Gary S Collins and Douglas G Altman

Identifying patients with undetected pancreatic cancer in primary care:

an independent and external validation of QCancer® (Pancreas)

Abstract

Background
Despite its rarity, the prognosis of pancreatic cancer is very poor and it is a major cause of cancer mortality; being ranked fourth in the world, it has one of the worst survival rates of any cancer.

Aim
To evaluate the performance of QCancer® (Pancreas) for predicting the absolute risk of pancreatic cancer in an independent UK cohort of patients, from general practice records.

Design and setting
Prospective cohort study to evaluate the performance QCancer (Pancreas) prediction models in 364 practices from the UK, contributing to The Health Improvement Network (THIN) database.

Method
Records were extracted from the THIN database for 2.15 million patients registered with a general practice surgery between 1 January 2000 and 30 June 2008, aged 30–84 years (3.74 million person-years), with 618 pancreatic cancer cases. Pancreatic cancer was defined as incident diagnosis of pancreatic cancer during the 2 years after study entry.

Results
The results from this independent and external validation of QCancer (Pancreas) demonstrated good performance data on a large cohort of general practice patients. QCancer (Pancreas) had very good discrimination properties, with areas under the receiver operating characteristic curve of 0.89 and 0.92 for females and males respectively. QCancer (Pancreas) explained 60% and 67% of the variation in females and males respectively. QCancer (Pancreas) over-predicted risk in both females and males, notably in older patients.

Conclusion
QCancer (Pancreas) is potentially useful for identifying undetected cases of pancreatic cancer in primary care in the UK.

Keywords
pancreatic cancer; primary care; risk prediction; validation.

INTRODUCTION

The incidence of pancreatic cancer is moderately low: it is the 13th most common cancer worldwide and the 9th most common in high-income countries.[1] Despite its rarity, the prognosis of pancreatic cancer is very poor and it is a major cause of cancer mortality; being ranked fourth in the world,[2] it has one of the worse survival rates of any cancer. In the UK in 2008, 8085 people were diagnosed with pancreatic cancer and 7781 died from this cancer.[3] The 1-year survival rate is less than 20%, while the 5-year survival rate is only around 3%.[3] The difficulty of early detection contributes to the poor prognosis in patients with pancreatic cancer, with most patients going undiagnosed until emergency admission. Pancreatic cancer has very few symptoms until the disease is at a relatively advanced stage, when curative treatment is limited. There are currently very few reliable tools to identify individuals at an increased risk of having or developing pancreatic cancer.[4,5]

QCancer® (Pancreas) is a pair of multivariable prediction models (one for males and one for females) that have been developed recently to predict the risk of having undiagnosed pancreatic cancer.[6] QCancer (Pancreas) was developed and internally validated on a large cohort of 3.6 million patients from the QResearch (www.qresearch.org) database.[6] This is a large database comprising over 12 million anonymised health records from 602 general practices throughout the UK, using the EMIS computer system. The models were derived on 2.4 million patients aged between 30 and 84 years, contributing 1415 incident cases of pancreatic cancer from 4.1 million person-years of observation between 1 January 2000 and 30 September 2010. The final prediction models, based on a Cox proportional hazards model, included eight risk factors for females and nine risk factors for males (Table 1). Open source code to calculate the QCancer (Pancreas) score is available from www.qcancer.org/, released under the GNU Lesser General Public License, version 3. The performance of QCancer (Pancreas) was assessed on separate sample of 1.2 million patients (781 incident cases of pancreatic cancer) from the same QResearch database (different general practices), with good discriminative ability and calibration.[6]

QCancer (Pancreas) is part of a suite of prediction models that form the QCancer scores (www.qcancer.org) that have been developed to predict the risk of having undiagnosed lung,[7] ovarian,[8] colorectal,[9] gastro-oesophageal,[10] renal,[11] or pancreatic cancer.[6] The authors are currently using identical methods to independently evaluate the performance of these six predictions models. To date, they have published validation studies for QCancer (Colorectal),[12] QCancer (Ovarian),[13] QCancer (Gastro-Oesophageal),[14] and QCancer (Renal),[15] to predict the risk of undiagnosed colorectal, ovarian, gastro-oesophageal and renal cancer, respectively. The description of the methods used within this study are thus substantially the same as the first study published, evaluating QCancer (Colorectal).[12]

This article describes the results from

GS Collins, PhD, senior medical statistician; **DG Altman**, DSc, director, professor of statistics in medicine, Centre for Statistics in Medicine, Oxford.
Address for correspondence
Gary S Collins, Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD.
E-mail: gary.collins@csm.ox.ac.uk

Submitted: 15 January 2013; **Editor's response:** 5 February 2013; **final acceptance:** 3 April 2013.

This is the full-length article (published online 27 Aug 2013) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2013; DOI: 10.3399/bjgp13X671623**

How this fits in

The prognosis of pancreatic cancer is poor and is a major cause of mortality. Contributing to the poor prognosis is that pancreatic cancer has very few symptoms until the disease is relatively advanced. A new diagnostic prediction model (QCancer® [Pancreas]) has recently been developed to predict the risk of undiagnosed pancreatic cancer in a primary care population. The prediction model has good predictive ability and could be integrated into clinical computer systems to identify individuals who are at increased risk of having undiagnosed pancreatic cancer.

an independent evaluation of QCancer (Pancreas) on a large dataset of general practice records in the UK that were not used to derive the prediction model.

METHOD

Cohort selection

Study participants were patients registered between 1 January 2000 and 30 June 2008 and recorded on The Health Improvement Network (THIN) database (www.thin-uk.com). The same exclusion criteria as in the original development paper were adopted.[6] Patients were excluded if they had a prior diagnosis of pancreatic cancer, were registered for <12 months with the general practice, had invalid dates, were aged ≤30 years, or ≥85 years. Entry to the cohort was defined as for the original development study,[6] as the latest of (1) the study start date; (2) the date the patient registered with the practice; and (3) for those patients with red flag symptoms (abdominal pain, appetite loss, dysphagia, weight loss, abdominal distension or constipation), the date of the first recorded onset of any red flag symptom within the study period.

Outcome measure

The outcome was diagnosis of pancreatic cancer as described in the original study that developed QCancer (Pancreas),[6] defined as incident diagnosis of pancreatic cancer during the 2 years after study entry. The only difference between this external validation and the original development of QCancer (Pancreas) is that there is no linkage to Office of National Statistics (ONS) death records, and thus the outcome was based solely on what was recorded on the patients' records (THIN). Patients not experiencing the study outcome were censored at the earliest date: either date of death, date of leaving the practice study, or 2 years of follow-up.

Statistical analysis

The 2-year predicted risk of pancreatic cancer was calculated for each patient in the THIN cohort using the QCancer (Pancreas) prediction model (www.qcancer.org/). Multiple imputation, using all predictors plus the outcome variable and censoring status, was used to replace missing values for smoking status. This involves creating multiple copies of the data and imputing the missing values with sensible values randomly selected from their predicted distribution. Ten imputed datasets were generated and results from analyses on each of the imputed datasets were combined using Rubin's rules to produce estimates and confidence intervals that incorporate the uncertainty of imputed values.[16] Smoking status was derived from combining two risk factors: (1) whether the patient was a non-smoker, ex-smoker or current smoker; and (2) the amount of cigarettes smoked, defined as light (<10 cigarettes/day), moderate (10–19 cigarettes/day), or heavy (≥20 cigarettes/day).

Predictive performance of the QCancer (Pancreas) prediction model on the THIN cohort was assessed by examining measures of calibration and discrimination. Calibration refers to how closely the predicted 2-year pancreatic cancer risk agrees with the observed proportions of patients with a diagnosis of pancreatic cancer within 2 years. This was assessed for each tenth of predicted risk, ensuring 10 equally sized groups, and each 5-year age band, by plotting observed proportions versus predicted risk.

Table 1. Risk factors in QCancer® (Pancreas)

Females	Males
Age (years)	Age (years)
Chronic pancreatitis (yes/no)	Chronic pancreatitis (yes/no)
Type 2 diabetes (yes/no)	Type 2 diabetes (yes/no)
Currently consulting GP with first onset of abdominal distension (yes/no)	Currently consulting GP with first onset of dysphagia (yes/no)
Currently consulting GP with first onset of abdominal pain (yes/no)	Currently consulting GP with first onset of abdominal pain (yes/no)
Currently consulting GP with first onset of appetite loss (yes/no)	Currently consulting GP with first onset of appetite loss (yes/no)
Currently consulting GP with first onset of weight loss (yes/no)	Currently consulting GP with first onset of weight loss (yes/no)
Smoking status (non-smoker; light smoker; moderate smoker; heavy smoker)	Smoking status (non-smoker; light smoker; moderate smoker; heavy smoker)
	Recently consulted a GP with first onset of constipation in the past 12 months (yes/no)

Table 2. Characteristics of participants aged 30 to 84 years in the QResearch development and THIN validation cohorts

	QResearch[6]		THIN (external validation)		
Risk predictor	**Development (*n* = 2 364 571)**	**Internal validation (*n* = 1 243 740)**	**Females (*n* = 1 082 730)**	**Males (*n* = 1 067 592)**	**Overall (*n* = 2 150 322)**
Median age, years (standard deviation)	50.1 (15.0)	50.1 (14.9)	49 (15.1)	47 (14.2)	48(14.7)
Smoking status, *n* (%)					
Non-smoker	1 200 385 (50.8)	627 868 (50.5)	488 238 (45.1)	373 3991 (35.0)	861 629 (40.1)
Ex-smoker	426 697 (18.0)	228 970 (18.4)	139 815 (12.9)	172 404 (16.2)	312 219 (14.5)
Current smoker, amount not recorded	71 668 (3.0)	39 438 (3.2)	150 891 (13.9)	133 753 (12.5)	284 644 (13.2)
Light smoker (<10/day)	149 044 (6.3)	80402 (6.5)	68 449 (6.3)	65 757 (6.2)	134 206 (6.2)
Moderate smoker (10–19/day)	180 887 (7.6)	96 443 (7.8)	103 501 (9.6)	101 098 (9.5)	204 599 (9.5)
Heavy smoker (≥20/day)	135 113 (5.7)	74 140 (6.0)	73 023 (6.7)	111 223 (10.4)	184 246 (8.6)
Smoking status not recorded	200 777 (8.5)	96 479 (7.8)	58813 (5.4)	109 966 (10.3)	168 779 (7.8)
Medical history, *n* (%)					
Type 2 diabetes	78 687 (3.3)	41 869 (3.4)	31 172 (2.9)	39 081 (3.7)	70 253 (3.3)
Chronic pancreatitis	2208 (0.1)	1206 (0.1)	880 (0.1)	1212 (0.1)	2092 (0.1)
Current symptoms and symptoms in the preceding year, *n* (%)					
Current appetite loss	10 351 (0.4)	5567 (0.4)	3444 (0.3)	2658 (0.3)	6102 (0.3)
Current weight loss	26 239 (1.1)	14 686 (1.2)	15 980 (1.5)	13 484 (1.3)	29 464 (1.4)
Current abdominal pain	232 586 (9.8)	129 924 (10.4)	148 290 (13.7)	106 768 (10.0)	255 058 (11.9)
Current abdominal distension	7985 (0.3)	4929 (0.4)	4457 (0.4)	—	6456 (0.3)
Current dysphagia	15 648 (0.7)	8507 (0.7)	—	9326 (0.9)	20 152 (0.9)
Constipation in the last year	15 094 (0.6)	8476 (0.7)	—	5326 (0.5)	13 523 (0.6)

Discrimination is the ability of the prediction model to differentiate between patients who experience an event during the study period and those who do not. This measure is quantified by calculating the c statistic: a value of 0.5 represents chance and 1 represents perfect discrimination.[17] The *D* statistic[18] and R^2 statistic[19] were also calculated; these are measures of discrimination and explained variation, respectively, and are tailored towards censored survival data. The *D* statistic is a measure of prognostic separation of survival curves and is closely related to the standard deviation of the prognostic index (the linear component from the Cox model). R^2 (explained variation) is the proportion of total variation in the outcome that is explained by the prediction model, ranging from 0 to 100%.

All statistical analyses were carried out in R (version 2.14.1, http://www.R-project.org) and the ICE (multiple imputation) procedure in Stata (version 11.2).

RESULTS

Between 1 January 2000 and 30 June 2008, 2 150 322 eligible patients from 364 general practices in the UK were registered in the THIN database. The 2 150 322 eligible patients contributed 3 744 567 person-years of observation (median follow-up was 2 years), among whom there were 618 cases of pancreatic cancer (287 females; 331 males). Table 2 details the characteristics of eligible patients. Compared with the original development cohort (QResearch),[6] members of the THIN cohort were marginally younger, proportionally more patients reported abdominal pain (11.9% versus 9.8%), and there were fewer non-smokers (40.1% versus 50.8%).

Complete data on smoking status were available for 80.6% of females (*n* = 873 026) and 77.2% of males (*n* = 823 873). There were noticeably more current smokers in the THIN cohort, for whom the number of cigarettes was not recorded (13.2% in THIN compared to around 3% in the original QResearch database).

Table 3 reports the age–sex incidence rates of each symptom included in the QCancer (Pancreas) prediction models. All the symptoms apart from abdominal pain in females tended to become more common with age. During the follow-up, the crude rate of pancreatic cancer was 17 per 100 000 person-years of observation (compared with 30 per 100 000 in the original QResearch cohort), with 18 per 100 000 person-years for females and 15 per 100 000 person-years for males. As in the original development cohort, incidence rates of pancreatic cancer increased sharply with age.

Performance data for QCancer (Pancreas) from the original development cohort and the THIN cohort (multiple imputation and complete-case) are presented in Table 4.

Table 3. Incidence rates of appetite loss, weight loss, abdominal pain, abdominal distension, and dysphagia per 100 000 person-years by sex and age in the THIN cohort

Symptom and age range, years	Incidence rate (95% CI)	
	Females	Males
Appetite loss		
All ages	52.0 (50.3 to 53.8)	41.5 (39.9 to 43.1)
<35	15.5 (12.9 to 18.6)	10.8 (8.7 to 13.4)
35–44	25.0 (22.6 to 27.5)	17.6 (15.7 to 19.7)
45–54	26.3 (23.7 to 29.1)	21.4 (19.1 to 23.8)
55–64	35.5 (32.2 to 39.1)	34.2 (31.0 to 37.8)
65–74	84.9 (79.1 to 91.1)	85.1 (78.8 to 91.7)
75–84	205.1 (194.2 to 216.4)	207.8 (194.2 to 222.1)
Weight loss		
All ages	241.5 (237.7 to 245.2)	210.6 (207.0 to 214.2)
<35	87.9 (81.4 to 94.9)	55.1 (50.1 to 60.4)
35–44	150.1 (144.3 to 156.2)	108.9 (104.1 to 114.0)
45–54	171.2 (164.6 to 178.1)	142.4 (136.4 to 148.6)
55–64	222.1 (213.7 to 230.8)	232.7 (224.1 to 241.6)
65–74	344.9 (333.0 to 357.1)	383.5 (370.1 to 397.3)
75–84	700.0 (679.8 to 720.7)	779.0 (752.4 to 806.2)
Abdominal pain		
All ages	2240.8 (2229.4 to 2252.2)	1667.4 (1657.4 to 1677.4)
<35	1901.6 (1870.7 to 1932.8)	877.8 (857.5 to 898.4)
35–44	2452.2 (2428.4 to 2476.2)	1542.3 (1523.8 to 1560.9)
45–54	2103.3 (2079.7 to 2127.1)	1560.9 (1540.8 to 1581.2)
55–64	2300.2 (2273.1 to 2327.7)	1885.2 (1860.4 to 1910.2)
65–74	2253.8 (2223.2 to 2284.7)	2165.0 (2132.9 to 2197.4)
75–84	2275.0 (2238.4 to 2312.0)	2529.4 (2481.4 to 2578.1)
Abdominal distension		
All ages	67.3 (65.4 to 69.4)	31.2 (29.9 to 32.6)
<35	29.6 (25.9 to 33.8)	5.9 (4.4 to 7.8)
35–44	52.7 (49.3 to 56.4)	15.3 (13.5 to 17.3)
45–54	60.8 (56.9 to 65.0)	25.5 (23.0 to 28.2)
55–64	71.8 (67.1 to 76.8)	36.4 (33.0 to 40.0)
65–74	91.3 (85.3 to 97.7)	62.2 (56.9 to 67.9)
75–84	121.2 (112.9 to 130.0)	92.2 (83.2 to 101.8)
Dysphagia		
All ages	163.6 (160.5 to 166.7)	145.6 (142.7 to 148.6)
<35	30.4 (26.6 to 34.6)	27.0 (23.5 to 30.8)
35–44	80.5 (76.3 to 85.0)	66.0 (62.2 to 70.0)
45–54	133.1 (127.2 to 139.2)	104.7 (99.5 to 111.0)
55–64	189.5 (181.8 to 197.5)	182.8 (175.1 to 190.7)
65–74	252.5 (242.3 to 263.0)	279.1 (267.7 to 290.9)
75–84	426.4 (410.6 to 442.5)	489.3 (468.3 to 511.0)

The values for the R^2 statistics (percentage of explained variation) and the D statistic were marginally higher in the THIN cohort, 60.0% and 2.51 (females) and 66.6% and 2.89 (males), compared to those reported in the original development paper: 58.7% and 2.44 (females) and 62.0% and 2.61 (males). These high values of the D statistic and also of the c statistic, 0.89 (females) and 0.92 (males), indicate good discrimination properties of QCancer (Pancreas). The value of using individual symptoms to identify patients with pancreatic cancer is compared to QCancer (Pancreas) in Table 5. Using a risk cut-off of 0.2% that identified the 10% of females and males at the highest risk, QCancer (Pancreas) clearly showed an improvement, with 64% and 74% of all new pancreatic cancers identified for females and males respectively. Using the presence of abdominal pain, 54% and 60% of new pancreatic cancers were identified in females and males respectively. The sensitivity of the other individual symptoms ranged between 2.5% (abdominal distention) and 9.1% (weight loss) in females and between 2.7% (dysphagia) and 16.9% (weight loss) in males.

Calibration plots of QCancer (Pancreas) for females and males by tenth of risk are presented in Figure 1. QCancer (Pancreas) increasingly over-predicts risk, with increases across the tenths of risk. Similarly, Figure 2 displays the calibration plots of QCancer (Pancreas) for females and males by 5-year age bands. Again, the QCancer (Pancreas) systematically over-predicts risk across all age groups in both females and males.

DISCUSSION

Summary

QCancer (Pancreas) is a new prediction model to identify individuals with undetected pancreatic cancer in a primary care setting. The prediction model was developed and internally validated on a large primary care electronic database (QResearch) of 3.6 million patients, contributing 2196 new cases of pancreatic cancer between 1 January 2000 and 30 September 2010.

This independent evaluation of the performance of QCancer (Pancreas) was carried out on the large separate database (THIN), based on general practices recording

Table 4. Performance data

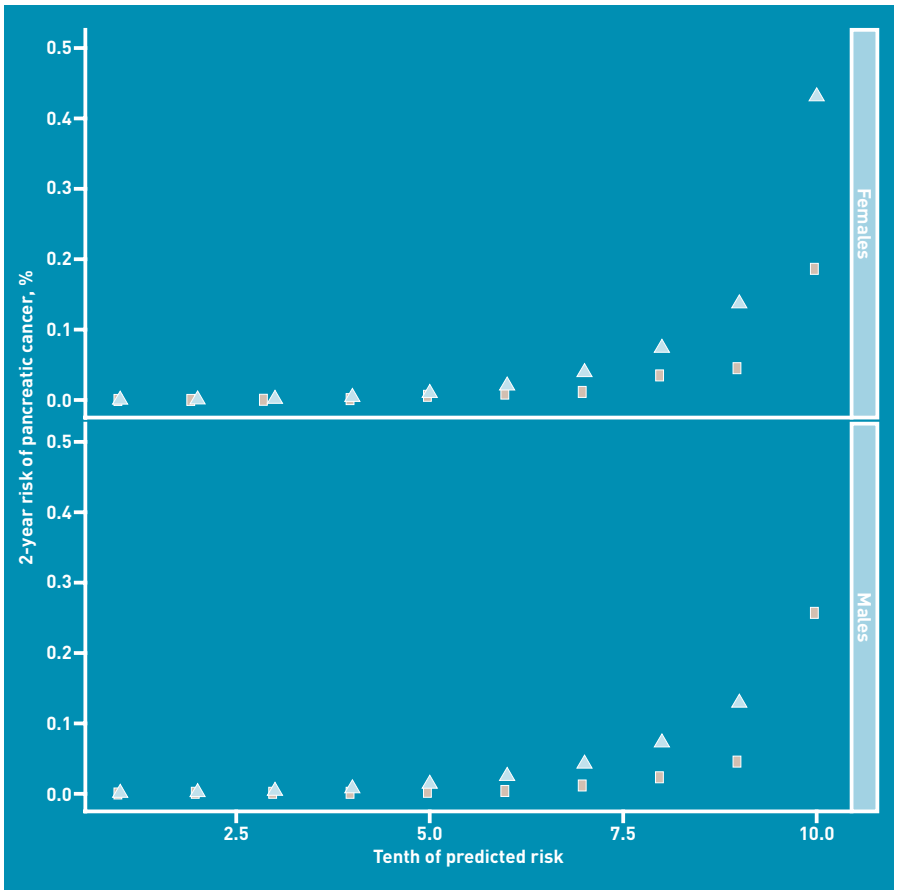
	QResearch (internal validation)		THIN (external validation)			
			Multiple imputation		Complete case	
	Females	Males	Females (n = 1 082 730)	Males (n = 1 067 592)	Females (n = 873 026)	Males (n = 823 873)
R^2 (95% CI)	58.7 (55.4 to 61.9)	62.0 (59.1 to 64.8)	60.0 (56.6 to 63.5)	66.6 (64.1 to 69.2)	62.0 (58.4 to 65.7)	66.1 (63.2 to 69.0)
D statistic (95% CI)	2.44 (2.27 to 2.60)	2.61 (2.45 to 2.77)	2.51 (2.32 to 2.70)	2.89 (2.72 to 3.07)	2.61 (2.40 to 28.3)	2.86 (2.66 to 3.05)
C statistic (95% CI)	0.84 (0.82 to 0.86)	0.87 (0.85 to 0.88)	0.89 (0.87 to 0.90)	0.92 (0.91 to 0.93)	0.90 (0.88 to 0.91)	0.91 (0.90 to 0.93)

Table 5. Comparison of strategies to identify patients having a diagnosis of pancreatic cancer in the next 2 years

Criteria	Risk threshold (%)	True negative	False negative	False positive	True positive	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Females									
Individual symptoms									
Abdominal pain	NA	934 307	133	148 136	154	53.7	86.3	0.1	100.0
Abdominal distension	NA	1 077 993	280	4450	7	2.5	90.6	0.2	100.0
Appetite loss	NA	1 079 010	276	3433	11	3.8	99.7	0.3	100.0
Weight loss	NA	1 066 489	261	15 954	26	9.1	98.5	0.2	100.0
QCancer (Pancreas)									
Top 10% risk	0.2	974 525	104	107 917	183	63.8	99.0	0.2	100.0
Top 5% risk	0.3	1 028 854	155	53 590	132	46.0	95.0	0.2	100.0
Males									
Individual symptoms									
Abdominal pain	NA	960 693	131	106 568	200	60.4	90.0	0.2	100.0
Dysphagia	NA	1 057 944	322	9317	9	2.7	99.1	0.1	100.0
Appetite loss	NA	1 064 616	318	2645	13	3.9	100.0	0.5	100.0
Weight loss	NA	1 053 833	275	13 428	56	16.9	98.7	0.4	100.0
Constipation	NA	1 061 946	320	5315	11	3.2	99.5	0.2	100.0
QCancer (Pancreas)									
Top 10% risk	0.2	961 130	85	106 130	246	74.3	90.1	0.2	100.0
Top 5% risk	0.3	1 014 153	139	53 104	191	57.9	95.0	0.4	100.0

NA = not available.

Figure 1. Observed versus predicted pancreatic cancer risks (triangles denote predicted risk, rectangles denote observed risk).

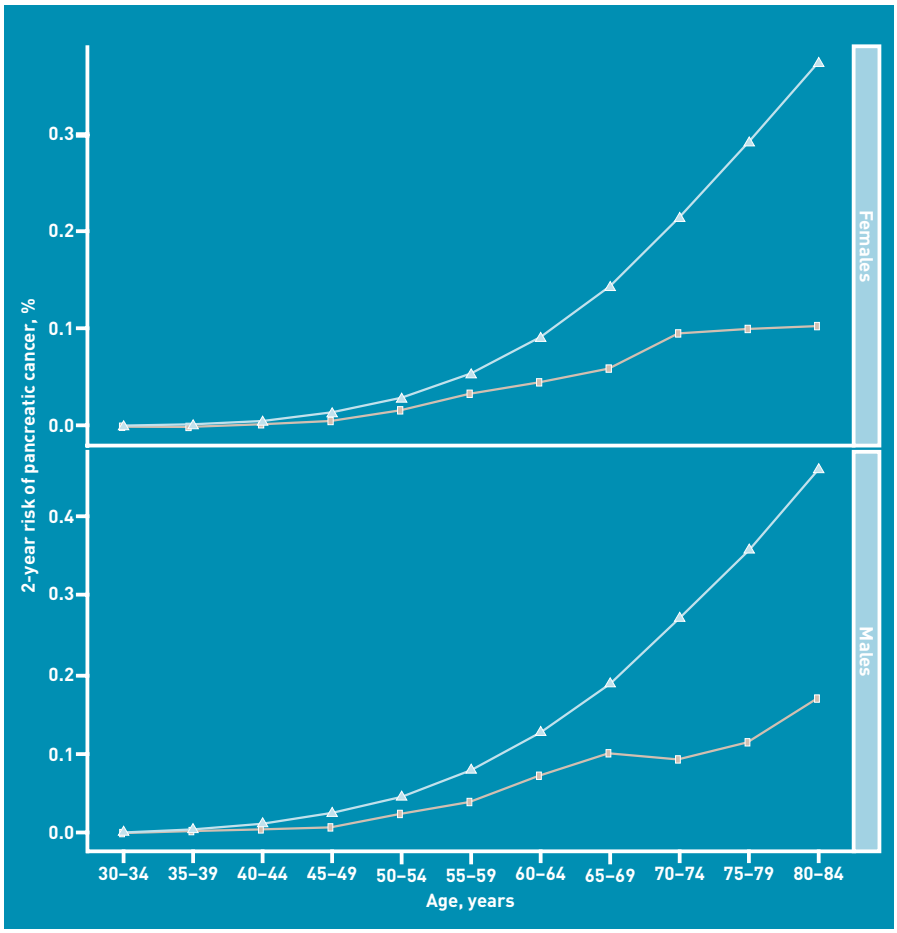


clinical data using the INPS Vision system that is used in 20% of UK general practices. The database comprised 2.15 million patients between 1 January 2000 and 30 June 2008, contributing 3.7 million person-years of observation and 618 cases of pancreatic cancer. The performance data presented in this paper on the THIN cohort provide good evidence to support the external validity of QCancer (Pancreas) in identifying patients with suspected pancreatic cancer, with good discrimination performance that is marginally better than that observed in the internal validation data. QCancer (Pancreas) also clearly outperformed the use of individual symptoms for identifying new cases of pancreatic cancer.

Strengths and limitations

The calibration was disappointingly moderate, with QCancer (Pancreas) systematically over-predicting the risk of undetected cancer. There was a suggestion of this effect in the last three or four tenths of risk in the original study developing the model.[6] The development cohort comprised 2.4 million patients and included only 1415 cases of pancreatic cancer, which arguably can be described as a rare event (despite there being more than the recommended 10 events per variable,[20] and could contribute to lack of calibration observed in the external validation. The case-mix of patients

Figure 2. Observed versus predicted pancreatic cancer risks by sex and age (triangles denote predicted risk, rectangles denote observed risk).



who had undetected pancreatic cancer may also be different between the original development data set and the THIN external validation cohort, which may also contribute to the observed miscalibration. The slight difference in outcome definition also possibly contributes to the miscalibration, with the original study including linkage to death records, while this external validation did not include the linkage. The slight difference in outcome definition meant a difference in the proportion diagnosed (and died) of pancreatic cancer of 0.06% in the original development cohort compared to 0.03% in this external validation.

Comparison with existing literature

QCancer (Pancreas) are the first multivariable prediction models to predict the risk of undiagnosed pancreatic cancer for use in UK primary care. Clearly the usefulness of the prediction models is predicated on having accurate information recorded in primary care electronic healthcare records. However, studies have shown good agreement between diagnoses recorded in general practice databases and other data sources.[21–24]

Implications for clinical practice

To date, the development, internal validation, and this external validation of QCancer (Pancreas) has used 5.8 million patients, contributing 10 million person-years of observation and 2814 cases of pancreatic cancer during the observation periods to develop and evaluate QCancer (Pancreas), in order to predict the risk of pancreatic cancer in adults aged 30–84 years. This study has provided an independent and external evaluation of the QCancer (Pancreas) prediction model on a large cohort of patients in the UK. It has assessed the performance of QCancer (Pancreas) against performance metrics presented in the internal validation of QCancer (Pancreas) and has provided evidence to support the external validity of QCancer (Pancreas).

The performance of QCancer (Pancreas) was similar to that in the internal validation of QCancer (Pancreas), with comparable results indicating excellent discrimination of the prediction model. However, QCancer (Pancreas) over-predicted risk in both females and males, which may be attributed to a slight difference in outcome definition.

Funding

None.

Ethical approval

Ethical approval was received from Trent multicentre research ethics committee.

Provenance

Freely submitted; externally peer reviewed.

Competing interests

The authors have declared no competing interests.

Acknowledgements

We thank Alyson Huntley for her assistance in review of data extraction.

Discuss this article

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

REFERENCES

1. Jemal A, Center MM, DeSantis C, *et al*. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol Biomarkers Prev* 2010; **19(8):** 1893–1907.
2. Hariharan D, Saied A, Kocher HM. Analysis of mortality rates for pancreatic cancer across the world. *HBP (Oxford)* 2008; **10(1):** 58-62.
3. Cancer Research UK. *Pancreatic cancer statistics*. http://info.cancerresearchuk.org/cancerstats/types/pancreas/ (accessed 17 Jul 2013).
4. Wang W, Chen S, Brune KA, *et al*. PancPRO: Risk Assessment for Individuals With a Family History of Pancreatic Cancer. *J Clin Oncol* 2007; **25(11):** 1417–1422.
5. Stapley S, Peters TJ, Neal RD, *et al*. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. *Br J Cancer* 2012; **106(12):** 1940–1944.
6. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2012; DOI: 10.3399/bjgp12X616355.
7. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X606627.
8. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ* 2011; **344:** d8009.
9. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2012; DOI: 10.3399/bjgp12X616346.
10. Hippisley-Cox J, Coupland C. Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X606609.
11. 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; DOI: 10.3399/bjgp12X636074.
12. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). *Br J Cancer* 2012; **107(2):** 260–265.
13. Collins GS, Altman DG. Identifying women with undetected ovarian cancer: independent and external validation of QCancer (Ovarian) prediction model. *Eur J Cancer Care* 2013; **22(4):** 423–429.
14. Collins GS, Altman DG. Identifying patients with undetected gastro-oesophageal cancer in primary care: external validation of QCancer® (Gastro-Oesophageal). *Eur J Cancer* 2013; **49(5):** 1040–1048.
15. Collins GS, Altman DG. Identifying patients with undetected renal tract cancer in primary care: an independent and external validation of QCancer® (Renal) prediction model. *Cancer Epidemiol* 2013; **37(2):** 115–120.
16. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons, 1987.
17. Harrell FE, Califf RM, Pryor DB, *et al*. Evaluating the yield of medical tests. *JAMA* 1982; **247(18):** 2543–2546.
18. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23(5):** 723–748.
19. Royston P. Explained variation for survival models. *Stata J* 2006; **6(1):** 83–96.
20. Peduzzi P, Concato J, Feinsten AR, *et al*. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995; **48(12):** 1503–1512.
21. Blak BT, Thompson M, Dattani H, Bourke A. Generlisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care* 2011; **19(4):** 251-255.
22. Shephard E, Stapley S, Hamilton W. The use of electronic databases in primary care research. *Fam Pract* 2011; **28(4):** 352–354.
23. 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.
24. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; DOI: 10.3399/bjgp10X483562.