1):** 4–14.
29. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; **60(572):** e128–136.