Julia Hippisley-Cox and Carol Coupland

Identifying patients with suspected gastro-oesophageal cancer in primary care:

derivation and validation of an algorithm

Abstract

Background
Gastro-oesphageal is one of the most common cancers worldwide. Evidence suggested that increased awareness of symptoms and earlier diagnosis could help improve treatment options and improve survival.

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

Design and setting
Cohort study of 375 UK QResearch® general practices for development, and 189 for validation.

Method
Included patients were aged 30–84 years, free at baseline of a diagnosis of gastro-oesophageal cancer, and without dysphagia, haematemesis, abdominal pain, appetite loss, or weight loss recorded in previous 12 months. The primary outcome was incident diagnosis of gastro-oesophageal cancer recorded in the next 2 years. Risk factors examined were age, body mass index, alcohol status, smoking status, deprivation, family history of gastrointestinal cancer, dysphagia, previous diagnosis of cancer apart from gastro-oesophageal cancer, haematemesis, abdominal pain, appetite loss, weight loss, tiredness, and anaemia. Cox proportional hazards models were used to develop risk equations. Measures of calibration and discrimination assessed performance in the validation cohort.

Results
There were 2527 incident cases of gastro-oesophageal cancer from 4.1 million person-years in the derivation cohort. Independent predictors were age, smoking, dysphagia, haematemesis, abdominal pain, appetite loss, weight loss, and anaemia. On validation, the algorithms explained 71% of the variation in females and 73% in males. The receiver operating curve statistics were 0.89 (females) and 0.92 (males). The D statistic was 3.2 (females) and 3.3 (males). The 10% of patients with the highest predicted risks included 77% of all gastro-oesophageal cancers diagnosed over the next 2 years.

Conclusion
The algorithm has good performance and could potentially be used to help identify those at highest risk of gastro-oesophageal cancer, to facilitate early referral and investigation.

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

INTRODUCTION

Gastro-oesophageal cancer is one of the most common cancers worldwide.[1] Evidence suggests that increased awareness of symptoms and earlier diagnosis could help improve treatment options and improve 5-year survival.[2] The National Awareness and Early Diagnosis Initiative (NAEDI) in the UK aims to make the public more aware of the signs and symptoms of cancer, and encourage those with symptoms to seek advice earlier.[3] It has been estimated that such an approach might save 5000 lives without any new medical advances.[4]

'Red flag' or 'alarm' symptoms such as haematemesis, dysphagia, weight loss, appetite loss, or abdominal pain might herald an existing condition of gastro-oesophageal cancer.[5] However, much of the information available to guide decision making is based on data from secondary care.[5–7] These studies demonstrate a wide variation in the sensitivity and specificity of alarm symptoms for upper gastrointestinal malignancies. More recent studies from primary care demonstrate that an approach focused on single symptoms alone such as dysphagia is likely to miss 40% of current gastro-oesophageal cancers.[8] A variety of factors, therefore, need to be combined to develop an algorithm to help clinicians better assess and prioritise patients at high risk of having gastro-oesophageal cancer for further investigation or referral.

It was decided to develop and validate an algorithm to estimate the individualised absolute risk of having gastro-oesophageal cancer incorporating both symptoms and baseline risk factors, to help identify those at highest risk for further investigation or referral. QResearch® (a large UK 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 and has been used successfully to develop and validate a range of prognostic models for use in primary care,[9–12] including a similar model to help detect lung cancer.[13] Once validated, the models could be integrated into clinical computer systems to help systematically identify those at high risk and alert clinicians to those who might benefit most from further assessment or interventions.[9–12] It could also be made available on the internet as a simple calculator for use by the general population to help support NAEDI.[3]

METHOD

Study design and data source

A prospective cohort study was carried out in a large population of primary care patients, using the QResearch database (version 30). All practices in England and Wales that had been using their EMIS (Egton Medical Information System) 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

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

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

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

Submitted: 21 June 2011; **Editor's response:** 12 July 2011; **final acceptance:** 19 July 2011.

This is the full-length article (published online 31 Oct 2011) of an abridged version published in print. Cite this article as: **Br J Gen Pract 2011; DOI: 10.3399/bjgp11X606609.**

How this fits in

Gastro-oesophageal cancer is one of the most common cancers worldwide. Evidence suggests that increased awareness of symptoms and earlier diagnosis could help improve treatment options and improve 5-year survival. 'Alarm' symptoms such as haematemesis, dysphagia, weight loss, appetite loss, or abdominal pain might herald an existing condition of gastro-oesophageal cancer but an approach focused on single symptoms alone such as dysphagia is likely to miss 40% of current gastro-oesophageal cancers. A simple algorithm based on age, smoking, dysphagia, haematemesis, abdominal pain, appetite loss, weight loss and anaemia was developed and validated to estimate absolute risk of a patient having gastro-oesophageal cancer in primary care. The algorithm has good discrimination and calibration and could be integrated into clinical computer systems to help identify those at highest risk for early referral and investigation

dataset. An open cohort of patients aged 30–84 years was identified, drawn from patients registered with practices between 1 January 2000 and 30 September 2010. The following were excluded: patients without a postcode-related Townsend score, those with a history of gastro-oesophageal cancer at baseline, and those with a recorded 'red flag symptom' in the 12 months prior to the study entry date, that is, symptoms of dysphagia, haematemesis, loss of appetite, weight loss, and abdominal pain, which might indicate gastro-oesophageal cancer.

Entry to the cohort was the latest of the study start date (1 January 2000), 12 months after the patient registered with the practice, and, for those patients with incident dysphagia, haematemesis, loss of appetite, weight loss, or abdominal pain, the date of first recorded onset within the study period. Where patients had new onset of multiple symptoms recorded, the entry date was the earliest recorded date of the new symptoms in the study period. Other symptoms were included if they occurred within a 60-day period of the entry date and before the diagnosis of gastro-oesophageal cancer or the date on which the patient left or died, or the study ended.

Clinical outcome definition

The study outcome was current gastro-oesophageal cancer, which was defined as incident diagnosis of either gastric cancer or oesophageal cancer during the 2 years after study entry recorded on either (a) the patient's GP record using the relevant UK diagnostic codes, or (b) their linked Office for National Statistics (ONS) cause-of-death record using the relevant International Classification of Diseases (ICD)-9 codes (150 or 151) or ICD-10 diagnostic codes (C15 or C16). A 2-year period was used, since this represents the period of time during which existing gastro-oesophageal cancers are likely to become clinically manifest.[8,14] It was assumed that where gastro-oesophageal cancer deaths occurred within 2 years, without a recorded diagnostic code in the GP record, the cancer would have been present at the start of the 2-year period.

Predictor variables

Established predictor variables were examined, focusing on those that are likely to be recorded in the patient's electronic record and that the patient is likely to know. 'Red flag' symptoms that might herald a diagnosis of gastro-oesophageal cancer were also included. For the purposes of this study, 'red flag' symptoms were defined as first onset of haematemesis, dysphagia, loss of appetite, weight loss, or abdominal pain. Separate analyses were carried out in males and females. The predictor variables examined were:

- currently consulting a GP with first onset of dysphagia (yes/no);
- currently consulting a GP with first onset of haematemesis (yes/no);
- currently consulting a GP with first onset of loss of appetite (yes/no);
- currently consulting a GP with first onset of weight-loss symptom (yes/no);
- currently consulting a GP with first onset of abdominal pain (yes/no);
- recently consulted a GP with tiredness in past 12 months (yes/no);
- age at baseline (continuous, ranging from 30 to 84 years);
- body mass index (continuous);
- smoking status (non-smoker; ex; light [1–9 cigarettes/day]; moderate [10–19 cigarettes/day]; heavy smoker [≥20 cigarettes/day]);
- alcohol status (non-drinker; trivial [<1 unit/day]; light [1–2 units/day]; moderate/heavy [≥3 units/day]);
- Townsend deprivation score (continuous);
- family history of gastrointestinal cancer (yes/no);
- previous diagnosis of cancer apart from

gastro-oesophageal cancer; and

- anaemia defined as recorded haemoglobin <11 g/dl in past 12 months (yes/no).

Derivation and validation of the models

The risk-prediction algorithm was developed and validated using established methods.[9–12,15–17] Multiple imputation was used to replace missing values for body mass index, and alcohol and smoking status, and these values were used in the main analyses.[18–21] Five imputations were carried out. Cox's proportional hazards models were used to estimate the coefficients for each risk factor for males and females separately, using robust variance estimates to allow for the clustering of patients within general practices. Rubin's rules were used to combine the results across the imputed datasets.[22] Fractional polynomials were used to model non-linear risk relationships with continuous variables.[23] A full model was fitted initially, and variables were retained if they had a hazard ratio of <0.80 or >1.20 (for binary variables) and were statistically significant at the 0.01 level. Interactions between predictor variables and age were examined and included in the final models if they were statistically significant at the 0.01 level.

The regression coefficients for each variable from the final model were used as weights, which were combined with the baseline survivor function evaluated at 2 years to derive absolute risk equations for 2 years of follow-up.[24] The baseline survivor function was estimated, based on zero values of centred continuous variables, with all binary predictor values set to zero, using the methods implemented in STATA.

Multiple imputation was used in the validation cohort to replace missing values for body mass index and alcohol and smoking status. The risk equations for males and females obtained from the derivation cohort were then applied to the validation cohort and measures of discrimination were calculated. R^2 (estimated variation in time to gastro-oesophageal cancer[25]), the D statistic[26] (a measure of discrimination where higher values indicate better discrimination), and the area under the receiver operating characteristic curve (ROC) statistic at 2 years were calculated. Calibration was assessed by comparing the mean predicted risks at 2 years with the observed risk by tenth of predicted risk. The observed risks were obtained using Kaplan–Meier estimates evaluated at 2 years.

The validation cohort was used to define the thresholds for the 0.5%, 1%, 5%, and 10% of patients at highest estimated risk of gastro-oesophageal cancer at 2 years. Sensitivity, specificity, and positive and negative predictive values were calculated using these thresholds, restricting the analyses to patients who had the outcome within 2 years or had at least 2 years of follow-up.

All the available data on the database were used to maximise the power and generalisability of the results. STATA (version 11) was used for all analyses.

Table 1. Baseline characteristics of patients in the derivation and validation cohorts; patients are free of a diagnosis of gastro-oesophageal cancer at baseline. (Figures are *n* [%] unless otherwise specified)

Characteristic	Derivation cohort (*n* = 2 355 719)	Validation cohort (*n* = 1 238 971)
Female	1 174 921 (49.9)	617 493 (49.8)
Male	1 180 798 (50.1)	621 478 (50.2)
Mean age (SD), years	50.1 (15.0)	50.1 (15.0)
Mean Townsend score (SD)	−0.3 (3.4)	−0.2 (3.6)
BMI recorded prior to study entry	1 869 779 (79.4)	1 005 846 (81.2)
Mean BMI (SD), kg/m^2	26.4 (4.6)	26.4 (4.7)
Smoking status		
Non-smoker	1 194 692 (50.7)	624 788 (50.4)
Ex-smoker	427 246 (18.1)	229 516 (18.5)
Current smoker, amount not recorded	71 416 (3.0)	39 231 (3.2)
Light smoker (<10/day)	148 063 (6.3)	79 844 (6.4)
Moderate smoker (10–19/day)	179 931 (7.6)	95 754 (7.7)
Heavy smoker (≥20/day)	133 980 (5.7)	73 554 (5.9)
Smoking status not recorded	200 391 (8.5)	96 284 (7.8)
Alcohol status		
None	511 397 (21.7)	275 795 (22.3)
Trivial (<1 unit/day)	657 782 (27.9)	356 394 (28.8)
Light (1–2 units/day)	493 275 (20.9)	257 800 (20.8)
Moderate or heavy (≥3 units/day)	176 350 (7.5)	93 310 (7.5)
Alcohol status not recorded	516 915 (21.9)	255 672 (20.6)
Medical history		
Family history of gastrointestinal cancer	29 636 (1.3)	17 742 (1.4)
Prior cancer apart from gastro-oesophageal cancer	53 971 (2.3)	28 520 (2.3)
Current symptoms and symptoms in the preceding year		
Current dysphagia	15 021 (0.6)	8165 (0.7)
Current haematemesis	12 952 (0.5)	7119 (0.6)
Current abdominal pain	225 543 (9.6)	126 161 (10.2)
Current appetite loss	9978 (0.4)	6133 (0.5)
Current weight loss	9998 (0.4)	5377 (0.4)
Tiredness in last year	25 200 (1.1)	14 119 (1.1)
Haemoglobin recorded in the last year	22 576 (1.0)	12 638 (1.0)
Haemoglobin <11 g/dl in the last year	406 410 (17.3)	218 862 (17.7)

BMI = body mass index. SD = standard deviation. Figures in the tables are counts (%) unless otherwise specified.

RESULTS

Overall study population

Overall, 564 QResearch practices in England and Wales met the study inclusion criteria, of which 375 were randomly assigned to the derivation dataset, with the remainder assigned to a validation cohort.

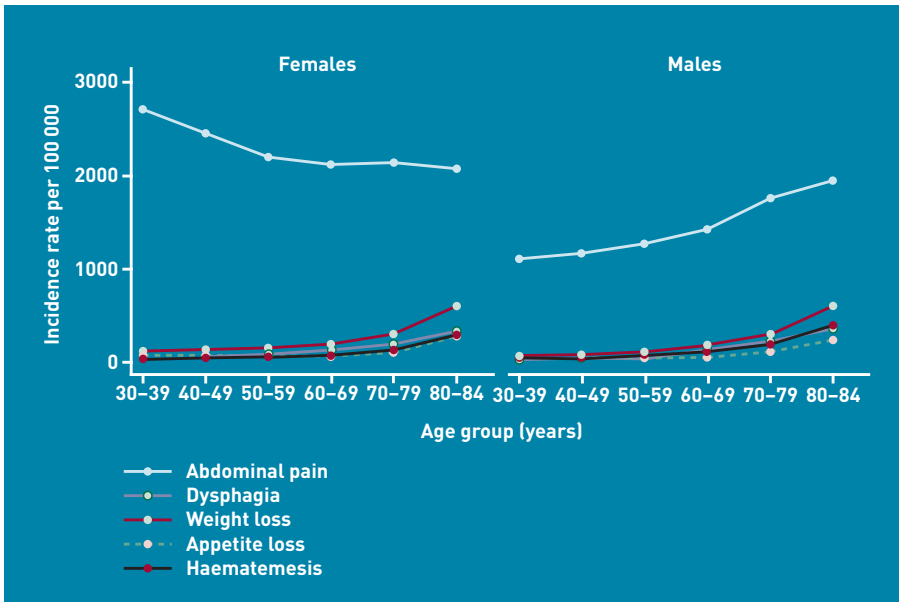

Figure 1. Incidence rates of dysphagia, haematemesis, appetite loss, weight loss, and abdominal pain in males and females per 100 000 person-years in the derivation cohort.

Of these a total of 2 538 615 patients aged 30–84 years were identified in the derivation cohort, and 124 458 patients (4.9%) without a recorded Townsend deprivation score were excluded; 839 (0.03%) patients with a history of gastro-oesophageal cancer were excluded, as well as a further 57 599 (2.3%) patients with at least one red flag symptom recorded in the 12 months prior to entry to the study at baseline, leaving 2 355 719 patients for analysis

A total of 1 342 329 patients aged 30–84 years were identified in the validation cohort; of these, 70 847 patients (5.3%) without a recorded Townsend score were excluded, as well as 538 (0.04%) with a history of gastro-oesophageal cancer, and 31 973 (2.4%) with at least one red flag symptom recorded in the 12 months prior to study entry, leaving 1 238 971 patients for analysis.

The baseline characteristics of each cohort were very similar, as shown in Table 1. As in previous studies,[9,11,27] the patterns of missing data supported the use of multiple imputation to replace missing values for alcohol and smoking status, and body mass index (not shown, available from the authors).

Table 2. Adjusted hazard ratios[a] (95% CI) for the final model for gastro-oesophageal cancer for males and females in the derivation cohort[b]

	Adjusted hazard ratios for women (95% CI)	Adjusted hazard ratios for men (95% CI)
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.33 (1.11 to 1.6)	1.38 (1.22 to 1.57)
Light smoker	1.96 (1.43 to 2.68)	1.89 (1.51 to 2.37)
Moderate smoker	2.51 (1.93 to 3.26)	2.18 (1.77 to 2.67)
Heavy smoker	3.11 (2.26 to 4.28)	2.00 (1.52 to 2.63)
Current symptoms and anaemia		
Current dysphagia[c]	131 (97.5 to 175.0)[d]	143 (108 to 189)[d]
Current abdominal pain[c]	4.74 (3.54 to 6.33)[d]	3.78 (3.32 to 4.30)
Current appetite loss[c]	10.0 (5.28 to 19.0)[d]	3.87 (2.82 to 5.32)
Current haematemesis[c]	25.2 (14.4 to 44.2)[d]	7.62 (6.08 to 9.55)
Current weight loss[c]	3.97 (3.06 to 5.16)	5.64 (4.67 to 6.81)
Haemoglobin <11 g/dl in last year[c]	2.32 (1.84 to 2.93)	1.79 (1.44 to 2.23)

[a]Hazard ratios were adjusted for all other terms in the table and also for age. [b]Models included fractional polynomial terms for age: for women the term was age$^{0.5}$; for men the terms were age^{-2}, age^{3}. The model for women also included interactions between the age term and dysphagia, abdominal pain, appetite loss, and haematemesis. The model for men included interactions between the age terms and dysphagia. [c]Compared with a person without this characteristic. [d]Hazard ratio evaluated at the mean age.

Incidence rates for red flag symptoms

Overall, in the derivation cohort, 15 021 patients were identified with incident dysphagia, 12 952 with haematemesis, 9978 with appetite loss, 9998 with weight loss, and 225 543 with abdominal pain; 4203 patients had multiple recorded symptoms. Figure 1 shows the age–sex incidence rates of each symptom. The incidence rates for dysphagia, haematemesis, appetite loss, and weight loss were similar in males and females, and increased steeply with age. Abdominal pain was more common in females and tended to decrease with age in females and increase with age in males.

Incidence rates of gastro-oesophageal cancer

Overall in the derivation cohort, during the 2-year follow-up period, a total of 2527 incident cases of gastro-oesophageal cancer arising from 4 122 629 person-years of observation were identified, giving a rate of 61 per 100 000 person-years. Of the 2527 incident cases, 1531 (60.6%) were oesophageal cancer and 996 (39.4%) gastric cancer. There were 2080 cases (82.3% of 2527) identified using the GP record and an additional 447 (17.7%) identified from the linked death record.

In the validation cohort, 1343 incident cases of gastro-oesophageal cancer were identified, arising from 2 169 715 person-years of observation, giving a rate of 62 per 100 000 person-years. Of these, 776 (57.8%) were oesophageal cancer and 567 (42.2%) were gastric cancer. There were 1126 cases (83.8% of 1339) identified using the GP record and an additional 217 (16.2%) from the linked death record.

Predictor variables

Table 2 shows the predictor variables selected for the final models for females and males. The predictors for both males and females were age, smoking status, body mass index, dysphagia,

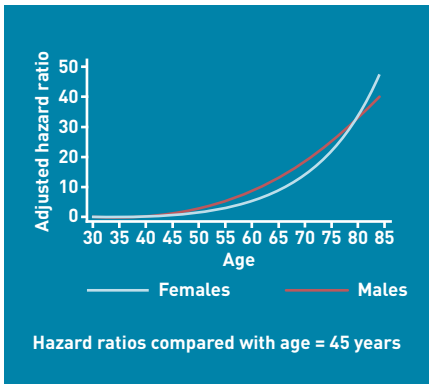

Figure 2. Hazard ratios for gastro-oesophageal cancer by age.

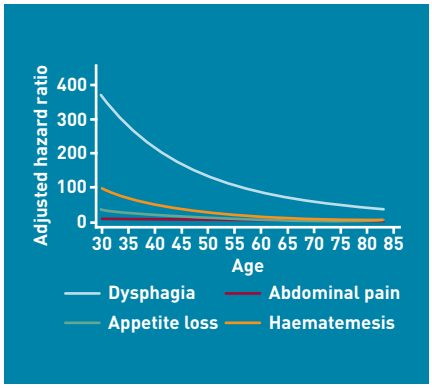

Figure 3. Hazard ratios for symptoms for gastro-oesophageal cancer by age in females.

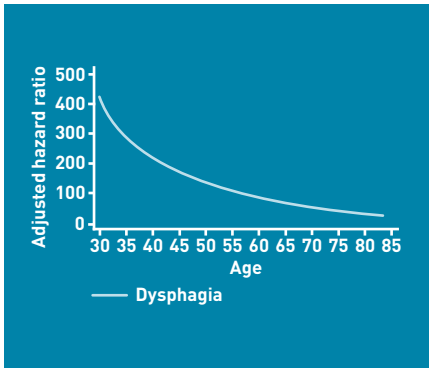

Figure 4. Hazard ratios for symptoms for gastro-oesophageal cancer by age in males.

haematemesis, abdominal pain, appetite loss, weight loss, and anaemia. The other variables examined were not independent risk factors in females or males, so were not included in the final models.

The risk of gastro-oesophageal cancer in females was significantly associated with increasing age (as shown in Figure 2) and the amount smoked. Compared with non-smokers, the risks were increased by 3.1-fold for heavy smokers, 2.5-fold for moderate smokers, 2.0-fold for light smokers and 1.3-fold for ex-smokers. The risks were also elevated in females with weight loss (4.0-fold higher) and anaemia (2.3-fold higher). There were significant interactions between age and four symptoms (dysphagia, abdominal pain, appetite loss, and haematemesis) as shown in Figure 3. For each symptom, the relative effect was more marked at younger ages and the most marked effect was observed for dysphagia. At the mean age in females, dysphagia was associated with a 131.0-fold higher risk, haematemesis with a 25.2-fold higher risk, abdominal pain with a 4.7-fold higher risk, and appetite loss with a 10.0-fold higher risk.

The magnitudes of the hazard ratios in males were similar to those found for females, as shown in Table 2, except that there was less of a gradient with smoking. There was a significant interaction between age and dysphagia, as shown in Figure 4.

Validation

The validation statistics (Table 3) showed that the risk-prediction equations explained 71% of the variation in time to diagnosis in females and 73% in males. The D statistic was 3.22 for females and 3.32 for males. The ROC statistics were 0.89 for females and 0.92 for males.

Figure 5 shows the mean predicted scores and the observed risks at 2 years within each tenth of predicted risk, in order to assess the calibration of the model in the validation cohort. Overall, the model was well calibrated. There was close correspondence between predicted and observed 2-year risks within each model tenth for males and females, with a small degree of over-prediction in the highest tenth in males and females.

Individual risk assessment and thresholds

One potential use for this algorithm is as a web calculator within consultations, with individual patients presenting with new onset of dysphagia, haematemesis, abdominal pain, weight loss or appetite loss, or anaemia. Some clinical examples are shown in Box 1. The results will quantify the risk of gastro-oesophageal cancer, which can be used to inform the urgency of further investigations such as gastroscopy or barium swallow. The web calculator could also be used by patients to prompt attendance at their GP.

The algorithm could also be used for systematic risk stratification for a population of patients aged 30–84 years. Software implementing the algorithm could calculate the risk of a patient having an existing, but as yet undiagnosed, gastro-oesophageal cancer, based on information already recorded in the patient's electronic

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

Statistic	Mean (95% CI)
Females	
R^2 statistic,[a] %	71.2 (69.2 to 73.2)
D statistic[b]	3.22 (3.06 to 3.37)
ROC statistic[c]	0.89 (0.87 to 0.91)
Men	
R^2 statistic,[a] %	72.5 (71.1 to 73.8)
D statistic[b]	3.32 (3.21 to 3.43)
ROC statistic[c]	0.92 (0.91 to 0.93)

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

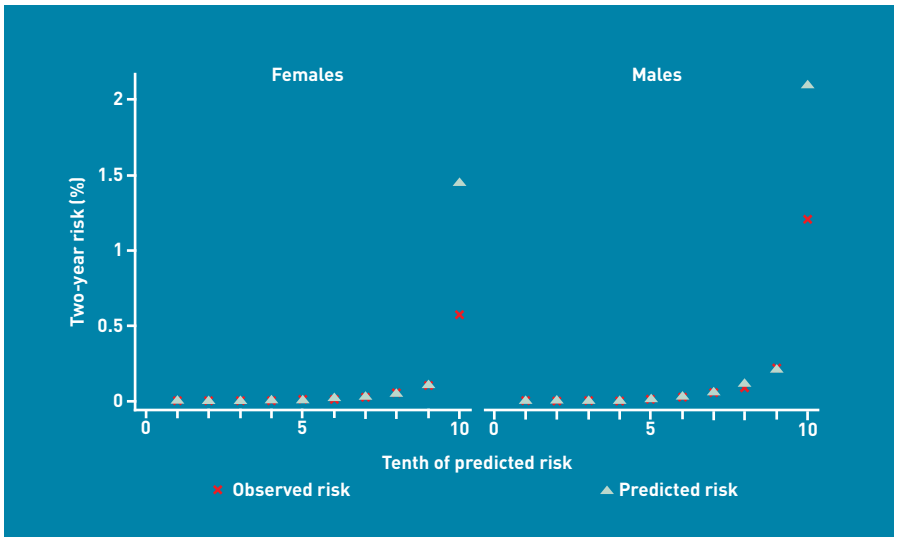

Figure 5. Mean predicted risk and observed risk of gastro-oesophageal cancer over 2 years by tenth of predicted risk applying the risk-prediction scores to the validation cohort.

Box 1. Clinical examples

- A 45-year-old male who is a non-smoker with dysphagia and no other symptoms has an estimated risk of gastro-oesophageal cancer of 1.4%. If he has also had anaemia in the last year and has abdominal pain, the estimated risk is 9%. If he has these symptoms and also loss of appetite, the estimated risk of gastro-oesophageal cancer is 31%.
- A 50-year-old female who is a heavy smoker with dysphagia has an estimated risk of gastro-oesophageal cancer of 3%. If she has also had anaemia in the last year, her estimated risk is 7%, and if she also has abdominal pain, the estimated risk of gastro-oesophageal cancer is 29%.
- A 70-year-old female who is an ex-smoker with abdominal pain, appetite loss, and weight loss but no dysphagia, has an estimated risk of gastro-oesophageal cancer of 3%. If she has also had anaemia in the last year, her estimated risk is 8%. If she also has haematemesis, her estimated risk of gastro-oesophageal cancer is 52%.

health record. Patients at highest risk could be identified for a clinical assessment.

The 90th centile defined a high-risk group with a 2-year risk score of >0.2% (Table 4). There were 1028 new cases of gastro-oesophageal cancer within this group, out of 1339 new cases identified in the validation cohort, which accounted for 77% of all new cases of gastro-oesophageal cancer (sensitivity). The positive predictive value (PPV) with this threshold was 1.2%. Alternatively, using a threshold based on the top 1% of risk (that is, a risk score >2.1%) had a sensitivity of 40% and a PPV of 7.7%. In contrast, the PPV of dysphagia alone was 7.8%, and only 32% of gastro-oesophageal cancers occurred in patients with a first onset of dysphagia. In other words, the sensitivity of an approach based only on dysphagia as a single symptom is low, since approximately 68% of cases of gastro-oesophageal cancer cases would be missed. Similarly, the PPV of anaemia on its own as a symptom is 1.1% and the sensitivity is 8.9%.

DISCUSSION

Summary

This research has developed and validated a new algorithm designed to quantify the absolute risk of having existing, but as yet undiagnosed, gastro-oesophageal cancer, based on a combination of symptoms and simple variables that the patient is likely to know or that can be easily ascertained. The algorithm, which included eight variables — age, smoking status, dysphagia, abdominal pain, appetite loss, haematemesis, weight loss, and anaemia — performed well in a separate validation sample, with good discrimination and calibration.

Strengths and limitations

Key strengths of the study include the size, 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.[28]

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

Criteria	Risk threshold %	True negative[a]	False negative[b]	False positive[c]	True positive[d]	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Current dysphagia	n/a	956 541	909	5156	434	32.3	99.5	7.8	99.9
Current haematemesis	n/a	957 321	1242	4376	101	7.5	99.5	2.3	99.9
Current abdominal pain	n/a	870 379	1034	91 318	309	23.0	90.5	0.3	99.9
Current appetite loss	n/a	958 441	1308	3256	35	2.6	99.7	1.1	99.9
Current weight loss	n/a	952 634	1236	9063	107	8.0	99.1	1.2	99.9
Anaemia	n/a	951 467	1224	10 230	119	8.9	98.9	1.1	99.9
Top 10% of risk	0.2	875 463	315	86 234	1028	76.5	91.0	1.2	100.0
Top 5% of risk	0.4	923 347	465	38 350	878	65.4	96.0	2.2	99.9
Top 1% of risk	2.1	955 297	812	6400	531	39.5	99.3	7.7	99.9
Top 0.5% of risk	5.7	958 399	942	3298	401	29.9	99.7	10.8	99.9

n/a = not applicable. [a]Criterion not met does not have disease. [b]Criterion not met does have disease. [c]Criterion met does not have disease. [d]Criterion met does have disease.

Funding

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

Ethical approval

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

Provenance

Freely submitted; externally peer reviewed.

Web calculator

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

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 consultant statistician for ClinRisk Ltd. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations.

Acknowledgements

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

Discuss this article

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

The authors consider this 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, and confirms established associations with smoking and symptoms. The algorithms have been developed in one cohort and validated in a separate cohort that is representative of the patients likely to be considered for preventative measures. The ROC values were 0.89 in females and 0.92 in males.

Limitations include a lack of formally adjudicated outcomes, potential information bias, and missing data. The study database has linked cause of death from the UK Office for National Statistics, and the study is therefore likely to have picked up the majority of cases of gastro-oesophageal cancer, thereby minimising ascertainment bias. Patients who die of gastro-oesophageal cancer will be included on the linked cause-of-death data. Patients diagnosed with gastro-oesophageal 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 incidence rate of gastro-oesophageal cancer in the study population was close to published UK data,[1] indicating that ascertainment of cases is likely to be good. While the study is reliant on the accuracy of information recorded by primary care physicians, the authors think that 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.[29,30] For example, one systematic review reported that, on average, 89% of diagnoses recorded on the GP electronic record are confirmed from other data sources.[29,31] Not all patients with symptoms will attend their GP, however, and not all symptoms will be reported or recorded. Some symptoms are likely to be coded (especially if they are the predominant ones), while others might only be recorded in the free text and so not be available for analysis in this database. The effect of this information or recording bias could be to over-inflate the hazard ratios if they relate to more severe symptoms (for example, major loss of appetite) or underestimate the hazard ratios if patients with the symptoms do not have them recorded. Integration of this utility into GP clinical systems could help improve the capture and recording of symptom data (significant positive and negative symptoms in the clinical record), as it could result in a template being presented to the clinician when an alarm symptom is entered into the clinical record. Over time, this would improve not only the medical record for clinical and medicolegal purposes but also the scope and quality of the data for refining this model.

While the validation cohort is derived from practices using the same clinical computer system (EMIS), they were physically discrete. Also, since this computer system is used in over half of UK general practices, the study results are likely to generalise well. A separate independent validation study using another GP database is planned and has not been included in the present study, so that it can be undertaken and published by an independent team.

Comparison with existing literature

This study builds on the approach by Jones *et al*, which quantified the PPV and sensitivity of dysphagia for oesophageal cancer,[8] and provides additional support to the concept of alarm symptoms in primary care.

Implications for research and practice

This study has developed and validated a prediction model that can be used to help identify patients with an existing but as yet undiagnosed gastro-oesophageal cancer. The algorithm is based on simple clinical variables that can be ascertained in clinical practice. The algorithm performed well in a separate validation sample with good discrimination and calibration. It could identify 10% of the population in which over 77% of all new gastro-oesophageal cancer cases arose over 2 years. Following external validation, this new algorithm could potentially be used to identify those at highest risk of gastro-oesophageal cancer, to facilitate early referral and investigation and so help earlier identification of patients with gastro-oesophageal cancer. However, 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 gastro-oesophageal cancer at diagnosis and subsequent survival.

REFERENCES

1. Ferlay J, Autier P, Boniol M, *et al.* Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol* 2007; **18(3):** 581–592.
2. Thomson CS, Forman D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EUROCARE results? *Br J Cancer* 2009; **101(Suppl 2):** S102–109.
3. Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009; **101(Suppl 2):** S1–4.
4. Department of Health. *The Cancer Reform Strategy.* London: Department of Health, 2007.
5. Vakil N, Moayyedi P, Fennerty MB, Talley NJ. Limited value of alarm features in the diagnosis of upper gastrointestinal malignancy: systematic review and meta-analysis. *Gastroenterology* 2006; **131(2):** 390–401; quiz 659–660.
6. Varadarajulu S, Eloubeidi MA, Patel RS, *et al.* The yield and the predictors of esophageal pathology when upper endoscopy is used for the initial evaluation of dysphagia. *Gastrointest Endosc* 2005; **61(7):** 804–808.
7. Meineche-Schmidt V, Jorgensen T. 'Alarm symptoms' in patients with dyspepsia: a three-year prospective study from general practice. *Scand J Gastroenterol* 2002; **37(9):** 999–1007.
8. 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.
9. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336(7659):** 1475–1482.
10. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 2009; **339:** b4229.
11. Hippisley-Cox J, Coupland C, Robson J, *et al.* Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009; **338:** b880.
12. Hippisley-Cox J, Coupland C. Predicting the risk of chronic kidney disease in men and women in england and Wales: prospective derivation and external validation of the QKidney® Scores. *BMC Fam Pract* 2010; **11(1):** 49.
13. 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.
14. 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.
15. Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing Type 2 diabetes. *Diabet Med* 2011; **28(5):** 599–607.
16. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; **340:** c2442.
17. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009; **339:** b2584.
18. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7:** 147–177.
19. Group TAM. Academic medicine: problems and solutions. *BMJ* 1989; **298:** 573–579.
20. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007; **60(9):** 979.
21. Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006; **59(10):** 1092.
22. Rubin DB. *Multiple imputation for non-response in surveys.* New York: John Wiley, 1987.
23. 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.
24. Hosmer D, Lemeshow S. *Applied logistic regression.* New York: John Wiley & Sons Inc., 1989.
25. Royston P. Explained variation for survival models. *Stata J* 2006; **6:** 1–14.
26. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23(5):** 723–748.
27. Hippisley-Cox J, Coupland C. Predicting the risk of osteoporotic fracture in England and Wales: prosepctive derivation and validation of the QFractureScore. *BMJ* 2009; **339:** b4229.
28. 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.
29. 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.
30. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; **60(572):** e128–136.
31. Jick H, Jick S, Derby LE, *et al.* Calcium-channel blockers and risk of cancer. *Lancet* 1997; **349(9051):** 525–528.