Research

Julia Hippisley-Cox and Carol Coupland

Symptoms and risk factors to identify men with suspected cancer in primary care:

derivation and validation of an algorithm

Abstract

Background
Early diagnosis of cancer could improve survival so better tools are needed.

Aim
To derive an algorithm to estimate absolute risks of different types of cancer in men incorporating multiple symptoms and risk factors.

Design and setting
Cohort study using data from 452 UK QResearch® general practices for development and 224 for validation.

Method
Included patients were males aged 25–89 years. The primary outcome was incident diagnosis of cancer over the next 2 years (lung, colorectal, gastro-oesophageal, pancreatic, renal, blood, prostate, testicular, other cancer). Factors examined were: 'red flag' symptoms such as weight loss, abdominal distension, abdominal pain, indigestion, dysphagia, abnormal bleeding, lumps; general symptoms such as tiredness, constipation; and risk factors including age, family history, smoking, alcohol intake, deprivation score and medical conditions. Multinomial logistic regression was used to develop a risk equation to predict cancer type. Performance was tested on a separate validation cohort.

Results
There were 22 521 cancers from 1 263 071 males in the derivation cohort. The final model included risk factors (age, BMI, chronic pancreatitis, COPD, diabetes, family history, alcohol, smoking, deprivation); 22 symptoms, anaemia and venous thrombo-embolism. The model was well calibrated with good discrimination. The receiver operator curve statistics values were: lung (0.92), colorectal (0.92), gastro-oesophageal (0.93), pancreas (0.89), renal (0.94), prostate (0.90) blood (0.83, testis (0.82); other cancers (0.86). The 10% of males with the highest risks contained 59% of all cancers diagnosed over 2 years.

Conclusion
The algorithm has good discrimination and could be used to identify those at highest risk of cancer to facilitate more timely referral and investigation.

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

BACKGROUND

The UK has one of the poorest survival rates for cancer in Europe.[1] This is thought to be partly related to late presentation and delays in diagnosis and treatment. Earlier diagnosis could improve with more targeted investigation of symptomatic patients and increased public awareness of symptoms as encouraged by the National Awareness and Early Diagnosis Initiative (NAEDI).[2] It has been estimated that such an approach might save 5000 lives a year without any new medical advances.[3] In general terms, the earlier the cancer is diagnosed, the more treatment options are available and the better the prognosis. The challenge is to make the correct diagnosis as early as possible despite the non-specific nature of cancer symptoms and signs. This is particularly the case for primary care where GPs need to differentiate those patients for whom further investigation is warranted from those who require reassurance or a 'watch and wait' policy.

QCancer® is an evolving set of prediction models designed to quantify the absolute risk that a patient has an existing cancer based on combinations of readily available risk factors and symptoms.[4–9] The initial approach was to develop separate algorithms for each individual cancer starting with five cancer outcomes: renal,[4] colorectal,[6] pancreatic,[5] gastro-oesophageal,[9] and lung cancer.[8] This approach has been successful in establishing a set of algorithms which are being validated on an external population by an independent team.[10] It's apparent that many of the general symptoms (for example, appetite, weight loss, anaemia, abdominal pain), and some of the more specific symptoms (for example, rectal bleeding), are predictive of multiple types of cancer. In addition, in clinical practice, patients generally consult with one or more symptoms rather than as a suspected case of a particular type of cancer. It is the clinician's job to decide whether a patient's symptoms might indicate serious disease such as cancer, which types of cancer are the most likely, what investigations and referrals might be needed, and the degree of urgency. With this in mind, the scientific approach used to develop the QCancer models was adapted from the individual 'cancer-based approach' towards a more 'symptoms-based approach' which incorporates multiple risk factors and symptoms in one model to predict risk of multiple types of cancer. A symptoms-based approach is more likely to emulate the clinical setting where the decision to investigate or refer is made and could also help optimise the use of scare diagnostic or secondary care resources. It could also help inform the update of the existing National Institute for Health and Clinical Excellence (NICE) guidelines on suspected cancer[11] which is currently underway.

A new risk prediction algorithm was developed and validated to estimate the individualised absolute risk of having different types of cancer incorporating both symptoms and other risk factors, to help

J Hippisley-Cox, MD, FRCGP, MRCP professor of clinical epidemiology & general practice; **C Coupland**, PhD, associate professor in medical statistics, Division of Primary Care, University Park, 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: 6 September 2012; **Editor's response:** 5 October 2012; **final acceptance:** 22 October 2012.

This is the full-length article (published online 12 Dec 2012) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2013; Advance online publication.**
DOI: 10.3399/bjgp13X660724

How this fits in

The UK has one of the worst records for cancer in Europe with late diagnoses and poor survival. Earlier diagnosis of cancer could improve with more targeted investigation of symptomatic patients. Risk assessment tools have the potential to help identify patients at risk of cancer for early referral and investigation although previous tools have tended to focus on individual cancers. Given that patients commonly present with symptoms and that symptoms map to multiple cancers, then a risk assessment tool that takes account of multiple symptoms and risk factors to predict risk of multiple cancers may better support clinical decisions regarding the need for referral or investigation. Primary care research databases can be used to develop prediction algorithms since they contain robust data on many of the relevant variables and outcomes. They also are representative of the populations where such models are likely to be used, especially when integrated into GP computer systems. The study has developed and validated a new algorithm to estimate an individual's overall cancer risk and risk of each type of cancer. The algorithm incorporates multiple symptoms and risk factors which the man is likely to know or which are routinely recorded in GP computer systems.

identify those at highest risk for further investigation or referral. The QResearch® primary care database was used to develop the risk prediction models since 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.[10] It has been used successfully to develop and validate a range of prognostic models[12,13] and models designed to help earlier detection of individual cancers.[4–9] This article describes the derivation and validation of the algorithm in males. The accompanying article describes the results for females.

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 33). All practices in England and Wales that had been using their Egton Medical Information Systems (EMIS) 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 25–89 years was identified, drawn from patients registered with practices between 1 January 2000 and 1 April 2012. Entry to the cohort was the latest of study start date (1 January 2000), 12 months after the patient registered with the practice and for those patients with one or more 'red flag' symptoms, the date of first recorded onset within the study period. Where patients had new onset of multiple red flag symptoms recorded, the entry date was the earliest recorded date of the new symptom in the study period

Patients without a postcode-related Townsend score and those with a recorded red flag symptom in the 12 months before the study entry date were excluded.

Symptoms

Red flag symptoms include symptoms which may herald cancer[4–6,8,9] such as abdominal distension, abdominal pain, appetite loss, heartburn, indigestion, dysphagia, haematemesis, rectal bleeding, haematuria, haemoptysis, neck lump, weight loss, night sweats, haematospermia, testicular lump, and testicular pain. A first occurrence of venous thrombo-embolism was also included as a red flag event as this can herald a previously undiagnosed cancer and recent NICE guidance recommends patients with venous thrombo-embolism have a cancer screen.[14,15]

Patients were also considered as having multiple red flag symptoms if the additional symptoms were recorded within 183 days after the earliest recorded symptom and before the diagnosis of cancer or the date on which the patient left, died, or the study period ended.

More general symptoms were considered for inclusion in the analysis if they were recorded within the 12 months before the cohort entry date. These included nausea, change in bowel habit, constipation, diarrhoea, back pain, bruising, cough, dyspnoea, fever, itching, tiredness, and headache. The following genitourinary symptoms were also included: incontinence, dribbling, impotence, hesitancy, poor stream, nocturia, urgency, frequency, and retention. Jaundice was not included as this is relatively rare, usually considered a sign, and would have its own pathway for investigation.

Baseline risk factors

Factors known to affect baseline cancer risk were as follows:

- age at baseline (continuous, ranging from 25 to 89 years);
- body mass index (BMI; continuous);
- smoking status (non-smoker; ex-smoker; light smoker (1–9 cigarettes/day); moderate smoker (10–19 cigarettes/day); heavy smoker (≥20 cigarettes/day);
- alcohol use (none, trivial (<1 unit/day); light (1–2 units/day); moderate or heavy (≥3 units/day);
- Townsend deprivation score, derived from patients' postcodes (continuous);
- previous diagnosis of cancer;
- anaemia defined as recorded haemoglobin <11 g/dl in the 12 months before study entry or the 60 days after (yes/no);
- family history of gastrointestinal cancer;
- family history of prostate cancer;
- chronic pancreatitis;
- type 1 diabetes; and
- type 2 diabetes.

Clinical outcome definition

The study's primary outcome was cancer which was defined as diagnosis of cancer within 2 years after study entry recorded either on the patients GP record using the relevant UK diagnostic Read Codes or on their linked Office of National Statistics (ONS) cause of death record using the relevant ICD 9 codes (183) or ICD 10 diagnostic codes (C56). The ONS data are currently linked deterministically within the NHS clinical computer system using NHS number, postcode, date of birth and date of death. A 2-year period was used, since this represents the period of time during which existing cancers are likely to become clinically manifest.[16,17] Cancer was subdivided into the following nine types chosen to represent the most common cancers and therefore likely to have sufficient numbers of events to ensure that there were at least 10 events per predictor tested.

- lung cancer;
- colorectal cancer;
- gastro-oesophageal cancer;
- pancreatic cancer;
- renal tract cancer (cancer of the bladder, kidney, or urethra);
- haematological (blood) cancer (leukaemia, lymphoma, and myeloma);
- prostate cancer;
- testicular cancer; and
- other cancers.

Derivation and validation of the models

Multinomial logistic regression was used to estimate the coefficients for each predictor variable for each type of cancer. In this model cancer type was used as the categorical outcome variable, which included the nine types listed above and a category for 'no cancer'. Multiple imputation was used to replace missing values for BMI, and alcohol and smoking status and these values were used in the main analyses.[18–20] Ten imputations were carried out. 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] Analyses were restricted to patients who had a cancer diagnosis within 2 years or had at least 2 years of follow-up. A full model was fitted initially and variables retained in the overall model if they were significant at the 0.01 level. Coefficients were constrained to equal zero for individual types of cancer within the overall model where the risk ratio was between 0.80 and 1.20 (for binary variables). Regression coefficients were combined for each variable from the final model with the constant terms to derive absolute risk equations for each type of cancer. Absolute risk of having any cancer was estimated by summing the absolute risks across the individual cancer types.

Multiple imputation was used in the validation cohort to replace missing values for BMI, alcohol, and smoking. Risk equations obtained from the derivation cohort were applied to the validation cohort to estimate absolute risk. Discrimination was assessed by calculating the receiver operating curve (ROC) statistic for each cancer type. Calibration was assessed by comparing the mean predicted risks with the observed risk by tenth of predicted risk for each individual cancer type.

The validation cohort was used to define the thresholds for the 1%, 5%, and 10% of patients at highest estimated risk of any cancer and each type of cancer. Thresholds for the 10% of patients at highest estimated risk of each type of cancer were also defined. Sensitivity, specificity, positive and negative predictive values were calculated using these thresholds restricting the analyses to males who had the outcome within 2 years or had at least 2 years of follow-up. For comparison, the sensitivity, specificity, positive and negative predictive

Table 1. Baseline characteristics of men in the derivation and validation cohorts.

	Derivation cohort (*n* =1 263 071)	Validation cohort (*n* =679 174)
Mean age (SD)	48 (16.1)	47.9 (16.1)
BMI recorded, *n* (%)	833 205 (66.0)	432 652 (63.7)
Mean BMI (SD)	26.2 (4.1)	26.2 (4.1)
Mean deprivation score, (SD)	–0.3 (3.4)	–0.1 (3.5)
Smoking status, *n* (%)		
Non-smoker	475 266 (37.6)	252 859 (37.2)
Ex-smoker	245 479 (19.4)	121 964 (18.0)
Current: amount not recorded	46 656 (3.7)	25 530 (3.8)
Light (<10/day)	82 564 (6.5)	42 330 (6.2)
Moderate (10–19/day)	90 649 (7.2)	48 434 (7.1)
Heavy (≥20/day)	74 769 (5.9)	40 021 (5.9)
Smoking not recorded	247 688 (19.6)	148 036 (21.8)
Alcohol status, *n* (%)		
None	169 927 (13.5)	89 840 (13.2)
Trivial <1 unit/day	271 223 (21.5)	136 079 (20.0)
Light 1–2 units/day	293 502 (23.2)	149 000 (21.9)
Moderate or heavy ≥3 units/day	138 294 (10.9)	70 110 (10.3)
Alcohol not recorded	390 125 (30.9)	234 145 (34.5)
Medical and family history, *n* (%)		
Prior cancer	26 884 (2.1)	13 977 (2.1)
Family history of gastrointestinal cancer	13 201 (1.0)	6646 (1.0)
Family history of prostate cancer	1323 (0.1)	654 (0.1)
Chronic pancreatitis	1340 (0.1)	747 (0.1)
Chronic obstructive pulmonary disease	25 428 (2.0)	13 217 (1.9)
Type 1 diabetes	4693 (0.4)	2364 (0.3)
Type 2 diabetes	46719 (3.7)	25 209 (3.7)
Anaemia	12 046 (1.0)	6031 (0.9)

BMI = body mass index. SD = standard deviation.

values of individual symptoms in relation to a combined cancer outcome were also calculated. All the available data on the database were used to maximise the power and also generalisability of the results. STATA (version 12) was used for all analyses.

RESULTS

Overall study population

Overall, 676 QResearch practices in England and Wales met the inclusion criteria, of which 452 were randomly assigned to the derivation dataset with the remainder assigned to the validation cohort. A total of 1 395 148 males aged 25–89 years were identified in the derivation cohort. The following were excluded 72 717 males (5%) without a recorded Townsend deprivation score, and 59 360 (4%) with at least one red flag symptom recorded in the 12 months prior to entry to the study leaving 1 263 071 males for analysis.

A total of 763 659 males aged 25–89 years in the validation cohort met the inclusion criteria. The following were excluded: 52 026 males (7%) without a recorded Townsend deprivation score, and 32 459 (4%) with at least one red flag symptom recorded in the 12 months prior to entry to the study leaving 679 174 males for analysis.

Baseline characteristics and symptoms

The baseline characteristics of males in the derivation and validation cohorts are shown in Table 1. Table 2 shows the frequency of symptoms in both cohorts at entry to the cohort. The five most common symptoms in the derivation cohort were: abdominal pain (7%), indigestion (3%), back pain (3%), cough (3%), and rectal bleeding (2%).

Cancer outcomes

There were 22 521 incident cases of cancer arising over 2 years in 1 263 071 males in the derivation cohort. There were 11 913 cancers arising in 679 174 males in the validation cohort. The types of cancer are shown in Table 3. The five most common cancers in the derivation cohort were prostate cancer (21%), lung cancer (15%), colorectal cancer (14%), real tract cancer (12%), and gastro-oesophageal cancer (10%). The pattern was similar in the validation cohort.

Multivariate analysis

Table 4 (available at www.qcancer.org) summarises which symptoms are associated with which cancers after adjustment for other symptoms and risk factors in the final multinomial model. The table shows the numbers of symptoms associated with a particular cancer and the numbers of cancers associated with a particular symptom. For example, blood cancers are associated with 13 symptoms (abdominal distension, abdominal pain, anaemia, appetite loss, dysphagia, haematuria, haemoptysis, indigestion, neck lump, night sweats, testicular lump, venous thrombo-embolism, and weight loss). Abdominal pain is associated with eight cancers (blood, colorectal, gastro-oesophageal, lung, pancreas, prostate, renal, and 'other cancers').

The following symptoms were not included in the final model since they did not meet the pre-specified inclusion criteria for the model overall or for individual cancer types within the model: tiredness, back pain, nausea, itching, dyspnoea, diarrhoea, fever, urinary incontinence, urinary hesitancy and urgency.

Table 5 (available at www.qcancer.org) shows the adjusted risk ratios for the final multinomial model incorporating the various risk factors and symptoms. The

Table 2. Frequency of red flag and recent general symptoms in men in the derivation and validation cohort

	Derivation cohort, *n* (%)	Validation cohort, *n* (%)
Red flag symptoms		
Abdominal distension	2340 (0.2)	1145 (0.2)
Abdominal pain	91 476 (7.2)	47 032 (6.9)
Appetite loss	4234 (0.3)	1887 (0.3)
Dysphagia	7088 (0.6)	3553 (0.5)
Haematemesis	7380 (0.6)	3803 (0.6)
Haematuria	26 230 (2.1)	13 527 (2.0)
Haemoptysis	7065 (0.6)	3559 (0.5)
Haematospermia	2848 (0.2)	1441 (0.2)
Heartburn	8846 (0.7)	4388 (0.6)
Indigestion	42 367 (3.4)	21 893 (3.2)
Neck lump	3758 (0.3)	1967 (0.3)
Night sweats	2264 (0.2)	1164 (0.2)
Rectal bleeding	27 446 (2.2)	14 006 (2.1)
Testicular pain	5571 (0.4)	2809 (0.4)
Testicular lump	8709 (0.7)	4329 (0.6)
Venous thrombo-embolism	9073 (0.7)	4603 (0.7)
Weight loss	11 197 (0.9)	5860 (0.9)
Recent general symptoms		
Back pain	41 216 (3.3)	21 725 (3.2)
Bruising	975 (0.1)	470 (0.1)
Change in bowel habit	2614 (0.2)	1321 (0.2)
Constipation	10 892 (0.9)	5638 (0.8)
Cough	37 745 (3.0)	19 338 (2.8)
Diarrhoea	17 936 (1.4)	9073 (1.3)
Dyspnoea	9898 (0.8)	4780 (0.7)
Fever	3830 (0.3)	1741 (0.3)
Headache	14 844 (1.2)	7882 (1.2)
Hesitancy	509 (0.0)	236 (0.0)
Impotence	12 299 (1.0)	6353 (0.9)
Itching	2066 (0.2)	1122 (0.2)
Nausea	3762 (0.3)	1636 (0.2)
Nocturia	3287 (0.3)	1806 (0.3)
Poor stream	357 (0.0)	188 (0.0)
Tiredness	12 370 (1.0)	6132 (0.9)
Urgency	1109 (0.1)	526 (0.1)
Urinary dribbling	457 (0.0)	230 (0.0)
Urinary frequency	5147 (0.4)	2638 (0.4)
Urinary incontinence	2141 (0.2)	1033 (0.2)
Urinary retention	3113 (0.2)	1594 (0.2)

risk factors in the final model included fractional polynomial terms for age and body mass index, smoking status, Townsend deprivation score, alcohol, family history of prostate cancer, family history of gastrointestinal cancer, chronic pancreatitis, chronic obstructive airways disease and type 2 diabetes.

Venous thrombo-embolism. On multivariate analysis, venous thrombo-embolism was associated with a significant increased risk of six cancers as shown in Table 5 (available at www.qcancer.org): testicular cancer (9-fold increased risk); pancreas and 'other' cancer (3-fold increase); prostate, lung and blood (2-fold increased risk).

General symptoms: appetite loss, weight loss, night sweats and anaemia. Appetite loss was associated with an increased risk of all cancers on multivariate analysis except testicular cancer and renal cancer: 4-fold increased risk of pancreatic cancer; 3-fold increased risk of blood, gastro-oesophageal and 'other cancer'; 2-fold increased risk of lung and colorectal cancer and a 1.4-fold increased risk of prostate cancer.

Weight loss was associated with an increased risk of all cancers on multivariate analysis except testicular cancer: 8-fold increased risk of pancreatic cancer; 4-fold increased risk of gastro-oesophageal and lung; 3-fold increased risk of blood, colorectal and 'other' cancer; 2-fold increased risk of prostate and renal cancer.

Night sweats were associated with increased risk of 3 cancers on multivariate analysis: blood (5-fold), lung (2-fold), and renal tract cancer (3-fold).

Anaemia was associated with increased risk of five cancers: blood (7-fold), colorectal (4-fold), gastro-oesophageal and 'other cancer' (3-fold); lung (2-fold).

Abdominal symptoms (dysphagia, pain, distension, indigestion and/or heartburn). Dysphagia was associated with increased risk of five cancers: gastro-oesophageal (46-fold); lung, other, pancreas (3-fold); blood cancer (2-fold).

Abdominal pain was associated with increased risk of all cancers except testicular cancer as shown in Table 5 (available at www.qcancer.org). Highest risks were for pancreatic cancer (9-fold), colorectal cancer (5-fold) and gastro-oesophageal cancer (3 fold).

Abdominal distension was associated with an increased risk of three cancers: colorectal (4-fold); blood and 'other' cancer (2-fold).

Indigestion was also associated with five cancers as shown in Table 5 (available at www.qcancer.org). These were gastro-oesophageal (7-fold); pancreatic cancer (4-fold); blood cancer (2-fold), lung (1.3-fold), other (1.4-fold). Heartburn was associated with a 3-fold increased risk of gastro-oesophageal cancer.

Haematemesis. Haematemesis was associated with increased risk of three cancers on multivariate analysis: gastro-oesophageal (6-fold); pancreas (3-fold); 'other cancer' (2-fold).

Haematuria. Haematuria was associated with increased risk of four cancers on multivariate analysis: renal tract cancer

Table 3. Numbers (%) men with cancer outcomes in the derivation and validation samples

	Derivation cohort		Validation cohort	
	n	**%**	***n***	**%**
Total patients	1 263 071		679 174	
No cancer	1 240 550	98.2	667 261	98.2
Any cancer	22 521	1.8	11 913	1.8
Cancer type				
Lung	3351	14.9	1761	14.8
Colorectal	3250	14.4	125	14.5
Gastro-oesophageal	2212	9.8	1174	9.9
Pancreatic	732	3.3	405	3.4
Renal tract	2579	11.5	1333	11.2
Prostate	4640	20.6	2477	20.8
Blood	1664	7.4	896	7.5
Testicular	437	1.9	225	1.9
Other	3656	16.2	1917	16.1

(64-fold); prostate (4-fold); blood cancer and 'other' cancer (2-fold).

Lumps in neck or testis. Neck lumps were associated with an increased risk of four cancers: blood (24-fold); 'other' (19-fold); lung (3-fold); gastro-oesophageal (2-fold).

Testicular lumps were associated with an increased risk of three cancers: testicular (185-fold), prostate (2-fold) and blood (3-fold). Testicular pain was associated with a 16-fold increased risk of testicular cancer and a 2-fold increased risk of prostate cancer.

Prostate cancer and urinary symptoms. Four genitourinary symptoms were predictive of prostate cancer only: these were urinary retention, frequency, nocturia, and impotence. The section of the model for prostate cancers included the following predictors: age, body mass index, Townsend deprivation score, family history of prostate cancer, abdominal pain, appetite loss, haematuria, rectal bleeding, testicular pain, testicular lump, venous thrombo-embolism, weight loss, impotence and nocturia, urinary frequency and urinary retention.

Validation: discrimination

Table 6 shows the ROC statistic values for each cancer type in the validation cohort using the algorithm from the multinomial model. All values were above 0.82 indicating very good discrimination. The highest ROC values were for renal tract cancer (0.94), gastro-oesophageal cancer (0.93), and lung cancer (0.92). The lowest was for testicular cancer (0.82).

Table 6 also shows the ROC values for the original QCancer models based on published equations for each separate cancer outcome where available.[4–9] Generally there were no significant differences observed between the ROC values for the new multinomial model compared with the original.

Validation: calibration

Figure 1 (available at www.qcancer.org) shows the mean predicted scores and the observed risks 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 for each cancer type with close correspondence between predicted and observed risks within each model tenth except for the 'other cancer' model which showed a degree of over prediction.

Sensitivity, specificity and predictive power of individual symptoms

Table 7 (available at www.qcancer.org) gives the sensitivity, specificity positive and negative predictive power of individual symptoms for predicting an overall outcome of 'any cancer'. Symptoms with the highest positive predictive values for any cancer (regardless of type) were anaemia (19%), urinary retention (14%), dysphagia (13%), haematuria (13%), weight loss (11%), neck lump (10%), haemoptysis (10%). The positive predictive value for venous thrombo-embolism was 6%. The sensitivity of single symptoms was generally low with the highest value being 16% for abdominal pain.

Table 7 (available at www.qcancer.org) also shows the sensitivity, specificity positive and negative predictive power for predicting a diagnosis of any cancer based on three risk thresholds. The 90th centile

Table 6. Multinomial prediction algorithms in men aged 25–89 years in the validation sample. The individual model figures refer to the published QCancer® models developed using individual cancer outcomes[4–9]

Site	Men (multinomial model)	Men (individual models)
Any cancer	0.87 (0.88 to 0.89)	n/a
Lung	0.92 (0.91 to 0.92)	0.92 (0.91 to 0.93)
Colorectal	0.90 (0.90 to 0.91)	0.91 (0.90 to 0.91)
Gastro-oesophageal	0.93 (0.92 to 0.93)	0.92 (0.91 to 0.93)
Pancreas	0.89 (0.87 to 0.91)	0.87 (0.85 to 0.88)
Renal tract	0.94 (0.93 to 0.95)	0.95 (0.94 to 0.96)
Prostate	0.90 (0.90 to 0.91)	n/a
Blood	0.83 (0.81 to 0.84)	n/a
Testis	0.82 (0.79 to 0.85)	n/a
Other	0.86 (0.85 to 0.87)	n/a

ROC = receiver operating curve.

Table 8. Comparison of strategies to identify men at risk of having a diagnosis of different types of cancer based on the top 10% at highest risk for each cancer in the validation cohort

Top 10% of risk	Risk threshold %	True negative	False negative	False positive	True positive	Sensitivity %	Specificity %	PPV %	NPV %
Lung cancer	0.67	610 755	502	66 658	1259	71.5	90.2	1.9	99.9
Colorectal cancer	0.45	610 731	526	66 718	1199	69.5	90.2	1.8	99.9
Gastro-oesophageal	0.29	610 982	275	67 018	899	76.6	90.1	1.3	100.0
Pancreatic cancer	0.10	611 136	121	67 633	284	70.1	90.0	0.4	100.0
Renal tract cancer	0.20	611 026	231	66 815	1102	82.7	90.1	1.6	100.0
Prostate cancer	1.30	610 254	1 003	66 443	1474	59.5	90.2	2.2	99.8
Blood cancer	0.27	610 800	457	67 478	439	49.0	90.1	0.6	99.9
Testicular cancer	0.02	611 183	74	67 766	151	67.1	90.0	0.2	100.0
Other cancer	0.66	611 065	192	67 846	71	27.0	90.0	0.1	100.0

NPV = negative predictive value. PPV = positive predictive value.

defined a high risk group with a cancer risk score of >4%. The positive predictive power was 10%, the sensitivity was 59%, and the specificity 91%. The 95th centile defined a high-risk group with a cancer risk score of >7%. The positive predictive power was 15%, the sensitivity was 42%, and the specificity 96%. The 99th centile defined a high risk group with a cancer risk score of >19%. The positive predictive power was 22%, the sensitivity was 13%, and the specificity 99%.

Table 8 shows the sensitivity, specificity, positive and negative predictive values for predicting cancer type based on the top 10% at risk of each individual cancer. For example, the 90th centile for prostate cancer defined a high risk group of >1%. The positive predictive power was 2%, the sensitivity 60%, and the specificity 90%. clinical examples are shown in Box 1.

Box 1. Clinical examples

- An 81-year-old male, who is a light drinker and a light smoker has abdominal pain, appetite loss, indigestion, and weight loss, and has had constipation and a cough recorded in the last 12 months. His overall cancer risk is 80.4% comprising pancreas (34.9%), gastro-oesophageal (20.8%), lung (9.0%), other (9.1%), colorectal (3.2%), prostate (1.4%), blood (1.7%), renal (0.2%), and other (9.2%).
- A 60-year-old male, who is a trivial drinker and a heavy smoker, has appetite loss, dysphagia and indigestion. His overall cancer risk is 64.4% comprising gastro-oesophageal (58.6%), lung (2.0%), pancreas (1%), other (2.1%), colorectal (0.1%), renal (0.1%), prostate (0.2%), and blood (0.5%). A similar male of 60 years who is a trivial drinker and heavy smoker but without any symptoms has an overall cancer risk of 2.3%.
- A 70-year-old male, who is a trivial drinker and a light smoker, has type 2 diabetes and a history of chronic pancreatitis. He has abdominal pain and night sweats and has had anaemia and constipation in the past 12 months. His overall cancer risk is (33.3%) comprising blood (8.1%), ´other cancer (6.2%), colorectal (5.2%), pancreas (4.5%), gastro-oesophageal (1.9%), prostate (1.1%), other cancer (6.3%).
- A 45-year-old male, who is a non-drinker and a non-smoker, with family history of prostate cancer and has abdominal pain and has had impotence, nocturia, and frequency in the last 12 months. His overall cancer risk is 1.37% comprising prostate (0.77%), other cancer (0.36%), colorectal (0.14%), and blood (0.10%).

DISCUSSION

Summary

This research has developed and validated a new algorithm designed to estimate the absolute risk of having existing but as yet undiagnosed cancer in men. The algorithm is based on a combination of symptoms and risk factors such as age and family history of cancer which the man is likely to know and which are recorded in GP electronic records. The original work has been extended by including multiple risk factors and symptoms as predictors for nine cancer types within one model. By modelling the cancer types simultaneously using multinomial logistic regression, the resulting algorithm will not only give the probabilities of each type of cancer for a given set of patient characteristics, but will also give an overall ´cancer risk´ as well as the risk that the patient does not have cancer. The trade-off is that the algorithm has more parameters although if the algorithms are embedded in GP clinical systems as intended, then much of the data needed for the calculation is already available, leaving the clinician to supplement the information at the point of care. It is important to note that the algorithm does not actually result in a diagnosis of cancer — rather it can be used to identify a subset of high-risk men suitable for targeted investigation or a subset of particularly low-risk men for whom reassurance might be appropriate. The 10% of patients with the highest risks contained 59% of all cancers diagnosed over the next 2 years.

Strengths and limitations

Strengths and limitations of the methods used in this study have been discussed in detail elsewhere[4–9] so are summarised here. 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 and prescribed medications.[23] 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. Algorithms have been developed in one cohort and validated in a separate cohort representative of the patients likely to be considered for referral and treatment. Lastly, the algorithm can be built into clinical systems. Electronic templates and alerts could be displayed when a red flag symptom is recorded in the patient's record. The template would then help structured data entry of other related symptoms including significant negative findings and the results generated automatically with suggestions on next steps (for example, suitability for further blood test, imaging, or referral) which potentially has a greater utility than a paper based flow chart which might be difficult for busy clinicians to remember in routine primary care. Over time integration into GP computer systems is likely to improve the accuracy and completeness of the electronic record and hence the underlying data used for future versions of this algorithm.

Limitations include lack of formally adjudicated outcomes, potential information bias, and missing data. The database has linked cause of death from the UK ONS and, therefore, this study is likely to have picked up the majority of cases of cancer, thereby minimising ascertainment bias. Patients diagnosed with 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. 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.[24,25] Recording of symptoms may be less complete or accurate than diagnostic codes since patients might not visit their GP with mild symptoms, may not report all symptoms to their GP when they do consult or GPs might not record all the symptoms in the electronic health record. The effect of this information or recording bias could be to underestimate risk ratios if symptoms are not reported and/or recorded or to over-inflate the risk ratios if only the more severe symptoms were reported and/or recorded. Similarly, family history of some types of cancer might be under-recorded since it is not routinely assessed and recorded in GP records.

Comparison with previous studies

This study has good clinical and content validity since the direction and magnitude of the risk ratios and predictive value of individual symptoms in the study are comparable to those reported elsewhere.[16,17,26–27] For example, a previous study examining clinical features of 217 patients with prostate cancer registered with 21 general practices in Exeter between 1998 and 2002 identified seven predictive symptoms (urinary retention, weight loss, impotence, frequency, hesitancy, nocturia, and haematuria).[27] The current study confirmed all of these with similar risk ratios except for hesitancy. In addition, the algorithm incorporated additional factors such as age, BMI, deprivation score, family history of prostate cancer, abdominal pain, appetite loss, rectal bleeding, testicular lump, and venous thrombo-embolism. Compared with the CAPER studies,[26,27] this study is much larger and nationally rather than locally based and has the potential to be updated as populations change, data quality improves and requirements evolve. Unlike CAPER, the QCancer algorithm includes established risk factors such as age, sex, family history, smoking, and other conditions and gives a combined individualised measure of absolute cancer risk for each type of cancer and for cancer overall.

The study found that risk of an existing testicular cancer increases with symptoms (testicular lump and testicular pain), venous thrombo-embolism and risk factors such age and body mass index. While a risk score for testicular cancer is unlikely to affect the decision to refer a male with a testicular lump, it could be useful for alerting the clinician to the possibility of cancer (testicular, prostate, and blood cancers) and for sharing the information on the risk with the patient. It might also be useful to include the information in a referral letter to help with prioritisation and future investigation once in the hospital setting.

Implications for clinical guidelines

This study is topical given the guidelines on referral of suspected cancer published by NICE in 2005 which are currently under review.[11] While it has been possible to confirm associations for many symptoms with cancer diagnoses, this study potentially provides new information on which to base guidance for GPs. It has also identified that some symptoms, such haematuria, generally thought to map to one main cancer each, actually map to multiple types of cancer. Other symptoms currently included in NICE guidelines, such as tiredness, itching and fever however, were not significant independent predictors in this analysis. Similarly, symptoms such as appetite loss and venous thrombo-embolism which are independently predictive of cancer on multivariate analysis and which are not included in the NICE guideline were identified. Importantly, the algorithm better accounts for age than the NICE guideline which simply dichotomises patients into those aged <50 or ≥50 years.[11] This is relevant since the risk of cancers generally increases with age. This study also quantified the risk associated with family history of cancer and incorporated it into the underlying algorithm so that it contributes to a patient's estimated absolute risk of cancer. Information has been provided on the sensitivity, specificity, positive and negative predictive powers at different thresholds of risk so that this can be used for cost-effectiveness modelling which is outside the scope of the present study. Such modelling, along with an evaluation of the performance of diagnostic investigations in symptomatic patients in primary care setting has the potential to inform future revisions of the NICE guideline.

The absolute risk of cancer in patients presenting with a first episode of venous thrombo-embolism has been quantified. This is relevant to the recent publication of NICE guidelines on thrombosis (2012) which recommend cancer screening in such patients if they are aged over 40.[15] The recommended tests include a chest X-ray, blood tests (full blood count, serum calcium, and liver function tests), urinalysis with further investigations as necessary.[18] These results confirm that venous thrombosis is predictive of nearly all cancer types except renal cancer although the risk ratios varied substantially with highest risks for abdominal malignancies. This tool will enable clinicians to quantify the risks each cancer for men with thrombo-embolism to ensure that the relevant investigations are undertaken.

This study has developed a model which can be used to estimate the absolute risk of patients having an existing but as yet undiagnosed cancer taking account of risk factors and symptoms. The algorithm predicts overall cancer risk and risk of each type of cancer. It is based on simple clinical variables which can be ascertained in clinical practice. While the algorithm itself does not make a diagnosis of cancer, it performed well to identify high risk patients in a separate validation sample with good discrimination and calibration. However, the early diagnosis of cancer remains a challenge. Further research is needed to assess how best to implement the algorithm, its cost-effectiveness and whether, upon implementation, it has any impact on the stage of cancer at diagnosis and subsequent survival.

Funding

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

Ethical approval

All QResearch® studies are 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 algorithm which will be publically available alongside the paper and open source software. http://qcancer.org/2013/male

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 software that implements the algorithms described in this paper is free for anyone to use under the terms of the GNU lesser GPL3. For those who wish to implement software in a closed source setting, then a licence fee is payable to ClinRisk Ltd.

Discuss this article

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

REFERENCES

1. Berrino F, De Angelis R, Sant M, *et al.* Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995–99: results of the EUROCARE-4 study. *Lancet Oncol* 2007; **8(9):** 773–783.
2. Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009; **101 Suppl 2:** S1–4.
3. Department of Health. The cancer reform strategy. London: Department of Health, 2007.
4. 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.
5. 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.
6. 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.
7. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ* 2012; **344:** d8009.
8. 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.
9. 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.
10. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). *Br J Cancer* 2012; **107(2):** 260–265.
11. National Institute for Health and Clinical Excellence. *Referral guidelines for suspected cancer*. London: NICE, 2005.
12. 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.
13. 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.
14. Oudega R, Moons KGM, Karel Nieuwenhuis H, *et al.* Deep vein thrombosis in primary care: possible malignancy? *Br J Gen Pract* 2006; **56(530):** 693–696.
15. National Institute for Health and Clinical Excellence. *Venous thromboembolic diseases: the management of venous thromboembolic disease and the role of thrombophilia testing.* NICE guidance no. CG144. London: NICE, 2012.
16. 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.
17. Jones R, Charlton J, Latinovic R, Gulliford MC. Alarm symptoms and identification of non-cancer diagnoses in primary care: cohort study. *BMJ* 2009; **339:** b3094.
18. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7(2):** 147–177.
19. The Academic Medical Group. Academic Medicine: problems and solutions. *British Medical Journal* 1989; **298:** 573–579.
20. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Clin Epidemiol* 2007; **60(9):** 979.
21. Rubin DB. *Multiple imputation for non-response in surveys.* New York: 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. 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.
24. 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..
25. 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.
26. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. *Br J Cancer* 2009; **101 Suppl 2:** S80–86.
27. Hamilton W, Sharp DJ, Peters TJ, Round AP. Clinical features of prostate cancer before diagnosis: a population-based, case-control study. *Br J Gen Pract* 2006; **56(531):** 756–762.