Julia Hippisley-Cox and Carol Coupland

Symptoms and risk factors to identify women 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 women 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 females aged 25–89 years. The primary outcome was incident diagnosis of cancer over the next 2 years (lung, colorectal, gastro-oesophageal, pancreatic, ovarian, renal tract, breast, blood, uterine, cervix, other). Factors examined were: 'red flag' symptoms including weight loss, abdominal pain, indigestion, dysphagia, abnormal bleeding, lumps; general symptoms including tiredness, constipation; and risk factors including age, family history, smoking, alcohol intake, deprivation, body mass index (BMI), 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 23 216 cancers from 1 240 864 females in the derivation cohort. The final model included risk factors (age, BMI, chronic pancreatitis, chronic obstructive pulmonary disease, diabetes, family history, alcohol, smoking, deprivation); 23 symptoms, anaemia and venous thrombo-embolism. The model was well calibrated with good discrimination. The receiver operating curve statistics were lung (0.91), colorectal (0.89), gastro-oesophageal (0.90), pancreas (0.87), ovary (0.84), renal (0.90), breast (0.88), blood (0.79), uterus (0.91), cervix (0.73), other cancer (0.82). The 10% of females with the highest risks contained 54% 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 may 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 six cancer outcomes: ovarian,[7] 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 is apparent that many of the general symptoms (for example, appetite, weight loss, anaemia, and abdominal pain), and some of the more specific symptoms (such as 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 may indicate serious disease such as cancer, which types of cancer are the most likely, what investigations and referrals may 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

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:** 24 October 2012; **final acceptance:** 2 November 2012.

©British Journal of General Practice

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 2012; Advance online publication.**

DOI: 10.3399/bjgp13X660733

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 woman is likely to know or which are routinely recorded in GP computer systems.

different types of cancer incorporating both symptoms and other risk factors, to help 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 females. The accompanying article describes the results for males.

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. Females from the age of 25 years were included to capture cancers which affect a younger age group such as cervical cancer, haematological malignancies, and breast cancer, and so that the algorithm can be used in younger females presenting with alarm symptoms.

Entry to the cohort was defined as the latest of study start date (1 January 2000) and 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 consultation with a symptom within the study period. Where patients had new onset of multiple red flag symptoms recorded, the entry date was the earliest recorded date of a 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 indicate cancer[4,5,7,8,10,14,15] such as abdominal distension, abdominal pain, appetite loss, heartburn, indigestion, dysphagia, haematemesis, rectal bleeding, haematuria, haemoptysis, neck lump, weight loss, night sweats, breast lump, breast pain, nipple discharge or breast skin changes, dyspareunia, inter-menstrual bleeding, post-menopausal bleeding, and post-coital bleeding. 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.[16,17]

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, headache, vaginal discharge, urinary incontinence, urinary retention, nocturia, urgency, and urinary frequency. These symptoms tend to be more common than the other red flag symptoms and are generally not considered to be alarm symptoms in quite the same way. 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 such as age, family history and medical conditions[4–9,15] 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 breast cancer;
- family history of gastrointestinal cancer;
- family history of ovarian cancer;
- benign breast disease;
- chronic pancreatitis;
- type 1 diabetes;
- type 2 diabetes;
- endometriosis;
- endometrial hyperplasia or polyp;
- fibroid;
- polycystic ovarian disease;
- rheumatoid arthritis;
- systemic lupus erythematosis;
- HIV or AIDS;
- oral contraceptive use; and
- hormone replacement therapy.

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.[18,19] Cancer was subdivided into the following 11 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);
- breast cancer;
- ovarian cancer;
- uterine cancer;
- cervical 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 11 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.[20–22] Ten imputations were carried out. Rubin's rules were used to combine the results across the imputed datasets.[23] Fractional polynomials were used to model non-linear risk relationships with continuous variables.[24] 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

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

	Derivation cohort (*n* = 1 240 864)	Validation cohort (*n* = 667 603)
Mean age (SD)	50.3 (17.5)	50.1 (17.4)
BMI recorded, *n* (%)	924 268 (74.5)	480 001 (71.9)
Mean BMI (SD)	25.8 (4.9)	25.8 (4.9)
Mean deprivation score, (SD)	–0.4 (3.3)	–0.2 (3.5)
Smoking status, *n* (%)		
Non-smoker	640 775 (51.6)	342 137 (51.2)
Ex-smoker	215 060 (17.3)	105 411 (15.8)
Current: amount not recorded	26 749 (2.2)	14 151 (2.1)
Light (<10/day)	68 059 (5.5)	35 608 (5.3)
Moderate (10–19/day)	92 337 (7.4)	50 146 (7.5)
Heavy (≥20/day)	50 831 (4.1)	27 711 (4.2)
Smoking not recorded	147 053 (11.9)	92 439 (13.9)
Alcohol status, *n* (%)		
None	325 730 (26.3)	169 033 (25.3)
Trivial <1 unit/day	402 453 (32.4)	203 775 (30.5)
Light 1–2 units/day	192 736 (15.5)	100 051 (15.0)
Moderate or heavy ≥3 units/day	25 003 (2.0)	13 039 (2.0)
Alcohol not recorded	294 942 (23.8)	181 705 (27.2)
Medical and family history, *n* (%)		
Prior cancer	34 324 (2.8)	17 863 (2.7)
Family history of breast cancer	45 621 (3.7)	22 043 (3.3)
Family history of gastrointestinal cancer	18 759 (1.5)	8780 (1.3)
Family history of ovarian cancer	2417 (0.2)	1192 (0.2)
Benign breast disease	41 728 (3.4)	20 687 (3.1)
Chronic pancreatitis	1042 (0.1)	539 (0.1)
Chronic obstructive pulmonary disease	21 516 (1.7)	11 358 (1.7)
Type 1 diabetes	3523 (0.3)	1921 (0.3)
Type 2 diabetes	37 827 (3.0)	20 372 (3.1)
Endometriosis	13 563 (1.1)	7153 (1.1)
Endometrial hyperplasia or polyp	3235 (0.3)	1621 (0.2)
Fibroids	18 796 (1.5)	10 291 (1.5)
Polycystic ovarian disease	10 756 (0.9)	5993 (0.9)
Rheumatoid arthritis	13 825 (1.1)	7153 (1.1)
Systemic lupus erythematosus	1443 (0.1)	687 (0.1)
HIV or AIDS	4186 (0.3)	2907 (0.4)
Oral contraceptive	120 840 (9.7)	61 830 (9.3)
Hormone replacement therapy	26 275 (2.1)	13 402 (2.0)
Anaemia	38 804 (3.1)	19 921 (3.0)

BMI = body mass index. SD = standard deviation.

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. Sensitivity, specificity, positive and negative predictive values were calculated using these thresholds restricting the analyses to females who had any cancer within 2 years or had at least 2 years of follow-up. For comparison, the sensitivity, specificity, positive and negative predictive 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 1 425 518 females aged 25–89 years were identified in the derivation cohort. The following were excluded: 77 335 (5%) without a recorded Townsend deprivation score and 107 319 (8%) with at least one red flag symptom recorded in the 12 months prior to entry to the study, leaving 1 240 864 females for analysis

A total 667 603 females aged 25–89 years were identified in the validation cohort and the following were excluded 55 610 (7%) without a recorded Townsend deprivation score, and 57 991 (7%) with at least one red flag symptom recorded in the 12 months prior to entry to the study, leaving 667 603 females for analysis

Baseline risk factors and symptoms

The baseline characteristics of the derivation and validation cohorts are shown in Table 1 (risk factors). Table 2 shows the frequency of all the red flag and general symptoms in both cohorts at study entry. The most common symptoms were abdominal pain (11%), indigestion (4%), back pain (4%), breast lump (4%), breast pain (3%), cough (4%), tiredness (2%), and rectal bleeding (2%).

Cancer outcomes

There were 23 216 incident cases of cancer arising over 2 years in 1 240 864 females in the derivation cohort and 12 292 cancers from 667 603 females in the validation cohort. The types of cancer are shown in

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

	Derivation cohort, *n* (%)	Validation cohort, *n* (%)
Red flag symptom		
Abdominal distension	5080 (0.4)	2489 (0.4)
Abdominal pain	133 715 (10.8)	70 703 (10.6)
Appetite loss	6130 (0.5)	2793 (0.4)
Breast lump	54 664 (4.4)	28 536 (4.3)
Breast pain	31 050 (2.5)	16 264 (2.4)
Breast skin/nipple changes	2102 (0.2)	1083 (0.2)
Dysphagia	7225 (0.6)	3511 (0.5)
Dyspareunia	7612 (0.6)	3837 (0.6)
Haematemesis	5598 (0.5)	2798 (0.4)
Haematuria	16 749 (1.3)	8640 (1.3)
Haemoptysis	4252 (0.3)	2117 (0.3)
Heartburn	11 981 (1.0)	5913 (0.9)
Indigestion	45 619 (3.7)	23 275 (3.5)
Inter-menstrual bleeding	13 098 (1.1)	6454 (1.0)
Neck lump	5743 (0.5)	2927 (0.4)
Night sweats	3658 (0.3)	1830 (0.3)
Post-coital bleeding	2705 (0.2)	1497 (0.2)
Post-menopausal bleeding	16 561 (1.3)	8578 (1.3)
Rectal bleeding	24 834 (2.0)	12 436 (1.9)
Venous thrombo-embolism	9605 (0.8)	4819 (0.7)
Weight loss	13 491 (1.1)	6874 (1.0)
Recent general symptoms		
Back pain	47 438 (3.8)	25 090 (3.8)
Bruising	2663 (0.2)	1166 (0.2)
Change in bowel habit	3383 (0.3)	1539 (0.2)
Constipation	17 101 (1.4)	8708 (1.3)
Cough	48 677 (3.9)	25 696 (3.8)
Diarrhoea	22 727 (1.8)	11 705 (1.8)
Dyspnoea	10 574 (0.9)	5267 (0.8)
Fever	4114 (0.3)	1788 (0.3)
Headache	28 145 (2.3)	14 899 (2.2)
Itching	3593 (0.3)	1975 (0.3)
Tiredness	27 006 (2.2)	13 169 (2.0)
Nausea	9477 (0.8)	4672 (0.7)
Nocturia	1245 (0.1)	582 (0.1)
Urgency	1532 (0.1)	627 (0.1)
Urinary frequency	6961 (0.6)	3188 (0.5)
Urinary retention	558 (0.0)	278 (0.0)
Urinary incontinence	8977 (0.7)	4627 (0.7)
Vaginal discharge	12 285 (1.0)	6541 (1.0)

Table 3. The five most common cancers in the derivation cohort were breast cancer (36% of all cancers), colorectal cancer (11%), lung cancer (9%), blood cancers (6%), and ovarian cancer (6%). 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 in females having adjusted for all 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 10 symptoms (abdominal pain, anaemia, bruising, change in bowel habit, haematuria, neck lumps, night sweats, post-menopausal bleeding, venous thrombo-embolism, and weight loss). Abdominal pain is associated with nine cancers.

The following symptoms were not included in the final model as they did not meet the inclusion criteria. These were tiredness, back pain, nausea, itching, dyspareunia, dyspnoea, diarrhoea, fever, vaginal discharge, urinary incontinence, urgency, frequency, urinary retention and nocturia.

Table 5 (available at www.qcancer.org) shows the adjusted risk ratios for the final multinomial model incorporating all 11 cancer types. The risk ratios were all adjusted for fractional polynomial terms for age and body mass index.

Venous thrombo-embolism. On multivariate analysis, venous thrombo-embolism was associated with a significant increased risk of all cancers except for renal tract cancers (where the risk was elevated but not significant at the 0.01 level). Venous thrombo-embolism was associated with a 5-fold increase in ovarian cancer risk; 4-fold increase in pancreatic cancer risk; 3-fold increase in risk of cervical, uterine and other cancers; 2 -fold increase in risk of blood, lung, colorectal and gastro-oesophageal cancer; 1.3 fold increase in risk of breast cancer.

General symptoms and anaemia. Appetite loss was associated with an increased risk of seven cancers on multivariate analysis. Weight loss was associated with an increased risk of eight cancers on multivariate analysis: for example a 5-fold increased risk of pancreatic cancer. Night sweats were associated with a 4-fold increased risk of blood cancer. Anaemia was associated with an increased risk of eight cancers.

Abdominal symptoms (dysphagia, rectal bleeding, pain, distension, heartburn and/or indigestion). Dysphagia was associated with increased risk of four cancers: 44-fold increased risk of gastro-oesophageal, 2-fold increased risk of lung and 'other cancers', 3-fold increased risk of pancreatic cancer. Rectal bleeding was associated with a 16-fold increased risk of colorectal cancer.

Abdominal pain was associated with increased risk of all cancers except

Table 3 Numbers (%) women with cancer outcomes in the derivation and validation cohorts

	Derivation cohort		Validation cohort	
	n	%	*n*	%
Total patients	1 240 864		667 603	
No cancer	1 217 648	98.1	655 311	98.2
Any cancer	23 216	1.9	12 292	1.8
Cancer type				
Lung	2043	8.8	1107	9.0
Colorectal	2607	11.2	1356	11.0
Gastro-oesophageal	1065	4.6	551	4.5
Pancreatic	693	3.0	380	3.1
Ovarian	1279	5.5	606	4.9
Renal tract	999	4.3	498	4.1
Breast	8412	36.2	4479	36.4
Blood	1384	6.0	703	5.7
Uterine	1015	4.4	523	4.3
Cervical	437	1.9	222	1.8
Other	3282	14.1	1867	15.2

breast cancer and lung cancer. For example, abdominal pain was associated with a 2-fold increase in cervical cancer risk, a 7-fold increased risk of pancreatic cancer; 6-fold increase in ovarian cancer risk, a 2-fold increase in uterine cancer (see Table 4 [available at www.qcancer.org] for other associations).

Abdominal distension was associated with an increased risk of three cancers: a 19-fold increased risk of ovarian cancer, a 2-fold increase of colorectal, 3-fold increase for other cancers.

Heartburn was associated with a 2-fold increased risk of gastro-oesophageal cancer. Indigestion was associated with six cancers: gastro-oesophageal, pancreas, ovarian, lung, renal tract, and 'other cancers' (Table 4 [available at www.qcancer.org]).

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

Site	Multinomial model ROC (95% CI)	Individual models ROC (95% CI)
Any cancer	0.85 (0.84 to 0.85)	n/a
Lung	0.91 (0.90 to 0.91)	0.92 (0.91 to 0.93)
Colorectal	0.89 (0.88 to 0.90)	0.89 (0.88 to 0.90)
Gastro-oesophageal	0.90 (0.89 to 0.92)	0.89 (0.87 to 0.91)
Pancreas	0.87 (0.85 to 0.89)	0.84 (0.82 to 0.86)
Ovary	0.84 (0.82 to 0.86)	0.84 (0.83 to 0.86)
Renal tract	0.90 (0.89 to 0.92)	0.91 (0.90 to 0.93)
Breast	0.88 (0.87 to 0.88)	n/a
Blood	0.79 (0.77 to 0.80)	n/a
Uterus	0.91 (0.90 to 0.93)	n/a
Cervix	0.73 (0.70 to 0.77)	n/a
Other	0.82 (0.81 to 0.83)	n/a

ROC = receiver operating curve.

Haematemesis. Haematemesis was associated with increased risk of gastro-oesophageal cancer (4-fold); pancreatic cancer (3-fold) and other cancers (1.5-fold).

Haematuria. While the strongest association for haematuria was with renal tract cancers (65-fold increased risk), haematuria was also associated with a 5-fold increased risk of cervical and uterine cancer and a 2-fold increased risk of ovarian cancer, blood cancer and 'other cancers'.

Abnormal vaginal bleeding. Post-menopausal bleeding was associated with increased risk of seven cancers: an 88-fold increase in uterine cancer; 29-fold increase in cervical cancer, a 5-fold increase in ovarian cancer; 4-fold increased risk of bladder cancer; 2-fold increase in breast cancer, blood cancer and 'other cancers'.

Inter-menstrual bleeding was associated with a 7-fold increased risk of cervical cancer and a 6-fold increased risk of uterine cancer.

Post-coital bleeding was associated with a 23-fold increased risk of cervical cancer despite adjustment for other risk factors and symptoms.

Lumps in neck or breast. Neck lumps were associated with an increased risk of three cancers: 19-fold increased risk of blood cancer; 9-fold increased risk of 'other' cancers and a 3-fold increased risk of lung cancer.

A breast lump was associated with a 51-fold increased risk of breast cancer. Breast skin or nipple changes were associated with a 9-fold increased risk of breast cancer. Breast pain was associated with a 2-fold increased risk of breast cancer.

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 0.79 except for cervix (0.73). The highest ROC values were for lung cancer (0.91) and uterine cancer (0.91).

Table 6 also shows the ROC values for the original QCancer models based on published equations for each separate

Table 8. Comparison of strategies to identify females 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	0.38	600 534	309	65 962	798	72.1	90.1	1.2	99.9
Colorectal cancer	0.35	600 412	431	65 835	925	68.2	90.1	1.4	99.9
Gastro-oesophageal cancer	0.14	600 705	138	66 347	413	75.0	90.1	0.6	100.0
Pancreas cancer	0.12	600 721	122	66 502	258	67.9	90.0	0.4	100.0
Ovarian cancer	0.18	600 610	233	66 387	373	61.6	90.0	0.6	100.0
Renal cancer	0.1	600 727	116	66 378	382	76.7	90.0	0.6	100.0
Breast cancer	0.72	599 414	1,429	63 710	3,050	68.1	90.4	4.6	99.8
Blood cancer	0.22	600 449	394	66 451	309	44.0	90.0	0.5	99.9
Uterine cancer	0.1	600 758	85	66 322	438	83.7	90.1	0.7	100.0
Cervical cancer	0.05	600 742	101	66 639	121	54.5	90.0	0.2	100.0
Other cancer	0.55	600 590	253	66 699	61	19.4	90.0	0.1	100.0

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

cancer outcome where available.[4–9] Generally the ROC values were very similar 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 within each model tenth except for 'other cancer' 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 values 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: breast lump (11%), haemoptysis (8%), dysphagia (8%), and post-menopausal bleeding (7%). The positive predictive value for anaemia was 6% and for venous thrombo-embolism was 5%.

Table 7 (available at www.qcancer.org) also shows the sensitivity, specificity, positive and negative predictive values for predicting cancer based on three risk thresholds. The 90th centile defined a high-risk group with a risk score for any cancer of >4.1%. The positive predictive power was 10%, the sensitivity was 54% and the specificity 91%. The 95th centile defined a high risk group with a cancer risk score of >6.9%. The positive predictive power was 14%, the sensitivity was 39% and the specificity 96%. The 99th centile defined a high risk group with a cancer risk score of >19.2%. The positive predictive power was 27%, the sensitivity was 14.7% and the specificity 99.3%.

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 breast cancer defined a high risk group with a risk score of >0.72%. The positive predictive power was 4.6%, the sensitivity 68% and the specificity 90.4%. Clinical examples of the algorithm are shown in Box 1.

Box 1. Clinical examples

- A 69-year-old female who is a light drinker and an ex-smoker. She has a breast lump and nipple discharge or breast skin changes. Her overall cancer risk is 73.9%, of which 73.3% is due to risk of breast cancer.
- A 50-year-old female, non-drinker, non-smoker with haematuria and post-menopausal bleeding has an 11.4% overall cancer risk, comprising uterus (5.7%), renal (2.6%), cervix (2.0%), ovary (0.4%), and other cancer (0.7%).
- A 68-year-old female, non-drinker, non-smoker with anaemia, abdominal distension, abdominal pain and recent constipation has an overall cancer risk of 37.8%, comprising ovary (25.8%), colorectal (5.6%), other cancer (4.3%), blood (0.8%), gastro-oesophageal (0.3%), renal (0.2%), breast (0.3%), pancreas (0.2%), lung (0.1%), uterus (0.1%), and cervix (0.1%).
- A 70-year-old female, heavy-smoker, trivial drinker, with family history of gastrointestinal cancer, night sweats, a lump in the neck has a 18.2% risk of any cancer, comprising blood (10.0%), lung (3.0%), other cancer (3.9%), breast (0.5%), colorectal (0.3%), gastro-oesophageal (0.2%), renal tract (0.1%), ovary (0.1%), and pancreas (0.1%).
- A 32-year-old female, non-smoker, heavy drinker with loss of appetite and abdominal pain has a 0.4% risk of any cancer and a 99.6% risk of no cancer.

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 women. The algorithm is based on a combination of symptoms and risk factors such as age and family history of cancer which the woman 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 11 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 women suitable for targeted investigation or a subset of particularly low-risk women for whom reassurance may be appropriate.

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.[25] 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 may 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.[26,27] In future, it is likely that QResearch will be linked to cancer registry data which will allow further validation of outcomes. Recording of symptoms may be less complete or accurate than diagnostic codes since patients may not visit their GP with mild symptoms, may not report all symptoms to their GP when they do consult, or GPs may 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 may 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.[14,18,19,28–30] Compared with the CAPER studies,[14] this study is larger and nationally, rather than locally, based and has the potential to be updated as populations change, data quality improves and requirements evolve. QCancer applies to a broader age range of patients (25–89 years) whereas CAPER can only be used in patients aged ≥40 years. This wider age range is an important advantage of the present study given the incidence of alarm symptoms in younger patients and the cancers which occur in younger patients, particularly cervix, breast and haematological malignancies. Unlike the CAPER studies and subsequent work by Hamilton *et al*,[28] the QCancer algorithm includes important risk factors such as sex, age,[18] family history, and anaemia alongside symptoms, which allows an individualised measure of risk for each type of cancer and for cancer overall.

Breast cancer. This study has 8412 breast cancer cases in the derivation cohort and 4479 in the validation cohort which is much larger than a recent study by McGowan *et al* to develop a clinical prediction rule for the diagnosis of breast cancer.[31] This prediction rule was derived from a secondary care setting and only had 59 cases in the derivation cohort and five in the validation cohort which the authors recognise was likely to be under-powered.[31] While both the current study and McGowan's found breast lump and skin tethering and increasing age were predictive of a diagnosis of breast cancer, this study's risk ratios were substantially higher for a breast lump (risk ratio = 51 compared with odds ratio = 15) although similar for skin tethering (risk ratio =9 compared with odds ratio = 8). Unlike the McGowan study, it was found that alcohol, family history of breast cancer, breast pain, post-menopausal bleeding, increasing affluence, and venous thrombo-embolism were independently predictive of a diagnosis of breast cancer. Some of these factors were tested in the McGowan study but were not significant which may reflect the very small sample size. While it is suspected that a risk score for breast cancer may not affect the decision to refer a woman with a breast lump, it could be useful for alerting the clinician to risk of breast cancer in women with breast pain and for the quantification and communication of risk with the patient. It may also be useful to include the risk assessment in a referral letter to help with prioritisation and future investigation once in the hospital setting.

Gynaecological cancer, vaginal bleeding, and haematuria. To the authors' knowledge this is the first study to develop a predictive algorithm that can quantify absolute risks of uterine cancer and cervical cancer in primary care. Post-menopausal bleeding is traditionally considered to be a sign of uterine cancer and haematuria to be a sign of renal tract cancer. These results support these assumptions with similar positive predictive powers to those reported elsewhere.[15] However, this study found post-menopausal bleeding was also predictive of other cancers apart from uterine cancer including, cervix, ovary, breast, bladder, and blood cancer. Similarly, haematuria was predictive of ovarian, cervix, uterine, blood cancer in addition to renal tract cancer. It is possible that these associations represent the progress of invasive disease (for example, uterine cancer invading the bladder and causing haematuria). Alternatively, it could be due to women misidentifying haematuria as vaginal bleeding or vice versa. Either way, this study suggests that clinicians and guidelines need to include the possibility of multiple types of cancer in the presence of such alarm symptoms and accordingly take a detailed history, pelvic, and breast examination and full blood count to inform further investigation or referral.

Haematological cancers. To the authors knowledge this is the first study to develop a predictive algorithm that can quantify absolute risk of blood cancer (leukaemia, lymphoma and myeloma) in primary care. The following symptoms were found to be predictive of blood cancers including anaemia, abdominal pain, haematuria, neck lumps, night sweats, venous thrombo-embolism, weight loss, change in bowel habit, and bruising. The majority of these features are included within the NICE guidelines for suspected cancer, although not all (venous thrombo-embolism, change in bowel habit).

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 as post-menopausal bleeding and haematuria, previously thought to map to one main cancer each, actually map to multiple types of cancer. However, other symptoms currently included in NICE guidelines, such as tiredness, itching and fever were not significant independent predictors in this analysis. Similarly, symptoms such as appetite loss, breast pain 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, this algorithm better accounts for age than the NICE guideline

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/female

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** which simply dichotomises patients into those aged <50 or ≥50 years.[11] This is relevant since the risk of cancer generally increases steeply with age. This study also quantified the risk associated with family history of cancer (where relevant) and incorporated it into the underlying algorithm so that it contributes to a patient's 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 years.[17] 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 (including mammograms and abdominal-pelvic CT scan).[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 gynaecological and abdominal malignancies. This tool will enable clinicians to quantify the risks of each cancer for women 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 cohort 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.

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. 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.
15. Shapley M, Mansell G, Jordan JL, Jordan KP. Positive predictive values of ≥5% in primary care for cancer: systematic review. *Br J Gen Pract* 2010; DOI: 10.3399/bjgp10X515412.
16. 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.
17. National Institute for Health Clinical Excellence. *Venous thromboembolic diseases: the management of venous thromboembolic disease and the role of thrombophilia testing.* NICE guidance no. CG144. London: NICE, 2012.
18. 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.
19. 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.
20. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7(2):** 147–177.
21. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Clin Epidemiol* 2007; **60(9):** 979.
22. Moons KGM, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006; **59(10):** 1092–1101.
23. Rubin DB. *Multiple imputation for non-response in surveys.* New York, NY: John Wiley, 1987.
24. 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.
25. 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.
26. 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.
27. 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.
28. Hamilton W, Peters TJ, Bankhead C, Sharp D. Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. *BMJ* 2009; **339:** b2998.
29. National Institute for Health and Clinical Excellence. *Ovarian cancer: the recognition and initial management of ovarian cancer.* London: NICE, 2011.
30. Goff BA, Mandel LS, Melancon CH, Muntz HG. Frequency of symptoms of ovarian cancer in women presenting to primary care clinics. *JAMA* 2004; **291(22):** 2705–2712.
31. McCowan C, Donnan PT, Dewar J, *et al.* Identifying suspected breast cancer: development and validation of a clinical prediction rule. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X572391.