Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm

Abstract

Background
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 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 187 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.6–7 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.3–12 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 3.0). 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 dataset.
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, 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.

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.

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. 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);
- 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...
The risk-prediction algorithm was developed and validated using established methods.\textsuperscript{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.\textsuperscript{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.\textsuperscript{22} Fractional polynomials were used to model non-linear risk relationships with continuous variables.\textsuperscript{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.\textsuperscript{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.\textsuperscript{25} 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), the D statistic\textsuperscript{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.

### 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.

### 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)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Derivation cohort (n = 2 355 719)</th>
<th>Validation cohort (n = 1 238 971)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td>1 174 921 (49.9)</td>
<td>617 493 (49.8)</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>1 180 798 (50.1)</td>
<td>621 478 (50.2)</td>
</tr>
<tr>
<td><strong>Mean age (SD), years</strong></td>
<td>50.1 (15.0)</td>
<td>50.1 (15.0)</td>
</tr>
<tr>
<td><strong>BMI recorded prior to study entry</strong></td>
<td>1 869 779 (79.4)</td>
<td>1 005 846 (81.2)</td>
</tr>
<tr>
<td><strong>Mean Townsend score (SD)</strong></td>
<td>$-0.3 (3.4)$</td>
<td>$-0.2 (3.6)$</td>
</tr>
<tr>
<td><strong>Alcohol status not recorded</strong></td>
<td>516 915 (21.9)</td>
<td>255 672 (20.6)</td>
</tr>
<tr>
<td><strong>Smoking status not recorded</strong></td>
<td>200 391 (8.5)</td>
<td>96 284 (7.8)</td>
</tr>
<tr>
<td><strong>Medical history</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Family history of gastrointestinal cancer</strong></td>
<td>29 636 (1.3)</td>
<td>17 742 (1.4)</td>
</tr>
<tr>
<td><strong>Prior cancer apart from gastro-oesophageal cancer</strong></td>
<td>53 971 (2.3)</td>
<td>28 520 (2.3)</td>
</tr>
<tr>
<td><strong>Current symptoms and symptoms in the preceding year</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Current dysphagia</strong></td>
<td>15 021 (0.6)</td>
<td>8 145 (0.7)</td>
</tr>
<tr>
<td><strong>Current haematemesis</strong></td>
<td>12 952 (0.5)</td>
<td>7 119 (0.6)</td>
</tr>
<tr>
<td><strong>Current abdominal pain</strong></td>
<td>225 543 (9.6)</td>
<td>126 161 (10.2)</td>
</tr>
<tr>
<td><strong>Current appetite loss</strong></td>
<td>9978 (0.4)</td>
<td>6 333 (0.5)</td>
</tr>
<tr>
<td><strong>Current weight loss</strong></td>
<td>9998 (0.4)</td>
<td>5 377 (0.4)</td>
</tr>
<tr>
<td><strong>Tiredness in last year</strong></td>
<td>25 200 (1.1)</td>
<td>14 119 (1.1)</td>
</tr>
<tr>
<td><strong>Haemoglobin recorded in the last year</strong></td>
<td>22 576 (1.0)</td>
<td>12 638 (1.0)</td>
</tr>
<tr>
<td><strong>Haemoglobin &lt;11 g/dl in the last year</strong></td>
<td>406 410 (17.3)</td>
<td>218 862 (17.7)</td>
</tr>
</tbody>
</table>

BMI = body mass index. SD = standard deviation. Figures in the tables are counts (%) unless otherwise specified.
Of these a total of 2,538,615 patients aged 30–84 years were identified in the derivation cohort, and 12,445 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, 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).

Incidence rates of red flag symptoms
Overall, in the derivation cohort, 15,021 patients were identified with incident dysphagia, 12,952 with haematemesis, 9,978 with appetite loss, 9,998 with weight loss, and 225,543 with abdominal pain; 4,203 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 2,527 incident cases, 1,531 (60.6%) were oesophageal cancer and 996 (39.4%) gastric cancer. There were 2,080 cases (82.3% of 2,527) identified using the GP record and an additional 447 (17.7%) identified from the linked death record.

In the validation cohort, 1,343 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 the 2527 incident cases, 1531 (60.6%) were oesophageal cancer and 996 (39.4%) gastric cancer. There were 1,126 cases (83.8% of 1,343) 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, appetite loss, weight loss, abdominal pain, haematemesis, and anaemia.
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 health record.

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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td></td>
</tr>
<tr>
<td>R² statistic, a %</td>
<td>71.2 (69.2 to 73.2)</td>
</tr>
<tr>
<td>D statistic b</td>
<td>3.22 (3.06 to 3.37)</td>
</tr>
<tr>
<td>ROC statistic c</td>
<td>0.89 (0.87 to 0.91)</td>
</tr>
<tr>
<td>Men</td>
<td></td>
</tr>
<tr>
<td>R² statistic, a %</td>
<td>72.5 (71.1 to 73.8)</td>
</tr>
<tr>
<td>D statistic b</td>
<td>3.32 (3.21 to 3.43)</td>
</tr>
<tr>
<td>ROC statistic c</td>
<td>0.92 (0.91 to 0.93)</td>
</tr>
</tbody>
</table>

aR² statistic shows explained variation in time to diagnosis of gastro-oesophageal cancer — higher values indicate more variation is explained. bD statistic is a measure of discrimination — higher values indicate better discrimination. cROC statistic is a measure of discrimination — higher values indicate better discrimination.
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**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Risk threshold %</th>
<th>True negative</th>
<th>False negative</th>
<th>False positive</th>
<th>True positive</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Positive predictive value (%)</th>
<th>Negative predictive value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current dysphagia</td>
<td>n/a</td>
<td>956 541</td>
<td>109</td>
<td>5156</td>
<td>434</td>
<td>32.3</td>
<td>99.5</td>
<td>7.8</td>
<td>199.9</td>
</tr>
<tr>
<td>Current haematemesis</td>
<td>n/a</td>
<td>957 321</td>
<td>1242</td>
<td>4376</td>
<td>101</td>
<td>7.5</td>
<td>99.5</td>
<td>2.3</td>
<td>99.9</td>
</tr>
<tr>
<td>Current abdominal pain</td>
<td>n/a</td>
<td>870 379</td>
<td>1034</td>
<td>91 318</td>
<td>309</td>
<td>23.0</td>
<td>90.5</td>
<td>0.3</td>
<td>99.9</td>
</tr>
<tr>
<td>Current appetite loss</td>
<td>n/a</td>
<td>958 441</td>
<td>1308</td>
<td>3256</td>
<td>35</td>
<td>2.6</td>
<td>99.7</td>
<td>1.1</td>
<td>99.9</td>
</tr>
<tr>
<td>Current weight loss</td>
<td>n/a</td>
<td>952 634</td>
<td>1236</td>
<td>9663</td>
<td>107</td>
<td>8.0</td>
<td>99.1</td>
<td>1.2</td>
<td>99.9</td>
</tr>
<tr>
<td>Anaemia</td>
<td>n/a</td>
<td>951 467</td>
<td>1224</td>
<td>10 230</td>
<td>119</td>
<td>8.9</td>
<td>98.9</td>
<td>1.1</td>
<td>99.9</td>
</tr>
</tbody>
</table>

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 | 40 | 29.9 | 99.7 | 10.8 | 99.9 |

n/a = not applicable. *Criterion not met does not have disease. **Criterion not met does have disease. *Criterion met does not have disease. **Criterion met does have disease.
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, 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. For example, one systematic review reported that, on average, 89% of diagnoses recorded on the GP electronic record are confirmed from other data sources.

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 medico-legal 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, 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

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): 1072.