BMJ Open

Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study

Julia Hippisley-Cox, Carol Coupland

To cite: Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015;**5**: e007825. doi:10.1136/bmjopen-2015-007825

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-007825).

Received 29 January 2015
Accepted 9 February 2015

Division of Primary Care, University Park, Nottingham, UK

Correspondence to
Professor Julia Hippisley-Cox; Julia.hippisley-cox@nottingham.ac.uk

ABSTRACT

Objective: To derive and validate a set of clinical risk prediction algorithm to estimate the 10-year risk of 11 common cancers.

Design: Prospective open cohort study using routinely collected data from 753 QResearch general practices in England. We used 565 practices to develop the scores and 188 for validation.

Subjects: 4.96 million patients aged 25–84 years in the derivation cohort; 1.64 million in the validation cohort. Patients were free of the relevant cancer at baseline.

Methods: Cox proportional hazards models in the derivation cohort to derive 10-year risk algorithms. Risk factors considered included age, ethnicity, deprivation, body mass index, smoking, alcohol, previous cancer diagnoses, family history of cancer, relevant comorbidities and medication. Measures of calibration and discrimination in the validation cohort.

Outcomes: Incident cases of blood, breast, bowel, gastro-oesophageal, lung, oral, ovarian, pancreas, prostate, renal tract and uterine cancers. Cancers were recorded on any one of four linked data sources (general practitioner (GP), mortality, hospital or cancer records).

Results: We identified 228 241 incident cases during follow-up of the 11 types of cancer. Of these 25 444 were blood; 41 315 breast; 32 626 bowel, 12 808 gastro-oesophageal; 32 187 lung; 4811 oral; 6635 ovarian; 7119 pancreatic; 35 256 prostate; 23 091 renal tract; 6949 uterine cancers. The lung cancer algorithm had the best performance with an R^2 of 64.2%; D statistic of 2.74; receiver operating characteristic curve statistic of 0.91 in women. The sensitivity for the top 10% of women at highest risk of lung cancer was 67%. Performance of the algorithms in men was very similar to that for women.

Conclusions: We have developed and validated a prediction models to quantify absolute risk of 11 common cancers. They can be used to identify patients at high risk of cancers for prevention or further assessment. The algorithms could be integrated into clinical computer systems and used to identify high-risk patients.

Web calculator: There is a simple web calculator to implement the Qcancer 10 year risk algorithm together with the open source software for download (available at http://qcancer.org/10yr/).

Strengths and limitations of this study

- The QCancer-10 year risk algorithms provide valid measures of absolute risk in the general population of patients as shown by the performance in a separate validation cohort.
- Key strengths include size, duration of follow up, representativeness, and lack of selection, recall and respondent bias.
- 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.
- The study used linked hospital, mortality and cancer records and is therefore likely to have picked up the majority of cancer diagnoses.
- The algorithms do not include some potential risk factors for cancer such as diet or physical activity since these are not routinely recorded in electronic health records.

INTRODUCTION

The UK has one of the worst records for cancer in Europe with late diagnoses and poor survival.[1] Each year around 230 000 people in England are diagnosed with cancer and around 125 000 die from it.[2] The early diagnosis and prevention of cancer is likely to remain a high priority especially with the global decline in other major causes of mortality such as cardiovascular disease and the ageing population.

Over the past few years, we developed and validated a series of risk prediction algorithms collectively known as the QCancer algorithms.[3–10] These were designed to quantify the absolute risk that a patient has an *existing* cancer based on combinations of symptoms and readily available risk factors and are intended to help inform decisions regarding further investigation and referral.[3–10] We decided to build on this work and derive a

set of risk prediction algorithms to quantify absolute risk of *future* cancer over a 10-year period using predictor variables recorded in the patient's primary care electronic record. In particular, we were interested to quantify the absolute risk of cancer in (1) patients with a positive family history of specific cancers, previous cancers or a chronic disease which might increase cancer risk and might require additional surveillance and (2) those with potentially modifiable risk factors (such as smoking and alcohol) for whom quantification of absolute risk might be useful to support efforts to lower risk. We decided to focus on the 11 most commonly occurring cancers in men and women in England. This paper reports the results of the derivation and validation of the new algorithms based on the QResearch database linked to cancer registrations, mortality and hospital episode statistics.

METHODS

Study design and data source

We undertook a prospective cohort study in a large population of primary care patients from an open cohort study using the QResearch database (V.38).

The QResearch database is a large pseudonymised database of electronic health records from over 750 general practices in the UK which has been described in detail elsewhere (http://www.qresearch.org). Over 99% of people in the UK are registered with general practices and have information routinely recorded on an ongoing basis when they consult their general practitioner (GP) or other primary care professional, receive prescriptions and from referrals to secondary care. The database includes event level detailed information on patient demographics (year of birth, sex, ethnicity, deprivation), medication, clinical diagnoses, referrals, clinical values (such as body mass index (BMI), systolic blood pressure), laboratory investigations. The QResearch database has data from primary care dating back to 1989 which has been linked at individual patient level to cancer registrations data (from 1990 onwards), mortality records (from 1997 onwards) and to hospital admissions data (from 1998 onwards). It has a population which is representative of that in UK and the database has been used extensively for epidemiological research including disease-based epidemiology, health services research, the development of risk prediction models and evaluation of drug safety.

We included all practices in England who had been using their Egton Medical Information Systems (EMIS) computer system for at least a year. We randomly allocated three-quarters of practices to the derivation data set and the remaining quarter to a validation data set.

We identified an open cohort of patients aged 25–84 years drawn from patients registered with practices between 1 January 1998 and 30 September 2013. We excluded patients who did not have a postcode-related Townsend score. For each type of cancer, we excluded patients with a history of the relevant cancer any time prior to the study start date. We determined an initial entry date to the cohort for each patient, which was the latest of the following dates: 25th birthday, date of registration with the practice plus 1 year, date on which the practice computer system was installed plus 1 year, and the beginning of the study period (1 January 1998). Patients were censored at the earliest date of the diagnosis of cancer, death, deregistration with the practice, last upload of computerised data, or the study end date (1 October 2013).

Cancer outcomes

Our outcomes were incident diagnosis of each type of cancer focusing on the 11 most common cancers in England in 2012 (excluding skin cancers) as identified by the Office of National Statistics (ONS).[11] The cancers were as follows (listed alphabetically):

1. Blood cancers (including leukaemias, Hodgkins lymphomas, non-Hodgkins lymphoma, myeloma)
2. Bowel cancer (including colon and rectal cancer)
3. Breast cancer (women only)
4. Gastro-oesophageal cancer
5. Lung cancer
6. Oral cancer
7. Ovarian cancer (women only)
8. Pancreatic cancer
9. Prostate cancer (men only)
10. Renal tract cancer (bladder or kidney)
11. Uterine cancer (women only).

We included cancer cases diagnosed on any of the four linked data sources (1) patients GP record (2) on their linked mortality record (3) hospital record or (4) cancer registry record. We used the earliest recorded date of cancer diagnosis on any of the four data sources as the index date.

The QResearch database is linked at individual patient level to the hospital admissions data, the Office for National Statistics mortality records and the Office for National Statistics Cancer Registry using a project specific pseudonymised National Health Service (NHS) number. The recording of NHS numbers is valid and complete for 99.8% of QResearch patients, 99.9% for ONS mortality and cancer records and 98% for hospital admissions records.[12 13] We defined patients as having the cancer of interest if there was a record of the relevant clinical code either in their GP record, their linked hospital record, their linked mortality record or their linked cancer registry record.

We used Read codes to identify cancer cases from the GP record. We used International Classification of Diseases 10th Edition (ICD 10) clinical codes to identify cancer cases from hospital, cancer registry and mortality records except for the period of 3 years between 1 January 1998 and 31 December 2000 where ICD 9 was in use for mortality records. Web extra table 1 lists all the clinical codes used to identify each cancer outcome.

Risk factors

We examined the following predictor variables based on established risk factors for each cancer as determined in other studies[14–17] or listed on the Cancer Research UK website.[18] We focused on variables which are likely to be recorded in the patient's electronic record and which the patient is likely to know.

Variables tested for all outcomes

The following variables were considered for all cancer outcomes:

- Age at baseline (continuous, ranging from 25 to 84 years)
- BMI (continuous)
- Smoking status (non-smoker; ex; light (1–9 cigarettes/day); moderate (10–19 cigarettes/day); heavy smoker (20+ cigarettes/day)
- Alcohol use (none, trivial (<1 unit/day); light (1–2 units/day); moderate (3–6 units/day); heavy (7–9 units/day); very heavy (>9 units/day))
- Townsend deprivation score (continuous)
- Ethnic group (White/not recorded, Indian, Pakistani, Bangladeshi, Other Asian, Caribbean, Black African, Chinese, Other)
- Type 1 diabetes
- Type 2 diabetes
- Manic depression or schizophrenia
- Use of antipsychotics at baseline
- Use of hormone replacement therapy (HRT, women) at baseline (progesterone only, oestrogen only; combined preparation; no HRT)
- Use of oral contraceptive (women) at baseline
- Previous diagnoses of cancer at baseline apart from the incident one under consideration recorded in the GP record. We included diagnoses of the following cancers as separate predictors for each cancer outcome—blood, bowel, brain, breast (in women), cervical (in women), gastro-oesophageal, lung, oral, ovarian (in women), pancreatic, prostate (in men), renal tract, uterine (in women).

Each type of previous cancer was considered as a separate predictor for each outcome.

Variables tested for specific cancers

Additional risk factors for individual cancer outcomes included:

- *Blood cancers*: family history of blood cancers
- *Bowel cancer*: family history of bowel cancer; ulcerative colitis; past colonic polyps
- *Breast cancer*: family history of breast cancer; family history of gynaecological cancer; benign breast disease (fibrocystic disease, intraductal papilloma, fibroadenoma)
- *Gastro-oesophageal cancer*: family history of bowel cancer; Barratt's oesophagus; peptic ulcer
- *Lung cancer*: family history of lung cancer; asbestos exposure; asthma; chronic obstructive pulmonary disease
- *Ovarian cancer*: family history of ovarian cancer; polycystic ovarian disease; endometriosis
- *Pancreatic cancer*: chronic pancreatitis; peptic ulcer
- *Prostate cancer*: family history of prostate cancer
- *Renal tract cancer*: family history of renal cancer; renal stones
- *Uterine cancer*: family history of gynaecological cancer; polycystic ovarian disease; endometriosis; fibroids; endometrial polyps or endometrial hyperplasia.

BMI, smoking status and alcohol use were obtained from values recorded closest to the baseline date and prior to cancer diagnosis. All other risk factors were based on the latest information recorded before entry to the cohort. Use of antipsychotic medication at baseline was defined as at least two prescriptions with the most recent one within 28 days of the date of entry to the cohort. For HRT and oral contraceptive use, the definition was at least two prescriptions with the most recent one within 6 months of the date of entry to the cohort since prescriptions are often issued for 6 months at a time.

Derivation and validation of the models

We developed and validated the risk prediction algorithms using established methods.[10 19 20–22] We used multiple imputation to replace missing values for BMI, alcohol and smoking status and used these values in our main analyses.[23–26] We carried out five imputations. We used Cox's proportional hazards models to estimate the coefficients for each risk factor for men and women separately, using robust variance estimates to allow for the clustering of patients within general practices. We used Rubin's rules to combine the results across the imputed data sets.[27] We used fractional polynomials[28] to model non-linear risk relationships with continuous variables (age, BMI, Townsend deprivation score). We fitted full models initially and retained variables if they had a hazard ratio (HR) of <0.90 or >1.10 (for binary variables) and were statistically significant at the 0.01 level. For previous diagnoses of cancer, we retained variables which were significant at the 0.05 level since some of the cancers are rare. In order to simplify the models we focused on variables for the most common conditions and medications and combined similar variables with comparable HRs where possible.

We examined interactions between predictor variables and age (focusing on family history of cancer and smoking status). We used the regression coefficients for each variable from the final model as weights which we combined with the baseline survivor function evaluated for each month up to 15 years to derive risk equations over a period of 15 years of follow-up.[29] This enabled us to derive risk estimates for each year of follow-up, with a specific focus on 10-year risk estimates. We estimated the baseline survivor function based on zero values of centred continuous variables, with all binary predictor values set to zero.

Validation of the models

We used multiple imputation in the validation cohort to replace missing values for BMI, alcohol and smoking

status. We carried out five imputations. We applied the risk equations for men and women obtained from the derivation cohort to the validation cohort and calculated measures of discrimination. As in previous studies,[30] we calculated R^2 (explained variation in time to cancer diagnosis[31]), D statistic[32] (a measure of discrimination where higher values indicate better discrimination) and the area under the receiver operating characteristic curve (ROC statistic) at 10 years and combined these across data sets using Rubin's rules. We assessed calibration (comparing the mean predicted risks at 10 years with the observed risk by tenth of predicted risk). The observed risks were obtained using the Kaplan-Meier estimate evaluated at 10 years. We applied the algorithms to the validation cohort to define the thresholds for the 10% of patients at highest estimated risk of each cancer at 10 years.

As an additional analysis we also calculated the validation statistics on the subset of the validation cohort with no missing data for BMI, alcohol and smoking status (a complete case analysis).

We used all the available data on the database to maximise the power and also generalisability of the results. We used STATA (V.13) for all analyses.

RESULTS

Overall study population

Overall, 753 QResearch practices in England met our inclusion criteria, of which 565 were randomly assigned to the derivation data set with the remainder assigned to a validation cohort. We identified 4 964 904 patients aged 25–84 years in the derivation cohort. We excluded 21 139 patients (0.43%) without a recorded Townsend score, leaving 4 943 765 for analysis. Table 1 shows numbers with diagnoses of each previous cancer excluded from the model for each type of cancer.

We identified 1 635 592 patients aged 25–84 years in the QResearch validation cohort. We excluded 10 796 patients (0.66%) without a recorded Townsend score, leaving 1 624 796 patients for the analysis. The numbers of patients with previous cancer diagnoses in the validation cohort are shown in table 1.

Table 1 also shows the baseline characteristics of men and women in the derivation and validation cohorts. For example, in the derivation cohort, smoking status was recorded in 94.1% of women, alcohol intake in 83.5%, ethnicity in 61.1% and BMI in 84.4%. These values were around 5% higher than recording levels in men and were similar to corresponding values for men and women in the validation cohort. As in previous studies[11 13] the patterns of missing data supported the use of multiple imputation to replace missing values for smoking status, alcohol intake and BMI.

Incidence rates of cancer

In the derivation cohort, we identified 228 241 incident cases of the 11 types of cancer on one or more of the four linked data sources during follow-up. Of these 25 444 were blood; 32 626 were bowel; 41 315 were breast; 12 808 were gastro-oesophageal; 32 187 were lung; 4811 were oral; 6635 were ovarian; 7119 were pancreatic; 35 256 were prostate cancer; 23 091 were renal tract; 6949 were uterine.

Table 2 shows the numbers of cases and age standardised incidence rates for each cancer in women in the derivation cohort. There were a total of 110 555 cases of the 10 types of cancer in women identified on one or more of the four linked data sources (GP, hospital, mortality or cancer registry). Of these, 79 863 (72.2%) were ascertained from the GP record, 89 927 (81.3%) from the GP or linked mortality record; 105 465 (95.4%) from the GP, mortality or hospital record. The highest ascertainment rate based on the GP record alone was for breast cancer (86.8%) and the lowest was for uterine cancer (47.4%).

Table 3 shows the corresponding figures for men. Of the 117 686 cases of the eight cancer types in men recorded on any of the four data sources, 84 787 (72.0%) were ascertained on the GP record, 95 933 (81.5%) from the GP or linked mortality record, 111 990 (95.2%) from the GP, mortality or hospital record. Ascertainment based on the GP record alone was highest for prostate cancer (85.1%) and lowest for renal tract cancer (55.3%).

Predictor variables

Table 4 shows the adjusted HRs for the final models for seven cancers occurring in men and women (blood, bowel, gastro-oesophageal, lung, oral, pancreas, renal tract). Table 5 shows the adjusted HRs for three cancers occurring in women (breast, ovary and uterus). Table 6 shows the adjusted HRs for prostate cancer. Figure 1 shows graphs of adjusted HRs from the fractional polynomial terms for age for each cancer. Figure 2 shows graphs of adjusted HRs from the fractional polynomial terms for BMI for each cancer. Figure 3 shows graphs of adjusted HRs for the fractional polynomial terms for Townsend deprivation score for each cancer. Figure 4 shows graphs of the adjusted HRs for the interactions between age and family history for each relevant cancer. Figure 5 show graphs of the adjusted HRs for the interactions between age and smoking status for each relevant cancer.

The text below describes the models for women in detail below though similar results were obtained for men as shown in the relevant tables.

Blood cancer

There were seven variables in the final model for women for blood cancer. These were age, BMI (linear), smoking status (33% higher risk in heavy smokers compared with non-smokers), type 1 diabetes (51% increased risk), family history of blood cancer (fourfold higher risk), prior brain cancer (fourfold higher risk), and prior ovarian cancer (59% increased risk). The final

Table 1 Baseline characteristics of patients in the derivation and validation cohorts aged 25–84 years

	Derivation men n (%)	Derivation women n (%)	Validation men n (%)	Validation women n (%)
Total (years)	2 447 866	2 495 899	802 437	822 359
25–34	805 109 (32.9)	870 317 (34.9)	278 480 (34.7)	309 425 (37.6)
35–44	613 038 (25.0)	551 468 (22.1)	201 950 (25.2)	179 621 (21.8)
45–54	425 824 (17.4)	394 520 (15.8)	132 729 (16.5)	121 668 (14.8)
55–64	298 448 (12.2)	297 981 (11.9)	91 781 (11.4)	91 285 (11.1)
65–74	203 714 (8.3)	227 761 (9.1)	64 283 (8.0)	71 086 (8.6)
75+	101 733 (4.2)	153 852 (6.2)	33 214 (4.1)	49 274 (6.0)
Mean age (SD)	44.3 (14.8)	44.9 (15.9)	43.8 (14.7)	44.1 (15.9)
Mean Townsend score (SD)	0.3 (3.6)	0.2 (3.6)	0.6 (3.6)	5 (3.5)
Body mass index recorded	1 903 519 (77.8)	2 105 539 (84.4)	621 882 (77.5)	691 753 (84.1)
Mean BMI (SD)	26.3 (4.2)	25.7 (5.0)	26.2 (4.1)	25.5 (5.0)
Ethnicity recorded	1 380 685 (56.4)	1 525 005 (61.1)	462 619 (57.7)	509 242 (61.9)
White/not recorded	2 231 641 (91.2)	2 271 520 (91.0)	725 421 (90.4)	743 043 (90.4)
Indian	42 771 (1.7)	37 773 (1.5)	13 192 (1.6)	12 376 (1.5)
Pakistani	22 004 (0.9)	16 893 (0.7)	8557 (1.1)	6436 (0.8)
Bangladeshi	17 169 (0.7)	13 170 (0.5)	5803 (0.7)	4235 (0.5)
Other Asian	24 494 (1.0)	27 750 (1.1)	7954 (1.0)	9034 (1.1)
Caribbean	18 553 (0.8)	23 920 (1.0)	6989 (0.9)	8691 (1.1)
Black African	37 003 (1.5)	40 742 (1.6)	14 911 (1.9)	15 729 (1.9)
Chinese	12 493 (0.5)	17 702 (0.7)	3564 (0.4)	5281 (0.6)
Other	41 738 (1.7)	46 429 (1.9)	16 046 (2.0)	17 534 (2.1)
Smoking recorded	2 188 935 (89.4)	2 349 027 (94.1)	715 821 (89.2)	774 509 (94.2)
Non-smoker	1 081 822 (44.2)	1 433 446 (57.4)	347 253 (43.3)	467 440 (56.8)
Ex-smoker	448 480 (18.3)	392 870 (15.7)	149 221 (18.6)	134 426 (16.3)
Light smoker (1–9 cigarettes/day)	351 559 (14.4)	284 482 (11.4)	117 904 (14.7)	96 014 (11.7)
Moderate smoker (10–19 Cigarettes/day)	167 089 (6.8)	152 115 (6.1)	55 987 (7.0)	49 627 (6.0)
Heavy smoker (20+ cigarettes /day)	139 985 (5.7)	86 114 (3.5)	45 456 (5.7)	27 002 (3.3)
Alcohol recorded	1 930 167 (78.9)	2 084 701 (83.5)	623 247 (77.7)	677 551 (82.4)
Non-drinker	433 515 (17.7)	753 150 (30.2)	137 452 (17.1)	238 102 (29.0)
Trivial drinker (<1 unit/day)	585 589 (23.9)	849 734 (34.0)	187 131 (23.3)	275 188 (33.5)
Light drinker (1–2 units/day)	358 713 (14.7)	295 009 (11.8)	118 697 (14.8)	101 967 (12.4)
Moderate drinker (3–6 units/day)	486 003 (19.9)	176 644 (7.1)	159 164 (19.8)	58 675 (7.1)
Heavy drinker (7–9 units/day)	41 223 (1.7)	5332 (0.2)	12 651 (1.6)	1782 (0.2)
Very heavy drinker (>9 units/day)	18 473 (0.8)	3743 (0.1)	5964 (0.7)	1446 (0.2)
Family history of cancer				
Family history of lung cancer	13 967 (0.6)	17 453 (0.7)	4302 (0.5)	5366 (0.7)
Family history of bowel cancer	29 877 (1.2)	43 741 (1.8)	9346 (1.2)	13 343 (1.6)
Family history of renal cancer	2465 (0.1)	2767 (0.1)	952 (0.1)	1070 (0.1)
Family history of breast cancer	NA	95 807 (3.8)	NA	32 725 (4.0)
Family history of uterine cancer	NA	2030 (0.1)	NA	638 (0.1)
Family history of ovarian cancer	NA	5412 (0.2)	NA	1722 (0.2)
Family history of prostate cancer	4230 (0.2)	NA	979 (0.1)	NA
Prior diagnosis of cancer				
Prior bowel cancer	4872 (0.2)	4330 (0.2)	1553 (0.2)	1370 (0.2)
Prior pancreatic cancer	143 (0.0)	140 (0.0)	57 (0.0)	47 (0.0)
Prior lung cancer	1488 (0.1)	977 (0.0)	485 (0.1)	293 (0.0)
Prior gastro-oesophageal cancer	976 (0.0)	562 (0.0)	303 (0.0)	134 (0.0)
Prior renal cancer	4069 (0.2)	1561 (0.1)	1263 (0.2)	499 (0.1)
Prior blood cancer	5953 (0.2)	4257 (0.2)	1906 (0.2)	1399 (0.2)
Prior oral cancer	964 (0.0)	571 (0.0)	315 (0.0)	215 (0.0)
Prior brain cancer	180 (0.0)	157 (0.0)	55 (0.0)	58 (0.0)
Prior breast cancer	NA	25 108 (1.0)	NA	7781 (0.9)
Prior uterine cancer	NA	1987 (0.1)	NA	669 (0.1)
Prior ovarian cancer	NA	2242 (0.1)	NA	725 (0.1)
Prior cervical cancer	NA	3582 (0.1)	NA	1194 (0.1)
Prior prostate cancer	7778 (0.3)	NA	2518 (0.3)	NA

Continued

Table 1 Continued

	Derivation men n (%)	Derivation women n (%)	Validation men n (%)	Validation women n (%)
Comorbidities				
Type 1 diabetes	9095 (0.4)	7207 (0.3)	2926 (0.4)	2391 (0.3)
Type 2 diabetes	68 727 (2.8)	53 070 (2.1)	21 772 (2.7)	16 959 (2.1)
Barratt's oesophagus	3611 (0.1)	1760 (0.1)	1083 (0.1)	543 (0.1)
Peptic ulcer disease	65 467 (2.7)	34 951 (1.4)	20 005 (2.5)	10 404 (1.3)
Ulcerative colitis	8956 (0.4)	8983 (0.4)	2923 (0.4)	2751 (0.3)
Chronic pancreatitis	2438 (0.1)	1701 (0.1)	818 (0.1)	512 (0.1)
Colonic polyp	3146 (0.1)	2447 (0.1)	1010 (0.1)	802 (0.1)
Exposure to asbestos	2960 (0.1)	341 (0.0)	885 (0.1)	89 (0.0)
Asthma	201 250 (8.2)	225 052 (9.0)	66 466 (8.3)	74 440 (9.1)
COPD	28 194 (1.2)	22 731 (0.9)	8929 (1.1)	7056 (0.9)
Renal stones	28 022 (1.1)	21 640 (0.9)	8511 (1.1)	6602 (0.8)
Manic depression/schizophrenia	18 455 (0.8)	17 100 (0.7)	6618 (0.8)	5927 (0.7)
Benign breast disease	NA	78 762 (3.2)	NA	24 157 (2.9)
Polycystic ovarian disease	NA	31 196 (1.2)	NA	10 914 (1.3)
Endometrial hyperplasia or endometrial polyps	NA	29 107 (1.2)	NA	9031 (1.1)
Uterine fibroids	NA	39 075 (1.6)	NA	12 265 (1.5)
Prescribed medication				
HRT (oestrogen only)	NA	159 516 (6.4)	NA	48 521 (5.9)
HRT (progesterone only)	NA	33 638 (1.3)	NA	9721 (1.2)
HRT (combined)	NA	56 096 (2.2)	NA	17 119 (2.1)
Oral contraceptive	NA	248 131 (9.9)	NA	84 565 (10.3)
Tamoxifen	NA	11 899 (0.5)	NA	3805 (0.5)
Antipsychotics	33 562 (1.4)	61 595 (2.5)	10 805 (1.3)	18 926 (2.3)

BMI, body mass index; COPD, chronic obstructive pulmonary disease; HRT, hormone replacement therapy; NA, not applicable.

model for men was similar except that prior renal cancer (46% increased risk) was a predictor instead of prior brain cancer.

Bowel cancer

There were 12 variables in the final model for bowel cancer for women. These were age, ethnicity (lower risk among non-white groups), smoking status (17% higher risk in heavy smokers), alcohol (36% higher risk in heavy drinkers), family history of bowel cancer (94% higher risk at the mean age of 45), ulcerative colitis (75% higher risk), colonic polyp (twofold higher risk), type2 diabetes (16% higher risk), previous breast cancer (16% higher risk), previous ovarian cancer (98% higher risk), previous uterine cancer (61% higher risk), previous cervical cancer (74% higher risk). There was an interaction between family history of bowel cancer and age which indicated higher HRs associated with a family history in younger women compared with older women (figure 4).

The direction and magnitude of the HRs was similar for men except for there were three types of prior cancer which were significant in men but not women (lung, blood, oral). BMI and Townsend deprivation score were also included as linear terms in the model for men.

Breast cancer

There were 13 variables in the final model for breast cancer. These were age, BMI, Townsend deprivation score, ethnicity (lower risks in non-white ethnic groups), alcohol (25% higher risk in heavy drinkers compared with non-drinkers), family history of breast cancer (93% higher risk at the mean age of 45 years), benign breast disease (51% higher risk), oral contraceptive pill (13% higher risk), oestrogen containing HRT (18% higher risk), manic depression or schizophrenia (16% higher risk), previous blood cancer (57% higher risk), previous lung cancer (86% higher risk), and previous ovarian cancer (42% higher risk). Increasing deprivation was associated with a lower risk of breast cancer (figure 3). There was an interaction between family history of breast cancer and age which indicated higher HRs associated with a family history of breast cancer in younger women compared with older women (figure 4).

Gastro-oesophageal cancer

There were 12 variables in the final model for gastro-oesophageal cancer in women. The 12 variables were age, BMI, Townsend deprivation score, smoking status (2.4-fold higher risk in heavy smokers), alcohol (trivial alcohol use was associated with an 11% lower risk and very heavy drinking with a twofold increased risk compared with non-drinkers), Barratt's oesophagus (3.8-fold higher risk), peptic ulcer disease (29% higher risk), type 2 diabetes (33% higher risk), and, previous lung cancer (2.3-fold higher risk), previous blood cancer (twofold

Table 2 Numbers of incident cases, age standardised incidence rates per 10 000 person years in the derivation cohort in women aged 25–84 years

	Cases on GP record			Cases on either GP or linked mortality record		Cases on either GP, linked mortality or hospital record		Cases on either GP, linked hospital, mortality or cancer record	
Cancer Type	Cases	Row % of total	Age standardised rate per 10 000	Cases	Age standardised rate per 10 000	Cases	Age standardised rate per 10 000	Total Cases	Age standardised rate per 10 000
Blood cancer	7519	67.4	3.95 (3.86 to 4.04)	8694	4.53 (4.44 to 4.63)	10 493	5.48 (5.38 to 5.59)	11 151	5.82 (5.71 to 5.93)
Breast cancer	35 874	86.8	19.9 (19.7 to 20.1)	37 057	20.5 (20.3 to 20.7)	40 251	22.3 (22.1 to 22.5)	41 315	22.9 (22.7 to 23.1)
Bowel cancer	10 230	70.6	5.29 (5.18 to 5.39)	11 443	5.89 (5.78 to 5.99)	13 824	7.12 (7.00 to 7.24)	14 496	7.46 (7.34 to 7.58)
Gastro-oesophageal	3104	70.4	1.57 (1.52 to 1.63)	3708	1.87 (1.81 to 1.93)	4251	2.15 (2.08 to 2.21)	4407	2.22 (2.16 to 2.29)
Lung cancer	8823	65.0	4.58 (4.48 to 4.67)	11 618	5.98 (5.87 to 6.09)	12 904	6.65 (6.54 to 6.77)	13 570	6.99 (6.87 to 7.10)
Oral cancer	968	56.4	0.52 (0.49 to 0.55)	1073	0.57 (0.54 to 0.61)	1556	0.83 (0.79 to 0.87)	1715	0.91 (0.87 to 0.95)
Ovarian cancer	4199	63.3	2.25 (2.18 to 2.32)	5006	2.66 (2.58 to 2.73)	6061	3.23 (3.15 to 3.31)	6635	3.54 (3.45 to 3.62)
Pancreatic cancer	2089	59.5	1.07 (1.02 to 1.12)	2975	1.51 (1.46 to 1.56)	3347	1.70 (1.64 to 1.76)	3512	1.78 (1.72 to 1.84)
Renal tract cancer	3762	55.3	1.94 (1.87 to 2.00)	4474	2.29 (2.22 to 2.35)	6438	3.31 (3.23 to 3.39)	6805	3.50 (3.41 to 3.58)
Uterine cancer	3295	47.4	1.76 (1.70 to 1.82)	3879	2.06 (2.00 to 2.13)	6340	3.39 (3.31 to 3.47)	6949	3.72 (3.63 to 3.80)
Total cancers	79 863	72.2		89 927		105 465		110 555	

Patients with existing diagnoses of each cancer at baseline were dropped from the relevant cohort.
Rates were age standardised to the overall QResearch population in 5 year age bands.
GP, general practitioner.

Table 3 Numbers of incident cases, age standardised incidence rates per 10 000 person years in the derivation cohort in men aged 25–84 years

	Cases on GP record			Cases on either GP or linked mortality record		Cases on either GP, linked mortality or hospital record		Cases on either GP, linked hospital, mortality or cancer record	
Cancer Type	Cases	Row % of total	Age standardised rate per 10 000	Cases	Age standardised rate per 10 000	Cases	Age standardised rate per 10 000	Total Cases	Age standardised rate per 10 000
Blood cancer	9614	67.3	5.75 (5.63 to 5.86)	11 230	6.77 (6.64 to 6.89)	13 499	8.12 (7.99 to 8.26)	14 293	8.62 (8.47 to 8.76)
Bowel cancer	13 286	73.3	8.11 (7.97 to 8.25)	14 736	9.05 (8.90 to 9.19)	17 396	10.7 (10.5 to 10.8)	18 130	11.2 (11.0 to 11.3)
Gastro-oesophageal	6167	73.4	3.76 (3.67 to 3.86)	7307	4.50 (4.39 to 4.60)	8146	5.01 (4.90 to 5.12)	8401	5.17 (5.06 to 5.28)
Lung cancer	12 534	67.3	7.68 (7.54 to 7.81)	16 254	10.05 (9.9 to 10.2)	17 784	10.99 (10.8 to 11.2)	18 617	11.5 (11.4 to 11.7)
Oral cancer	1957	63.2	1.12 (1.07 to 1.17)	2150	1.24 (1.19 to 1.29)	2881	1.67 (1.61 to 1.73)	3096	1.80 (1.73 to 1.86)
Pancreatic cancer	2134	59.2	1.29 (1.23 to 1.34)	2985	1.82 (1.75 to 1.89)	3429	2.09 (2.02 to 2.16)	3607	2.20 (2.12 to 2.27)
Prostate cancer	29 989	85.1	18.8 (18.5 to 19.0)	30 817	19.3 (19.1 to 19.6)	33 287	20.9 (20.7 to 21.1)	35 256	22.2 (22.0 to 22.4)
Renal tract cancer	9106	55.9	5.59 (5.47 to 5.71)	10 454	6.46 (6.33 to 6.58)	15 568	9.61 (9.46 to 9.76)	16 286	10.1 (9.9 to 10.2)
Total cancers	84 787	72.0		95 933		111 990		117 686	

Patients with existing diagnoses of each cancer at baseline were dropped from the relevant cohort.
Rates were age standardised to the overall QResearch population in 5 year age bands.
GP, general practitioner.

Table 4 Adjusted HRs with 95% CIs for seven cancers which occur in men and women in the derivation cohort (blood, bowel, gastro-oesophageal, lung, oral, pancreatic and renal)

Cancer type	Adjusted HRs in women (95% CI)	Adjusted HRs in men (95% CI)
Blood*		
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.01 (0.96 to 1.06)	1.02 (0.98 to 1.07)
Light smoker (1–9/day)	1.20 (1.12 to 1.28)	1.10 (1.05 to 1.17)
Moderate smoker (10–19/day)	1.22 (1.12 to 1.33)	1.17 (1.08 to 1.27)
Heavy smoker (20+/day)	1.33 (1.20 to 1.47)	1.23 (1.14 to 1.34)
Other baseline factors		
Type 1 diabetes	1.51 (1.03 to 2.22)	1.61 (1.21 to 2.13)
Family history of blood cancer	4.08 (1.53 to 10.87)	3.92 (1.47 to 10.45)
Prior brain cancer	4.12 (1.03 to 16.49)	NS
Prior ovarian cancer	1.59 (1.08 to 2.33)	NA
Prior renal cancer	NS	1.46 (1.18 to 1.80)
Bowel†		
Ethnic group		
White/not recorded	1.00	1.00
Indian	0.35 (0.25 to 0.49)	0.56 (0.44 to 0.70)
Pakistani	0.47 (0.28 to 0.78)	0.56 (0.38 to 0.81)
Bangladeshi	0.85 (0.54 to 1.34)	0.41 (0.25 to 0.69)
Other Asian	0.59 (0.40 to 0.89)	0.62 (0.44 to 0.87)
Caribbean	0.71 (0.56 to 0.90)	0.70 (0.57 to 0.86)
Black African	0.69 (0.49 to 0.99)	0.75 (0.55 to 1.03)
Chinese	0.61 (0.35 to 1.05)	0.81 (0.51 to 1.28)
Other	0.80 (0.62 to 1.03)	0.59 (0.45 to 0.77)
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.07 (1.03 to 1.12)	1.06 (1.02 to 1.09)
Light smoker (1–9/day)	1.11 (1.04 to 1.17)	1.07 (1.01 to 1.13)
Moderate smoker (10–19/day)	1.21 (1.12 to 1.31)	1.03 (0.93 to 1.13)
Heavy smoker (20+/day)	1.17 (1.06 to 1.30)	1.13 (1.05 to 1.22)
Alcohol		
Non-drinker	1.00	1.00
Trivial drinker (<1 unit/day)	1.02 (0.98 to 1.06)	1.05 (1.00 to 1.10)
Light drinker (1–2 units/day)	1.05 (1.00 to 1.11)	1.14 (1.08 to 1.20)
Moderate drinker (3–6 units/day)	1.08 (1.01 to 1.16)	1.30 (1.24 to 1.36)
Heavy drinker (7–9 units/day)	1.38 (1.00 to 1.90)	1.62 (1.47 to 1.79)
Very heavy drinker (>9 units/day)	1.36 (0.80 to 2.32)	1.56 (1.33 to 1.83)
Other baseline factors		
Family history of bowel cancer	1.94 (1.62 to 2.32)‡	2.18 (1.84 to 2.59)‡
Ulcerative colitis	1.75 (1.48 to 2.08)	1.83 (1.58 to 2.11)
Colonic polyp	2.11 (1.62 to 2.76)	1.51 (1.19 to 1.90)
Type 2 diabetes	1.16 (1.07 to 1.26)	1.27 (1.20 to 1.35)
Prior breast cancer	1.16 (1.05 to 1.29)	NA
Prior uterine cancer	1.61 (1.24 to 2.11)	NA
Prior ovarian cancer	1.98 (1.48 to 2.65)	NA
Prior cervical cancer	1.74 (1.36 to 2.21)	NA
Prior lung cancer	NS	1.87 (1.32 to 2.65)
Prior blood cancer	NS	1.53 (1.26 to 1.87)
Prior oral cancer	NS	1.62 (1.06 to 2.49)
Gastro-oesophageal§		
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.09 (0.99 to 1.19)	1.25 (1.18 to 1.32)
Light smoker (1–9/day)	1.78 (1.62 to 1.96)	1.73 (1.61 to 1.85)
Moderate smoker (10–19/day)	2.05 (1.82 to 2.31)	1.85 (1.68 to 2.04)
Heavy smoker (20+/day)	2.40 (2.08 to 2.78)	1.91 (1.73 to 2.11)

Continued

Table 4 Continued

Cancer type	Adjusted HRs in women (95% CI)	Adjusted HRs in men (95% CI)
Alcohol		
Non-drinker	1.00	1.00
Trivial drinker (<1 unit/day)	0.89 (0.83 to 0.96)	0.94 (0.88 to 1.00)
Light drinker (1–2 units/day)	0.91 (0.81 to 1.02)	0.89 (0.83 to 0.95)
Moderate drinker (3–6 units/day)	0.98 (0.86 to 1.11)	0.94 (0.88 to 1.00)
Heavy drinker (7–9 units/day)	2.00 (1.28 to 3.13)	1.24 (1.08 to 1.41)
Very heavy drinker (>9 units/day)	2.00 (0.99 to 4.01)	1.63 (1.33 to 1.98)
Other baseline factors		
Barratt's oesophagus	3.83 (2.63 to 5.57)	4.05 (3.30 to 4.97)
Peptic ulcer disease	1.29 (1.12 to 1.49)	1.25 (1.15 to 1.36)
Type 2 diabetes	1.33 (1.17 to 1.52)	1.18 (1.08 to 1.29)
Prior lung cancer	2.28 (1.02 to 5.08)	NS
Prior blood cancer	2.14 (1.43 to 3.20)	NS
Prior breast cancer	1.31 (1.09 to 1.56)	NA
Prior oral cancer	3.84 (1.92 to 7.70)	2.65 (1.65 to 4.27)
Prior pancreatic cancer	NS	4.16 (1.04 to 16.65)
Lung¶		
Ethnic group		
White/not recorded	1.00	1.00
Indian	0.54 (0.36 to 0.80)	0.36 (0.27 to 0.48)
Pakistani	0.37 (0.16 to 0.82)	0.40 (0.27 to 0.60)
Bangladeshi	1.18 (0.74 to 1.87)	0.84 (0.65 to 1.10)
Other Asian	0.69 (0.43 to 1.10)	0.36 (0.24 to 0.57)
Caribbean	0.52 (0.37 to 0.71)	0.58 (0.47 to 0.71)
Black African	0.55 (0.32 to 0.95)	0.45 (0.29 to 0.69)
Chinese	1.16 (0.69 to 1.96)	0.64 (0.40 to 1.03)
Other	0.62 (0.45 to 0.85)	0.51 (0.39 to 0.67)
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.84 (1.60 to 2.13)	2.37 (2.04 to 2.76)
Light smoker (1–9/day)	5.67 (5.08 to 6.33)	5.93 (5.19 to 6.78)
Moderate smoker (10–19/day)	6.57 (5.84 to 7.40)	7.13 (6.15 to 8.27)
Heavy smoker (20+/day)	11.02 (9.84 to 12.33)	10.24 (8.94 to 11.74)
Alcohol		
Non-drinker		1.00
Trivial drinker (<1 unit/day)	NS	0.91 (0.87 to 0.95)
Light drinker (1–2 units/day)	NS	0.92 (0.87 to 0.97)
Moderate drinker (3–6 units/day)	NS	0.97 (0.93 to 1.02)
Heavy drinker (7–9 units/day)	NS	1.13 (1.03 to 1.25)
Very heavy drinker (>9 units/day)	NS	1.25 (1.09 to 1.43)
Other baseline factors		
Family history of lung cancer	1.32 (1.10 to 1.58)	1.28 (1.08 to 1.52)
Asthma	1.33 (1.24 to 1.41)	1.18 (1.11 to 1.26)
Chronic obstructive pulmonary disease	1.97 (1.80 to 2.15)	1.92 (1.76 to 2.08)
Asbestos	NS	1.85 (1.54 to 2.22)
Prior renal cancer	1.74 (1.31 to 2.31)	1.50 (1.30 to 1.74)
Prior blood cancer	1.93 (1.51 to 2.48)	1.91 (1.61 to 2.26)
Prior oral cancer	2.83 (1.76 to 4.56)	2.86 (2.15 to 3.80)
Prior breast cancer	1.53 (1.39 to 1.69)	NA
Prior uterine cancer	1.53 (1.12 to 2.09)	NA
Prior ovarian cancer	1.64 (1.17 to 2.29)	NA
Prior cervical cancer	1.58 (1.26 to 1.97)	NA
Prior bowel cancer	NS	1.29 (1.09 to 1.52)
Prior gastro-oesophageal cancer	NS	1.79 (1.28 to 2.49)
Oral**		
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.22 (1.06 to 1.41)	1.15 (1.04 to 1.28)

Continued

Table 4 Continued

Cancer type	Adjusted HRs in women (95% CI)	Adjusted HRs in men (95% CI)
Light smoker (1–9/day)	2.12 (1.83 to 2.46)	2.31 (2.09 to 2.56)
Moderate smoker (10–19/day)	2.52 (2.11 to 3.00)	2.35 (2.04 to 2.70)
Heavy smoker (20+/day)	3.52 (2.86 to 4.33)	2.95 (2.60 to 3.35)
Alcohol		
Non-drinker	1.00	1.00
Trivial drinker (<1 unit/day)	1.03 (0.91 to 1.16)	0.89 (0.79 to 1.00)
Light drinker (1–2 units/day)	1.18 (0.99 to 1.40)	1.02 (0.90 to 1.15)
Moderate drinker (3–6 units/day)	1.60 (1.35 to 1.90)	1.36 (1.22 to 1.53)
Heavy drinker (7–9 units/day)	2.86 (1.66 to 4.91)	2.59 (2.18 to 3.09)
Very heavy drinker (>9 units/day)	4.38 (2.25 to 8.52)	3.71 (2.99 to 4.59)
Other baseline factors		
Prior blood cancer	4.54 (2.73 to 7.56)	2.34 (1.51 to 3.63)
Prior bowel cancer	NS	1.62 (1.00 to 2.61)
Prior lung cancer	NS	2.87 (1.29 to 6.40)
Prior ovarian cancer	4.14 (2.15 to 7.97)	NA
Pancreas††		
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.03 (0.94 to 1.13)	1.09 (1.00 to 1.18)
Light smoker (1–9/day)	1.77 (1.59 to 1.97)	1.56 (1.41 to 1.73)
Moderate smoker (10–19/day)	1.89 (1.65 to 2.17)	1.96 (1.70 to 2.27)
Heavy smoker (20+/day)	2.02 (1.71 to 2.39)	1.94 (1.68 to 2.24)
Other baseline factors		
Chronic pancreatitis	3.64 (2.11 to 6.28)	5.43 (3.72 to 7.94)
Type 2 diabetes	1.51 (1.31 to 1.74)	1.85 (1.65 to 2.07)
Prior renal cancer	1.97 (1.14 to 3.40)	NS
Prior breast cancer	1.38 (1.13 to 1.67)	NA
Prior blood cancer	NS	1.71 (1.12 to 2.60)
Renal Tract‡‡		
Smoking status		
Non-smoker	1.00	1.00
Ex-smoker	1.27 (1.19 to 1.37)	1.23 (1.18 to 1.29)
Light smoker (1–9/day)	1.74 (1.61 to 1.89)	1.64 (1.56 to 1.72)
Moderate smoker (10–19/day)	2.23 (2.02 to 2.45)	2.05 (1.91 to 2.20)
Heavy smoker (20+/day)	2.35 (2.08 to 2.64)	2.24 (2.09 to 2.40)
Other baseline factors		
Type 2 diabetes	1.34 (1.21 to 1.50)	1.21 (1.13 to 1.28)
Prior bowel cancer	1.44 (1.07 to 1.92)	1.25 (1.05 to 1.50)
Prior lung cancer	NS	1.78 (1.23 to 2.58)
Prior prostate cancer	NA	1.46 (1.26 to 1.68)
Prior blood cancer	1.63 (1.11 to 2.40)	NS
Prior brain cancer	10.18 (3.28 to 31.58)	NS
Prior uterine cancer	2.12 (1.51 to 2.98)	NA
Prior ovarian cancer	2.62 (1.81 to 3.77)	NA
Prior cervical cancer	2.56 (1.94 to 3.40)	NA

For fractional polynomial terms and interactions see footnotes and figures 1–5.
*Blood cancer models also included age (2 FP terms) and body mass index (linear) in women and men.
†Bowel cancer models also included age (2 FP terms) and interaction between age and family history in women, and age (2 FP terms), BMI (linear,positive), Townsend (linear, positive) and interaction between age and family history in men.
‡Adjusted HR evaluated at mean age.
§Gastro-oesophageal models also included age (2 FP terms), BMI (2 FP terms) and Townsend (linear, positive) in women and men.
¶Lung cancer models included age (2 FP terms), Townsend (2 FP terms) and BMI (2 FP terms) in women and men?
**Oral cancer models included age (1 FP term for women, 2 FP terms for men), Townsend (linear and positive in women and men) and BMI (2 FP terms for men only).
††Pancreatic cancer models included age (1 FP term in women and men), BMI (linear for women, 2 FP terms for men) and Townsend (linear and positive—women only).
‡‡Renal cancer models included age (2 FP terms in women and men), Townsend (2 FP terms women) and BMI (linear and positive in women and men).
NA, not applicable; NS, not significant.

Table 5 Adjusted HRs with 95% CIs for cancers occurring in women in the derivation cohort (breast, ovary, uterus)

Cancer type	Adjusted HRs (95% CI)
Breast cancer*	
Ethnic group	
White/not recorded	1.00
Indian	0.73 (0.64 to 0.83)
Pakistani	0.71 (0.58 to 0.88)
Bangladeshi	0.39 (0.27 to 0.56)
Other Asian	0.78 (0.66 to 0.93)
Caribbean	0.82 (0.73 to 0.93)
Black African	0.75 (0.63 to 0.88)
Chinese	0.73 (0.56 to 0.94)
Other	0.83 (0.73 to 0.94)
Alcohol	
Non-drinker	1.00
Trivial drinker (<1 unit/day)	1.05 (1.03 to 1.08)
Light drinker (1–2 units/day)	1.11 (1.07 to 1.15)
Moderate drinker (3–6 units/day)	1.21 (1.16 to 1.26)
Heavy drinker (7–9 units/day)	1.31 (1.07 to 1.61)
Very heavy drinker (>9 units/day)	1.25 (0.92 to 1.71)
Other baseline factors	
Family history of breast cancer	1.93 (1.83 to 2.04)†
Benign breast disease	1.51 (1.45 to 1.57)
Current oral contraceptive	1.13 (1.07 to 1.20)
Current oestrogen containing HRT	1.18 (1.15 to 1.22)
Manic depression or schizophrenia	1.16 (1.04 to 1.30)
Prior lung cancer	1.86 (1.21 to 2.85)
Prior blood cancer	1.57 (1.31 to 1.88)
Prior ovarian cancer	1.42 (1.12 to 1.80)
Ovarian cancer‡	
Family history of ovarian cancer	3.81 (2.72 to 5.33)†
Current oral contraceptive	0.65 (0.54 to 0.79)
Prior breast cancer	1.62 (1.41 to 1.88)
Prior cervical cancer	1.60 (1.08 to 2.38)
Uterine cancer§	
Smoking status	
Non-smoker	1.00
Ex-smoker	0.82 (0.77 to 0.87)
Light smoker (1–9/day)	0.83 (0.76 to 0.92)
Moderate smoker (10–19/day)	0.74 (0.65 to 0.84)
Heavy smoker (20+/day)	0.66 (0.56 to 0.77)
Other baseline factors	
Manic depression or schizophrenia	1.55 (1.25 to 1.92)
Type 2 diabetes	1.35 (1.21 to 1.49)
Endometrial hyperplasia or polyp	2.35 (1.83 to 3.01)
Polycystic ovarian disease	1.98 (1.43 to 2.76)
Prior bowel cancer	1.56 (1.13 to 2.17)
Prior breast cancer	2.49 (2.22 to 2.79)

For fractional polynomial terms and interactions see footnotes and figures 1–5.
*Breast cancer model also included terms for age (2 FP terms), BMI (2 FP terms), Townsend (2 FP terms) and interaction between age and family history.
†Adjusted HR evaluated at mean age in women.
‡Ovarian cancer model also included terms for age (2 FP terms), BMI (linear and positive), and interaction between age and family history.
§Uterine cancer model also included terms for age (2 FP terms), and BMI (1 FP term).
HRT, hormone replacement therapy.

Table 6 Adjusted HRs with 95% CIs for prostate cancer in men in the derivation cohort

Prostate cancer*	Adjusted HR (95% CI)
Ethnic group	
White/not recorded	1.00
Indian	0.60 (0.51 to 0.71)
Pakistani	0.42 (0.30 to 0.58)
Bangladeshi	0.29 (0.18 to 0.47)
Other Asian	0.46 (0.34 to 0.62)
Caribbean	2.84 (2.61 to 3.09)
Black African	1.98 (1.69 to 2.33)
Chinese	0.50 (0.32 to 0.76)
Other	1.48 (1.30 to 1.69)
Smoking status	
Non-smoker	1.00
Ex-smoker	1.02 (0.92 to 1.12)†
Light smoker (1–9/day)	0.78 (0.70 to 0.87)†
Moderate smoker (10–19/day)	0.74 (0.63 to 0.87)†
Heavy smoker (20+/day)	0.79 (0.68 to 0.91)†
Other baseline factors	
Family history of prostate cancer	7.65 (6.11 to 9.57)†
Manic depression or schizophrenia	0.64 (0.54 to 0.77)
Type 1 diabetes	0.57 (0.40 to 0.81)
Type 2 diabetes	0.90 (0.85 to 0.94)

For fractional polynomial terms and interactions see footnotes and figures 1–5.
*Prostate cancer model also included terms for age (2 FP terms), BMI (2 FP terms), Townsend (linear and negative) and interactions between age and family history and age and smoking status.
†Adjusted HR evaluated at mean age.

higher risk), previous breast cancer (31% higher risk), and previous oral cancer (3.8-fold higher risk).

The direction and magnitude of the HRs was similar for men except for prior cancers where prior pancreatic cancer was a significant predictor for men not women (fourfold higher risk) and prior lung and blood cancers were significant predictors in women but not in men.

Lung cancer

There were 15 variables in the final model for lung cancer in women. The 15 variables were age, BMI, Townsend deprivation score (figure 3), ethnicity (lower risk among non-white groups), smoking status, family history of lung cancer (32% higher risk), asthma (33% higher risk), chronic obstructive pulmonary disease (97% higher risk), previous blood cancer (93% higher risk), previous breast cancer (53% higher risk), previous cervical cancer (58% higher risk), previous oral cancer (2.8-fold higher risk), previous ovarian cancer (64% higher risk), previous renal cancer (74% higher risk), and previous uterine cancer (53% higher risk). There was a 'dose response' association for risk of lung cancer with smoking status—compared with a non-smoker: at the mean age of 45 years, there was a 5.7-fold higher risk for a light smoker; 6.6-fold higher risk for a moderate smoker and an 11-fold higher risk for a heavy smoker.

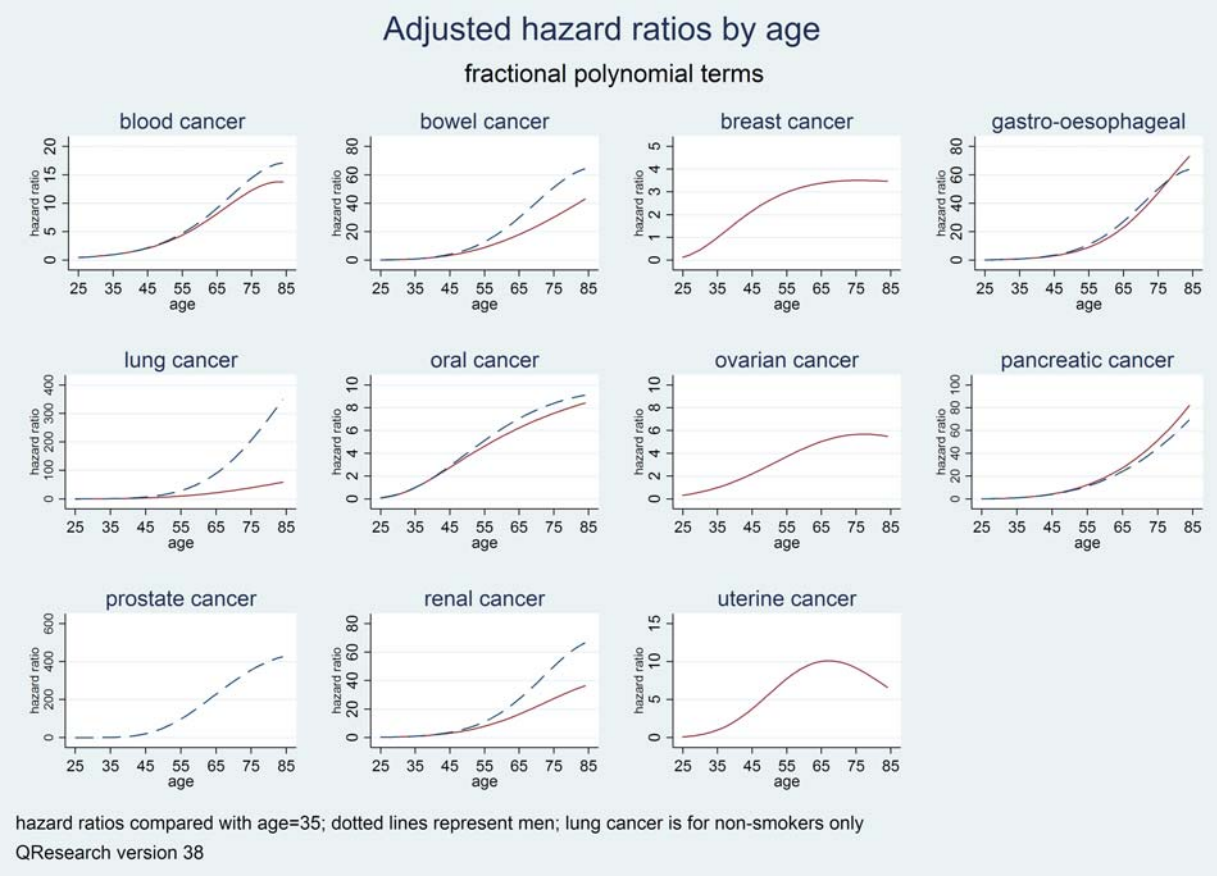

Figure 1 Showing graphs of the adjusted HRs for the fractional polynomial terms for age for each cancer.

There was also an interaction between smoking status and age (figure 5) such that the 'dose response' effect was most marked in women aged 60–70.

The direction and magnitude of the HRs was similar for men except that three additional variables reached significance and were included in the final model (alcohol, prior bowel and gastro-oesophageal cancers).

Oral cancer

There were six variables in the final model for oral cancer in women. These were age, Townsend deprivation score, smoking status (3.5-fold higher risk for heavy smokers) alcohol (4.4-fold higher risk for very heavy drinkers), previous blood cancer (4.5-fold higher risk) and previous ovarian cancer (4.1-fold higher risk).

The direction and magnitude of the HRs was similar for men except BMI, prior bowel and prior lung cancer were significant and so were included in the final model for men but not women.

Ovarian cancer

There were six variables in the final model for ovarian cancer which were age, BMI, family history of ovarian cancer (3.8-fold higher risk at the mean age of 45 years), oral contraceptive use (35% reduced risk), previous breast cancer (62% higher risk) and previous cervical cancer (60% higher risk). There was an interaction between age and family history as shown in (figure 4) with higher HRs at both extremes of age.

Pancreatic cancer

There were eight variables in the final model for pancreatic cancer in women. These were age, BMI, Townsend score, smoking status (twofold higher risk in heavy smokers), chronic pancreatitis (3.6-fold higher risk), type 2 diabetes (51% higher risk), previous breast cancer (38% higher risk) and previous renal cancer (97% higher risk).

The direction and magnitude of the HRs was similar for men except prior blood cancer reached significance and prior renal cancer did not.

Renal tract cancer

There were 11 variables in the final model for renal tract cancer in women. These were age, Townsend deprivation score, BMI, smoking status (2.4-fold higher risk in heavy smokers), type 2 diabetes (34% higher risk), previous blood cancer (63% higher risk), previous brain cancer (10-fold higher risk), previous cervical cancer (2.6-fold higher risk), previous bowel cancer (44% higher risk), previous ovarian cancer (2.6-fold higher risk), previous uterine cancer (2.1-fold higher risk). Increasing deprivation was associated with a lower risk as shown in figure 3.

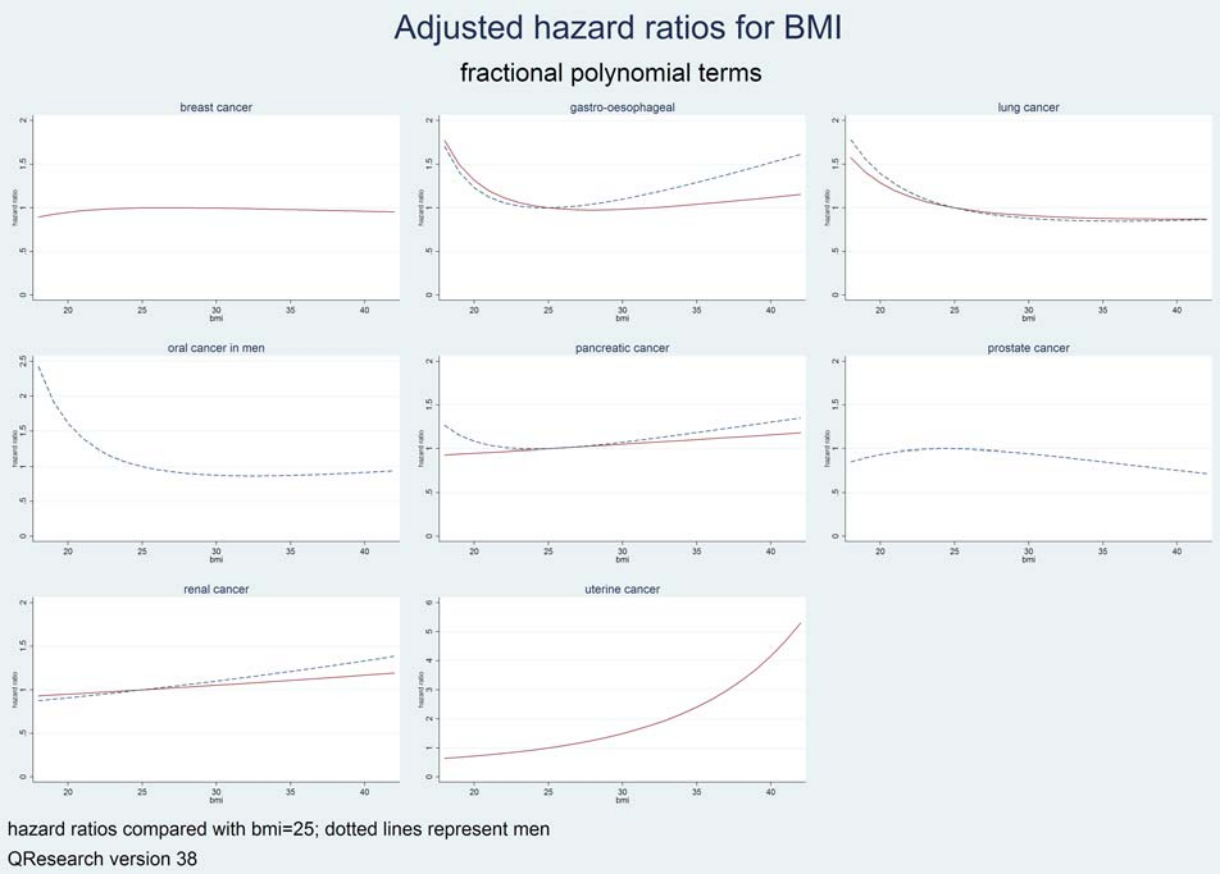

Figure 2 Showing graphs of the adjusted HRs for the fractional polynomial terms for body mass index for each cancer.

The direction and magnitude of the HRs was similar for men except Townsend score was not significant. Also prior lung cancer and prostate cancer were significantly associated with increased risk of renal cancer but prior blood cancer and brain cancer were not.

Uterine cancer

There were nine variables in the final model for uterine cancer. These were age, BMI, smoking status (heavy smokers had a 34% lower risk), manic depression or schizophrenia (55% higher risk), type 2 diabetes (35% higher risk), endometrial hyperplasia or polyp (2.4-fold higher risk), polycystic ovarian syndrome (98% higher risk), previous breast cancer (2.5-fold higher risk) and previous bowel cancer (56% higher risk).

Prostate cancer

There were nine variables in the final model for prostate cancer. These were age, BMI, Townsend deprivation score, ethnicity, smoking status (reduced risks in current smokers at the mean age of 44 years), family history of prostate cancer (7.7-fold higher risk at the mean age of 44 years), manic depression/schizophrenia (36% lower risk), type 1 diabetes (43% lower risk) and type 2 diabetes (10% lower risk). There were marked differences in risk between different ethnic groups with the South Asian and Chinese men having the lowest risks and Black African and Caribbean men having the highest risks. Caribbean men had a 2.8-fold higher risk of prostate cancer than White men.

Figures 1 and 2 show the adjusted HRs for the fractional polynomial terms for age and BMI. Figure 4 shows the HRs for the interaction between age and family history of prostate cancer where HRs for family history are highest among men under the age of 30. Figure 5 shows the interaction between age and smoking status—at younger ages heavy smoking is associated with around a fourfold to fivefold increased risk of prostate cancer then the HRs decrease with advancing age.

Risk of incident cancers in patients with prior cancers

Table 7 summarises the adjusted HRs for each type of cancer associated with different types of prior cancer at baseline. For example, women with an existing diagnosis of breast cancer have a significantly increased risk of bowel cancer (16% higher), lung cancer (53% higher), pancreatic cancer (38% higher), gastro-oesophageal cancer (31% higher), ovarian cancer (62% higher) and uterine cancer (2.5-fold higher). Men with an existing diagnosis of blood cancer have a significantly increased risk of bowel cancer (53% higher), lung cancer (91% higher), oral cancer (2.3-fold higher) and pancreatic cancer (71% higher). The other associations are shown in table 7.

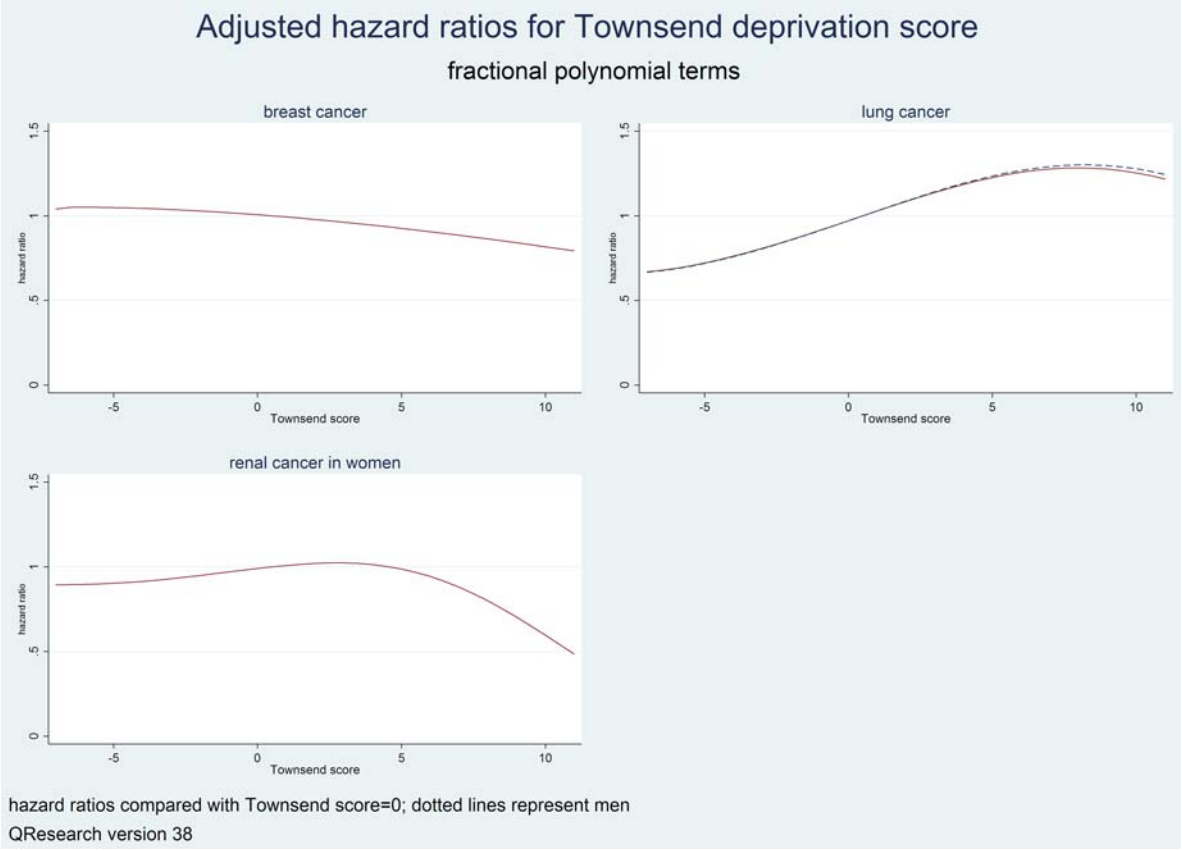

Figure 3 Showing graphs of the adjusted HRs for fractional polynomial terms for Townsend deprivation score for each cancer.

Validation

Discrimination

Table 8 shows the performance of each algorithm in the validation cohort for women and men. The lung cancer algorithm had the highest values for all three performance measures evaluated over 10 years in men and women—the algorithm explained 64.2% of the variation in time to cancer diagnosis in women (R^2), the D statistic was 2.74 and the ROC value was 0.91. Apart from breast cancer and ovarian cancer, the ROC values for all the other algorithms exceeded 0.8 in men and women.

The algorithm for breast cancer had the lowest values with an R^2 value of 22.0%, D statistic of 1.09 and ROC value of 0.76. The performance of the algorithm for ovarian cancer was marginally better than that for breast cancer with an R^2 value of 29.1%, D statistic of 1.31 and ROC value of 0.77.

Performance of the algorithms in men was very similar to that for women. For prostate cancer, the performance was good with an R^2 of 54.8%, D statistic of 2.25 and ROC value of 0.89. The algorithm for oral cancer had the lowest performance among men although the R^2 was 45.8%, D statistic was 1.88 and ROC value was 0.81.

Web extra table 2 shows the performance statistics based on the 1 197 426 (73.7%) of patients in the validation cohort with complete data for BMI, smoking status and alcohol use rather than using imputed values for missing data. The results are similar though the absolute values tend to be marginally lower than the analyses based on imputed data.

Thresholds

Table 9 shows the classification statistics of each algorithm in the full validation cohort for the 10% of women at highest predicted risk of each cancer. For example, for the 10% of women at highest predicted risk of lung cancer (ie, those with a 10-year predicted risk score of 1.43% or higher), the sensitivity was 67.3% and the observed risk at 10 years was 3.9%. For breast cancer, the corresponding sensitivity was 27.5% with an observed risk of 4.4% at 10 years for the top 10% of women at highest predicted risk. Table 10 shows the corresponding results for men. For prostate cancer, for the top 10% of men at highest predicted risk (a 10-year predicted risk score of 5.89% or higher) the sensitivity was 56% and the observed risk at 10 years was 8.4%.

Calibration

Figure 6 shows the mean predicted risks and the observed risks at 10 years within each tenth of predicted risk in order to assess the calibration of the model in women in the validation cohort. Figure 7 shows the corresponding calibration graph for men. There was close correspondence between the mean predicted risks and the observed risks within each model tenth in women and men indicating that the algorithms were well calibrated.

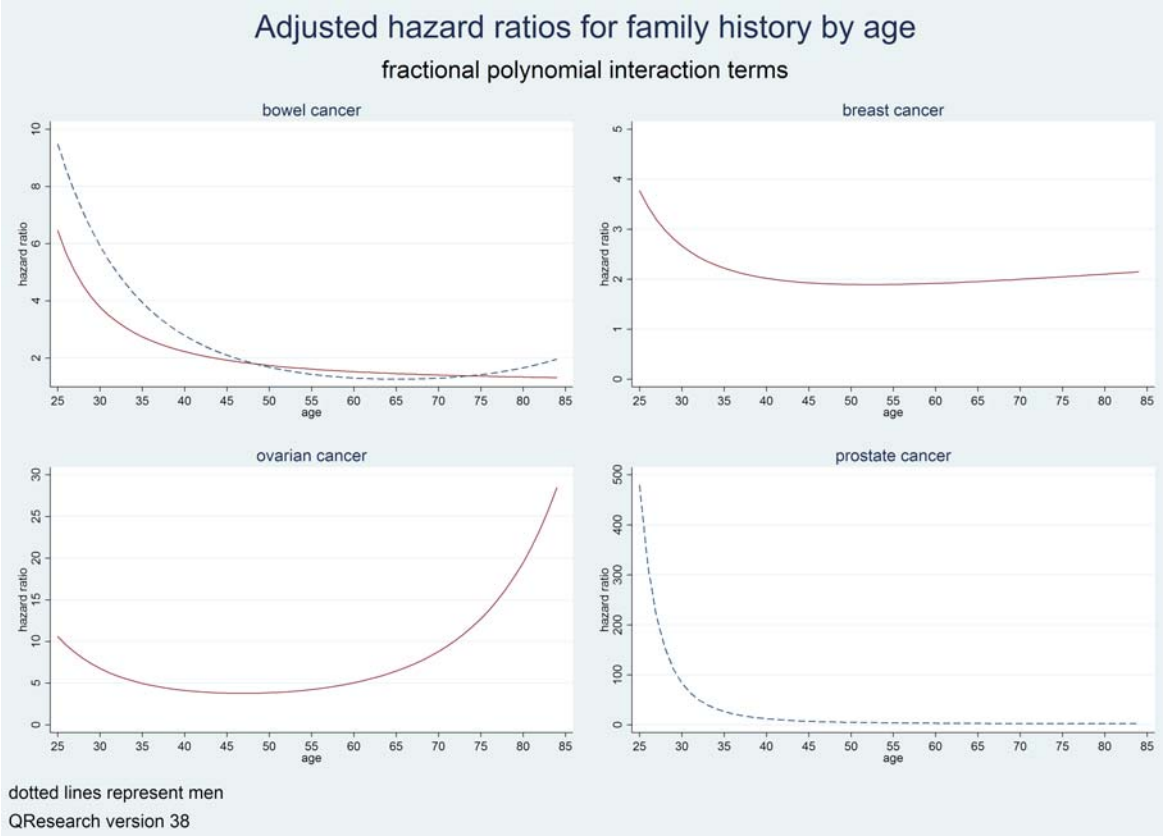

Figure 4 Showing graphs of the adjusted HRs for the interactions between age and family history for each relevant cancer.

Figure 8 shows the web calculator for an example patient which is a 64-year-old man who is a heavy smoker, has type 2 diabetes and a family history of bowel cancer. His 10-year predicted risks of the following cancers are: blood cancer (2%) bowel cancer (5.8%), gastro-oesophageal cancer (3.3%) lung cancer (9.4%), oral cancer (2%), renal cancer (4.6%), pancreatic cancer (1.3%), prostate cancer (3.9%).

DISCUSSION

Summary of key findings

We have developed and validated a series of risk prediction algorithms—collectively known as the QCancer 10 year risk algorithms—to quantify future absolute risk of 10 common cancers in women and eight common cancers in men. The algorithms incorporate predictor variables which are associated with risk of cancer including sociodemographic variables, lifestyle, morbidity, medications, family history and previous diagnoses of other cancers. The algorithms can be applied to any adult aged 25–84 years in a primary care setting regardless of whether they have had a prior cancer. For nine of the 11 cancers, the ROC values exceeded 0.8 which is generally considered to be very good. For two of the cancers in women (breast and ovarian) the ROC values were lower at 0.76 and 0.77 though this is still considered acceptable.

The QCancer 10 year risk algorithms are designed to quantify *future* risk of cancer and differ from the existing QCancer algorithms[3–10] which combine symptoms such as appetite loss and abdominal pain with risk factors to quantity the absolute risk that a patient has an *existing* cancer as yet undiagnosed to help inform the decision for further diagnostic tests. These algorithms to predict *existing* cancer are designed to be used when patients present with symptoms that might be indicative of cancer.

Comparisons with the literature

We included established predictor variables in our analysis and found that our HRs were similar in magnitude and direction to those reported in other studies. This increases the clinical face validity of the variables included as predictors in the final algorithms. We have summarised the key relationships below from the perspective of the relevant set of risk factors (rather than by each cancer as described in our results section above).

Smoking and cancer risk

We found that smoking was associated with a significantly increased risk of seven cancers (blood, bowel,[33] gastro-oesophageal,[34 35] lung,[36] oral,[37] pancreatic[38 39] and renal tract[18]) with some evidence of a 'dose

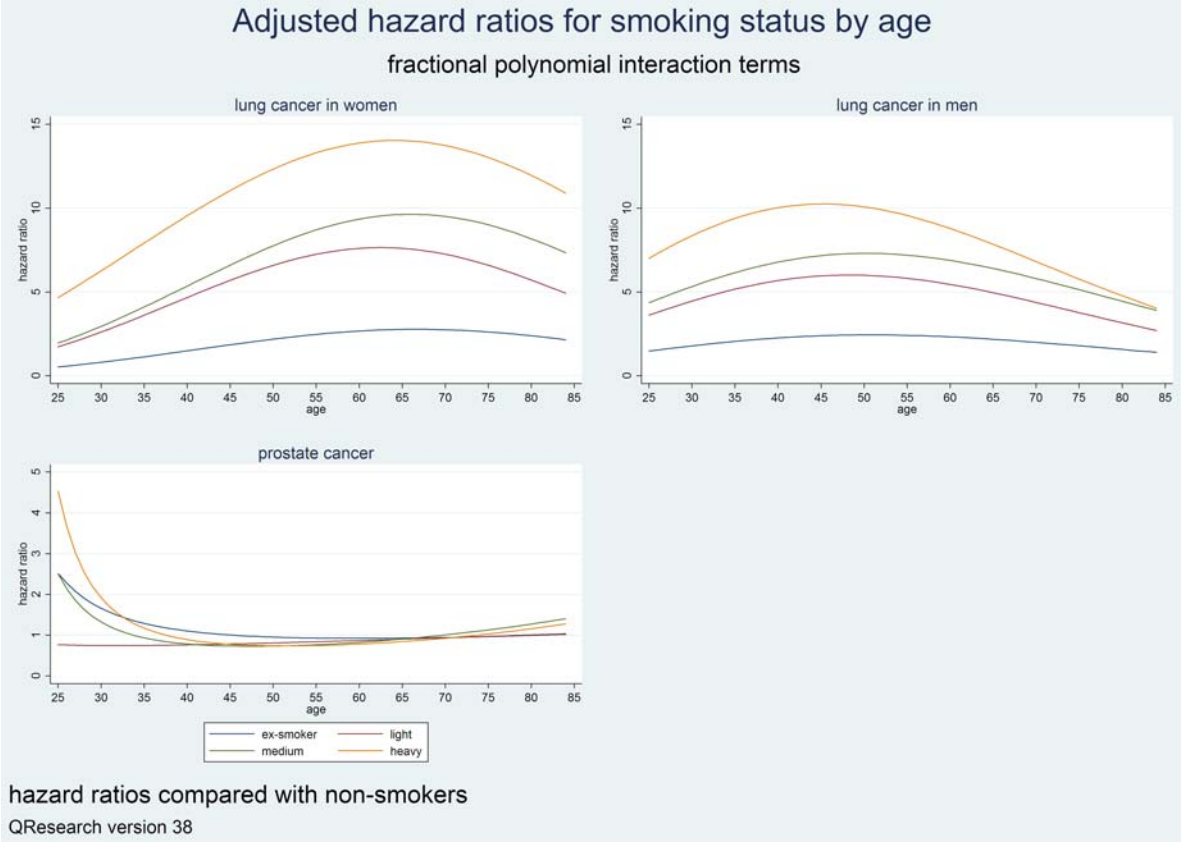

Figure 5 Showing graphs of the adjusted HRs for the interactions between age and smoking status for each relevant cancer.

response' relationship with higher levels of smoking associated with higher levels of risk. Smoking was associated with a decreased risk of uterine cancer which is consistent with other studies.[40] We found no significant association between smoking and ovarian or breast cancer. We found an interaction between age and smoking for risk of prostate cancer. At younger ages heavy smoking is associated with around a fourfold to fivefold increased risk of prostate cancer then the HRs decrease with advancing age.

Alcohol and cancer risk

Alcohol intake was associated with an increased risk of four cancers (oral,[41] breast[41 42] bowel,[41] gastro-oesophageal[34 41 43]) with a tendency for higher volumes of alcohol consumption to be associated with higher levels of cancer risk. We found no significant association between alcohol and lung cancer in women but there was an association for men.

Family history and cancer risk

Family history was associated with a higher risk of six cancers (prostate,[44] breast[45 46] bowel[47 48] blood, lung[49] and ovarian cancer[50 51]). The magnitude of the risk of ovarian cancer associated with a family history of ovarian cancer in this study was 3.8 fold higher at the mean age which is similar to the threefold to fourfold higher risk reported elsewhere.[50 51]

Hormonal treatments and cancer risk

There was a small increased risk of breast cancer with oral contraceptive use[52] and oestrogen containing HRT.[53 54] There was a decreased risk of ovarian cancer associated with use of oral contraceptives in line with other studies.[55 56] There was no clear association between use of the oral contraceptive pill or HRT for the remaining cancers.

Comorbidities and cancer risk

Ulcerative colitis and previous colonic polyps were associated with a higher risk of bowel cancer.[57 58] Barratt's oesophagus was associated with a fourfold higher risk of gastro-oesophageal cancer. This is consistent with, but lower than the 11-fold increase in risk of adenocarcinoma of the oesophagus.[59] Chronic pancreatitis was associated with a higher risk of pancreatic cancer.[60] We found a higher risk of breast cancer among patients with benign breast disease.[61] Manic depression or schizophrenia was associated with a higher risk of uterine cancer and a marginal higher risk of breast cancer and a reduced risk of prostate cancer.[62] There was a twofold higher risk of uterine cancer with polycystic ovarian disease in line with other studies.[63]

Diabetes and cancer risk

Type 1 diabetes was associated with a higher risk of blood cancer. Type 2 diabetes was associated with a

Table 7 Summary of adjusted HRs in the derivation cohort for risk of future cancers significantly associated with prior cancers at baseline

Women	Blood	Bowel	Lung	Oral	Pancreas	Renal tract	Gastro-oesophageal	Breast	Ovarian	Uterine cancer
Prior blood cancer	NA	–	1.93	4.54	–	1.63	2.14	1.57	–	–
Prior bowel cancer	–	NA	–	–	–	1.44	–	–	–	1.56
Prior brain cancer	4.12	–	–	–	–	10.18	–	–	–	–
Prior breast cancer	–	1.16	1.53	–	1.38	–	1.31	NA	1.62	2.49
Prior cervical cancer	–	1.74	1.58	–	–	2.56	–	–	1.60	–
Prior lung cancer	–	–	NA	–	–	–	2.28	1.86	–	–
Prior oral cancer	–	–	2.83	NA	–	–	3.84	–	–	–
Prior ovarian cancer	1.59	1.98	1.64	4.14	–	2.62	–	1.42	NA	–
Prior renal cancer	–	–	1.74		1.97	NA	–	–	–	–
Prior uterine cancer	–	1.61	1.53	–	–	2.12	–	–	–	NA
Men	**Blood**	**Bowel**	**Lung**	**Oral**	**Pancreas**	**Renal tract**	**Gastro-oesophageal**	**Prostate**		
Prior blood cancer	NA	1.53	1.91	2.34	1.71	–	–	–		
Prior bowel cancer	–	NA	1.2	1.62	–	1.25	–	–		
Prior gastro-oesophageal cancer	–	–	1.79	–	–	–	NA	–		
Prior lung cancer	–	1.87	NA	2.87	–	1.78	–	–		
Prior oral cancer	–	1.62	2.86	NA	–	–	2.65	–		
Prior pancreatic cancer	–	–	–	–	NA	–	4.16	–		
Prior prostate cancer	–	–	–	–	–	1.46	–	NA		
Prior renal cancer	1.46	–	1.50	–	–	NA	–	–		

NA, not applicable.

higher risk of five cancers (bowel,[64–67] gastro-oesophageal,[68] pancreatic,[69 70] renal tract[69] and uterine cancer[69 71]). Both type 1 and type 2 diabetes were associated with a reduced risk of prostate cancer in line with previous studies.[72 73]

Ethnicity and cancer risk

Ethnic groups other than the white/not recorded group tended to be associated with decreased risk of three cancers (breast cancer, bowel cancer, lung cancer) which is consistent with other studies.[74] We found no significant association between ethnicity and risk of the other cancers included in our study except for prostate cancer where black African and Caribbean men had significantly higher risks compared with white men, and South Asian and Chinese men had significantly lower risks. This is consistent with other studies examining risk of prostate cancer among different ethnic groups.[75]

Prior cancers

We have identified and quantified a number of associations between previous cancers and risk of future cancer by cancer type. We think our findings have reasonable face validity as they are consistent with those reported elsewhere[76 77] and we have been able to adjust for potential confounding variables. For example, we found that a previous diagnosis of lung cancer in men was associated with an increased risk of three cancers (oral, bowel and renal tract).[77] Also, we found that a previous diagnosis of blood cancer in women was associated with an increased risk of five cancers (lung, oral, renal tract, gastro-oesophageal and breast cancer).[77 78] Similarly we found that prior breast cancer was associated with an increased risk of six cancers (uterine, bowel, lung, pancreatic, gastro-oesophageal and ovarian cancer[76 77]). We found that prior ovarian cancer was associated with an increased risk of five cancers (bowel, lung, oral, renal tract and breast). Some of the associations between prior cancer and risk of future cancer may reflect common aetiologies between different cancers not fully adjusted for in our multivariate model (eg, lifestyle factors or genetic predisposition). Alternatively some may represent secondary cancers directly related to the first but which have not been correctly coded as metastases. It is important to note that apparent lack of associations between some types of prior cancer and future cancer may reflect small numbers especially where specific cancers are rare and/or have a poor 5-year survival (such as pancreatic cancer). Additional research would be needed to determine the potential utility of enhanced screening among patients with an existing malignancy who are at increased risk of a second primary cancer.

Thresholds

Generally we envisage that cancer risk prediction values would be kept continuous for assessment of an individual although at some point there needs to be a cut-off if

Table 8 Performance of each algorithm in the validation cohort in men and women (including patients with imputed data)

Statistic	Women: mean (95% CI)	Men: mean (95% CI)
Blood cancer		
D statistic	1.639 (1.578 to 1.699)	1.726 (1.673 to 1.78)
R^2 (%)	39.1 (37.3 to 40.8)	41.6 (40.1 to 43.1)
ROC statistic	0.803 (0.796 to 0.811)	0.8 (0.793 to 0.807)
Breast cancer		
D statistic	1.088 (1.058 to 1.119)	NA
R^2 (%)	22 (21.1 to 23)	NA
ROC statistic	0.761 (0.758 to 0.765)	NA
Bowel cancer		
D statistic	1.974 (1.922 to 2.027)	2.139 (2.091 to 2.188)
R^2 (%)	48.2 (46.9 to 49.5)	52.2 (51.1 to 53.3)
ROC statistic	0.847 (0.842 to 0.852)	0.862 (0.858 to 0.866)
Gastro-oesophageal cancer		
D statistic	2.277 (2.181 to 2.372)	2.241 (2.174 to 2.308)
R^2 (%)	55.3 (53.2 to 57.4)	54.5 (53 to 56)
ROC statistic	0.873 (0.864 to 0.881)	0.868 (0.862 to 0.874)
Lung cancer		
D statistic	2.742 (2.687 to 2.797)	2.797 (2.75 to 2.844)
R^2 (%)	64.2 (63.3 to 65.1)	65.1 (64.4 to 65.9)
ROC statistic	0.905 (0.901 to 0.91)	0.911 (0.908 to 0.914)
Oral cancer		
D statistic	1.817 (1.67 to 1.964)	1.881 (1.771 to 1.991)
R^2 (%)	44.1 (40.1 to 48.1)	45.8 (42.9 to 48.7)
ROC statistic	0.795 (0.775 to 0.814)	0.808 (0.794 to 0.823)
Ovarian cancer		
D statistic	1.311 (1.237 to 1.385)	NA
R^2 (%)	29.1 (26.8 to 31.4)	NA
ROC statistic	0.769 (0.76 to 0.778)	NA
Pancreas cancer		
D statistic	2.235 (2.126 to 2.345)	2.225 (2.119 to 2.33)
R^2 (%)	54.4 (52 to 56.8)	54.2 (51.8 to 56.5)
ROC statistic	0.865 (0.855 to 0.875)	0.857 (0.847 to 0.867)
Prostate cancer		
D statistic	NA	2.252 (2.219 to 2.285)
R^2 (%)	NA	54.8 (54 to 55.5)
ROC statistic	NA	0.895 (0.893 to 0.897)
Renal tract cancer		
D statistic	2.005 (1.923 to 2.086)	2.234 (2.181 to 2.287)
R^2 (%)	49 (46.9 to 51)	54.4 (53.2 to 55.5)
ROC statistic	0.851 (0.843 to 0.859)	0.863 (0.858 to 0.867)
Uterine cancer		
D statistic	1.758 (1.677 to 1.839)	NA
R^2 (%)	42.5 (40.2 to 44.7)	NA
ROC statistic	0.828 (0.819 to 0.837)	NA

Notes on understanding validation statistics.
Discrimination is the ability of the risk prediction model to differentiate between patients who experience a admission event during the study and those who do not. This measure is quantified by calculating the area under the receiver operating characteristic curve (ROC) statistic; where a value of 1 represents perfect discrimination.
The D statistic is also a measure of discrimination which is specific to censored survival data. As with the ROC, higher values indicate better discrimination.
R^2 measures explained variation and higher values indicate more variation is explained.
NA, not applicable.

a clinician is going to take action for an individual patient. At national level, policymakers and commissioners tend to make recommendations around absolute thresholds to ensure equitable access and consistent management across a health community. This is already common place for cardiovascular disease risk where the latest National Institute for Health and Care Excellence (NICE) guidelines from 2014 recommend treatment with statins for patients with a 10-year cardiovascular disease risk above 10%.[79] However, in this paper, we have not provided definite comment on what threshold of absolute risk should be used to define 'high risk'

Table 9 Classification statistics for each algorithm in the validation cohort based on the top 10% of patients at highest predicted risk of each cancer in women

Type of cancer	Cut-off 10 year risk (top 10%)	Sensitivity (%)	Specificity (%)	Observed risk at 10 years (%)
Blood cancer	1.48	38.1	90.1	1.88
Bowel cancer	2.04	45.5	90.3	2.97
Breast cancer	3.40	27.5	90.2	4.42
Gastro-oesophageal cancer	0.65	53.3	90.1	1.07
Lung cancer	1.43	67.3	90.2	3.91
Oral cancer	0.17	42.3	90.0	0.32
Ovarian cancer	0.70	28.1	90.0	0.87
Pancreas cancer	0.50	50.0	90.0	0.76
Renal tract cancer	0.94	46.6	90.1	1.37
Uterine cancer	0.72	42.6	90.1	1.20

since that would require (1) consideration of the balance of risk/benefit for an individual and their choice and (2) cost-effectiveness analyses which are outside the scope of this study. We have, however, provided analyses using the top 10% of absolute risk as a threshold of risk which can be used to help inform future analyses. Sensitivity is important as it is a measure of how well the algorithm performs in finding cases that might be suitable for intervention. If the risk threshold is set too high, then the sensitivity will be low and a large number of patients at increased risk of cancer will be 'missed' by the algorithm. Although a high-risk threshold is likely to result in a higher positive predictive value which means a higher proportion of those identified are likely to go on to develop cancer over the next 10 years. A lower risk threshold would have higher sensitivity but could lead to unnecessary interventions and anxiety in people who will not develop cancer over 10 years. So, at the population level, there is a balance to be struck between the sensitivity and positive predictive value of the score which depends on the risk threshold selected, resources available; likely risks and benefits of the interventions and the chance that patients might become anxious about being classified as 'high risk'.

Methodological considerations

The methods to derive and validate these models are the same as for a range of other clinical risk prediction tools derived from the QResearch database.[19 80–83] The strengths and limitations of the approach have already been discussed in detail[21 80 83–86] including information on multiple imputation of missing data. In summary, key strengths include size, duration of follow-up, representativeness, and lack of selection, recall and respondent bias. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and prescribed medications.[87] We think our 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. Our database has linked hospital, mortality and cancer records for nearly all patients and is therefore likely to have picked up the majority of cancer diagnoses thereby minimising ascertainment bias. We excluded patients without a valid deprivation score since this group may represent a more transient population where follow-up could be unreliable or unrepresentative. Their deprivation scores are unlikely to be missing at random so we did not think it would be appropriate to impute them.

The present validation has been carried out on a completely separate set of practices and individuals to those which were used to develop the score although the practices all use the same GP clinical computer system (EMIS—the computer system used by 55% of UK GPs). An independent validation study would be a more stringent test and should be carried out, but when such

Table 10 Classification statistics for each algorithm in the validation cohort based on the top 10% of patients at highest predicted risk of each cancer in men

Type of cancer	Cut-off 10 year risk (top 10%)	Sensitivity (%)	Specificity (%)	Observed risk at 10 years (%)
Blood cancer	1.96	40.0	90.1	2.75
Bowel cancer	2.82	50.6	90.2	4.20
Lung cancer	2.73	66.6	90.3	6.17
Gastro-oesophageal	1.34	52.9	90.1	2.27
Oral cancer	0.33	46.3	90.0	0.70
Pancreas cancer	0.52	49.3	90.0	0.86
Prostate cancer	5.89	56.2	90.4	8.42
Renal tract cancer	2.54	50.4	90.2	3.80

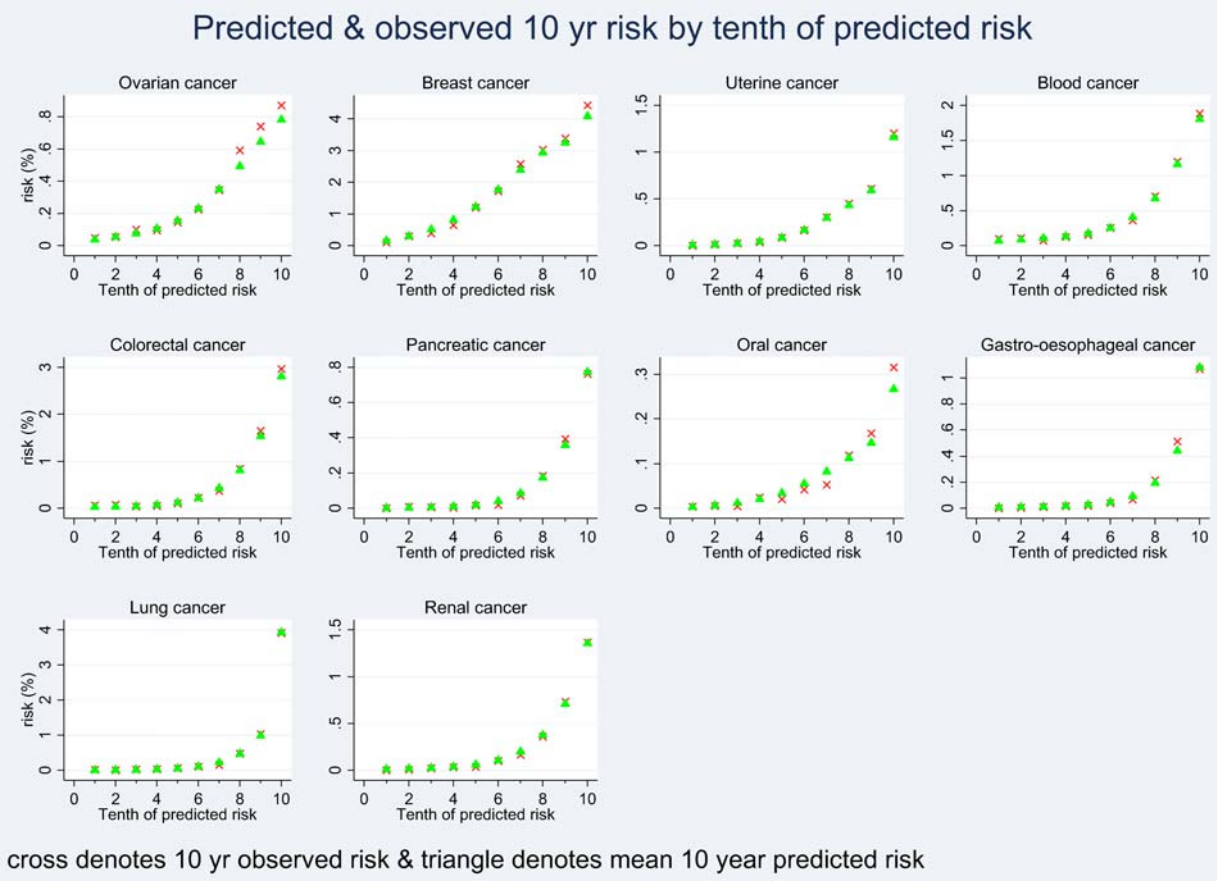

Figure 6 Showing the mean predicted risks and observed risks at 10 years by tenth of predicted risk applying each algorithm to all women in the validation cohort.

independent studies have examined other risk algorithms,[85 86 88 89] they have demonstrated comparable performance compared with the validation in the QResearch database.[19 80 84] We (or another academic team) intend to conduct a separate validation on an independent database using a different clinical system (CPRD) as part of a separate project.

Other limitations of our study include the lack of formal adjudication of cancer diagnoses, information bias, and potential for bias due to missing data. However, we think our ascertainment of cancer is likely to be high given the combination of the four data sources and the similarity between our rates and those published by the Office for National Statistics.[11] While we acknowledge that there may be discrepancies between self-reported features (such as number of cigarettes smoked), the information recorded on the GP record has a face validity as it is still predictive of relevant outcomes, exhibits a 'dose-response' relationships (with higher doses being associated with increased cancer risks) and similar to HRs from the published literature as we describe above. We have also only based behavioural variables (for smoking and alcohol) on a simple assessment of current exposure rather than more complex measures reliant on recall of past exposure which would be less reliable.

We measured family history of cancer based on information which may have been recorded opportunistically rather than systematically. Patients with a positive family history of cancer may be more likely to report this to their GP and for this to be recorded and we have assumed that where no positive family history is recorded then the patients has a negative family history which will lead to some misclassification. Also, there is no Read code to record family history of some types of cancer (eg, pancreatic cancer) so we were unable to include this in our list of predictor variables as this information is not captured in coded form and available on the QResearch database. Overall, the recording of family history is therefore likely to be subject to both ascertainment and recording bias. However, it is likely that the recording of family history will improve over time particularly if (1) clinicians can be prompted to ask for this information by the use of structured clinical templates which can be offered to the clinician for use during the consultation or if (2) patients can record this information directly once they are able to access their medical record electronically.

Genetic information is likely to be important in a full assessment of cancer risk although the information is not currently routinely recorded in electronic health records. Genetic information cannot therefore be used

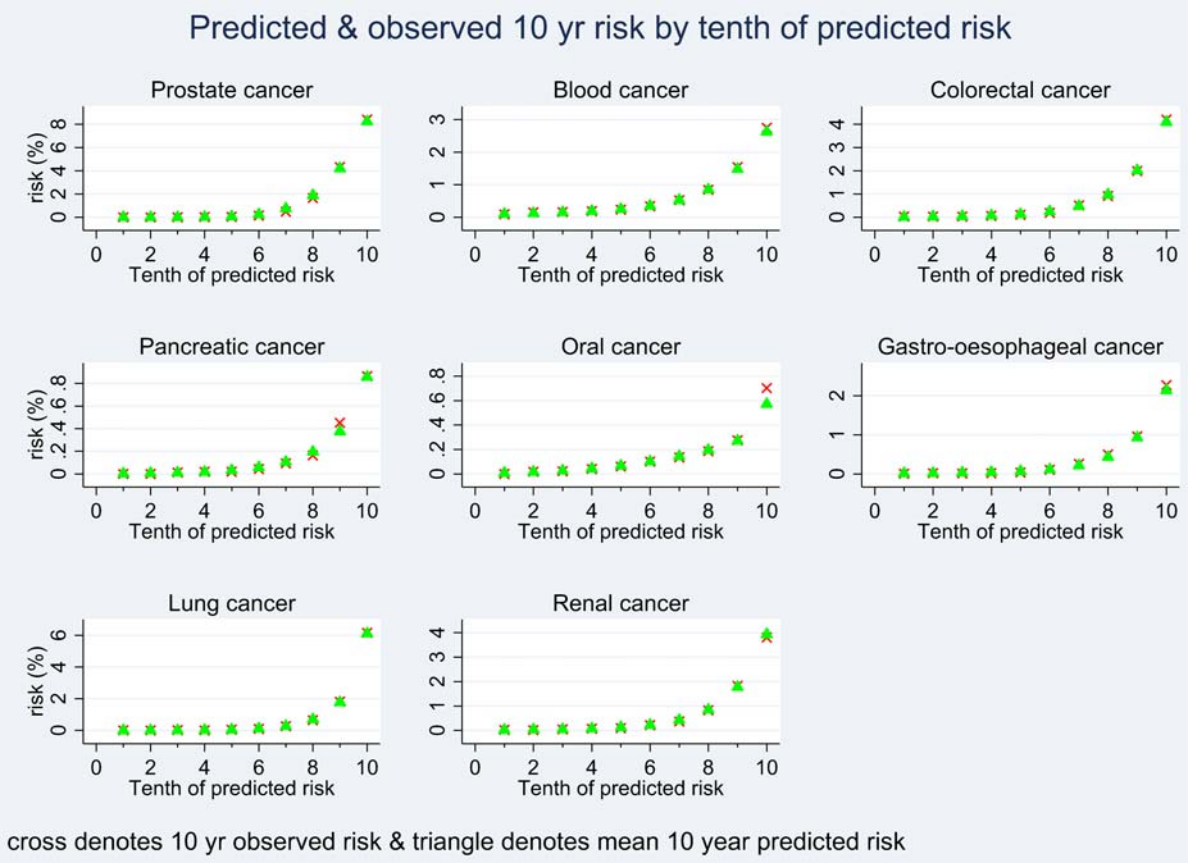

Figure 7 Showing the mean predicted risks and observed risks at 10 years by tenth of predicted risk applying each algorithm to all men in the validation cohort.

either to derive or validate a new score using large representative UK clinical databases. The lack of routinely available genetic results would also cause practical problems in implementing a tool dependent on such variables into clinical practice. Previous studies examining the benefits of adding genetic variations to risk assessment tools have shown only modest improvements in the predictive power. For example adding genetic variant information to the Gail Breast Cancer Risk assessment tool only increased the ROC value from 0.58 to 0.62.[90] The ROC value for the breast cancer algorithm based on clinical data in our study was 0.76 which is significantly higher than the ROC value of 0.62 reported for the generic variant of the Gail model.[90]

A recent study by Tomasetti *et al*[91] concluded that "only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions. The majority is due to 'bad luck', that is, random mutations arising during DNA replication in normal, noncancerous stem cells". Our study focuses on methods to predict absolute risk of cancer taking account of the individual's age, sex, ethnicity, lifestyle, family history and comorbidities. We found that our models explained well over 33% of the variation in cancer risk for 9 of the 11 cancers studied as shown by the R^2 values presented in table 8. This would tend to refute the hypothesis by Tomasetti *et al*[91] that only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions with the remainder being due to random mutations for the majority of cancers. However, our models only explained 22% of the variation for breast cancer and 29% for ovarian cancer which would support it.

The prediction models in our study have been derived from routinely recorded electronic health data using variables which are accessible in everyday practice in order that they can be applied in real world clinical settings to identify high-risk patients for further screening or prevention. Overall the validation statistics indicate that performance of all the algorithms is either good (breast, oral and ovarian cancers have ROC values 0.76–0.79); very good (blood, bowel, gastro-oesophageal, pancreas prostate, renal, uterine cancers have ROC values of 0.80 to 0.89) or excellent (lung cancer has an ROC value >0.9). Our prediction models have included clinical and lifestyle variables (such as age/sex/ethnicity/coexisting illnesses/smoking/alcohol and medication). They also include family history which was a significant predictor for six cancers (prostate, breast, bowel, blood, lung, ovarian). While there is clearly a

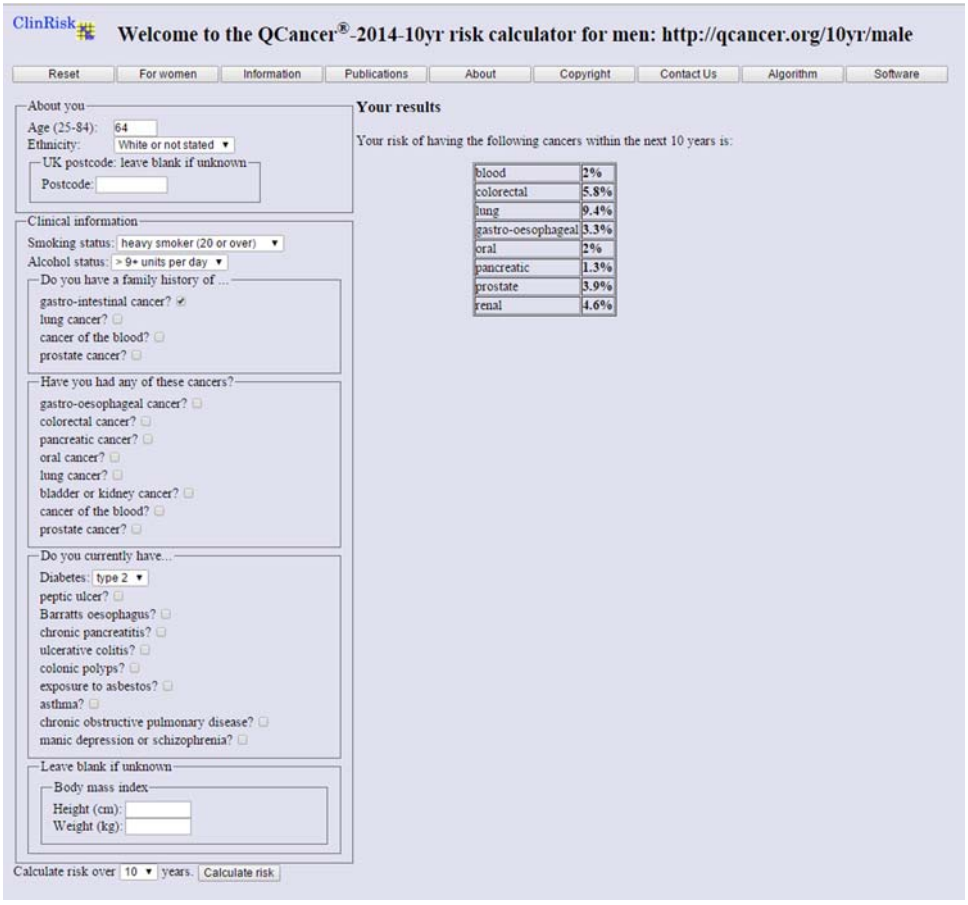

Figure 8 Showing the web calculator for an example patient.

difference between family history information and a genetic sample, family history does reflect genetic factors and is routinely recorded in general practice settings. It would be impractical, both from a cost and governance perspective to undertake genetic tests on all patients so that this information could be included in derivation of the models. It would also be impractical to require genetic information to be available routinely for when the scores are applied. It would be possible, however, to use the QCancer 10 year risk tool to identify a subset of high-risk patients for whom further genetic testing might be warranted and to quantify the level of absolute risk to help inform patient choice.

The QCancer 10 year risk algorithms have been developed using linked data from general practices in England and some of them (breast, gastro-oesophageal, lung, oral, pancreatic, prostate and renal tract) include a postcode related deprivation score (Townsend score). We included the Townsend score in the algorithms since there were clear relationships between deprivation and risk of some cancers which is captured by this variable. If we omitted it, it would tend to under-estimate cancer risk in patients from deprived areas for most cancers except breast cancer and renal cancer in women and prostate cancer in men, where the risk may be slightly over-estimated. In terms of international use of these algorithms where the Townsend score is not available a locally relevant deprivation score could be constructed or adapted for the relevant country to have a range between −7 (most affluent) and +11 (most deprived) to correspond to the range of Townsend scores which could be used instead (subject to local validation). For other predictor variables such as age, smoking, alcohol, family history, prior cancer, then we have compared the magnitude of the HRs to other international studies and found them to be similar in direction and magnitude (see Comparison with literature). This would tend to support the utility and face validity of the algorithms outside the UK although best practice would be that they are externally and independently validated in the settings where they would be used to ensure they are appropriately calibrated and have good discrimination.

Lastly changes in environmental factors may occur over time and this underlines the need to update the prediction algorithms on an ongoing basis as has been carried out with QRISK2[92] cardiovascular score and other related prediction scores.[93] Regular updates, with a moving time window, will also help ensure that the algorithms will benefit from improvements in the scope and quality of the underlying database which is likely to occur over time. This is an important strength of using routine databases for the development of risk prediction algorithms that is not feasible with prospective study cohorts that are assembled at one point in time.

Clinical implications

The algorithms have been designed to work in a primary care setting, making use of information which is already recorded on the GP clinical computer system. The algorithms can be integrated into the clinical computer system alongside similar algorithms which already quantify risk of other clinical conditions in everyday clinical practice such as QRISK2,[19] QDiabetes,[80] QStroke,[94] QFracture,[81] QThrombosis,[83] QBleed[95] and QAdmissions.[12] They can be used in 'batch process' mode to generate a list of patients at high risk of each cancer for further assessment or they could be used within the consultation. For example, policy makers and commissioners are likely to want to use the tumour site specific algorithms in 'batch mode' to risk stratify populations to better target screening programmes.

To indicate the potential value of using the QCancer tool for identifying a high-risk population for screening compared to an approach based on risk factors alone, we calculated the sensitivity and predictive values for lung cancer in heavy and moderate smokers. In the validation data set for women there were 3007 lung cancer cases over 10 years, and 812 occurred in moderate/heavy smokers (9.3% of cohort) giving a sensitivity of 27%, this contrasts with a sensitivity of 67% for the top 10% of women at highest predicted risk of lung cancer using the QCancer score (table 9). In men the corresponding sensitivity values are 22% in moderate/heavy smokers compared with 67% using QCancer (table 10). So QCancer is better at identifying future lung cancer cases than an approach targeting moderate/heavy smokers would be. Further work would be needed to determine any risk thresholds for screening, which as we state in the paper would require cost-effectiveness analyses which are outside the scope of this study and consideration of available resources and impacts on patients.

The algorithms could help inform the discussions between doctor and patient within the consultation regarding the future risk of cancer associated with existing diagnoses (such as the risk of bowel cancer among patients with ulcerative colitis); prior cancers (which might indicate an increased risk of a second primary cancer); family history of cancer (where additional screening may be justified) or lifestyle related variables (such as BMI, smoking and alcohol intake which can be moderated). For example, the doctor could use the algorithms to assess the patient's 10 year risk of cancer to highlight the higher risk associated with heavy or moderate smoking compared with that for lighter smokers or ex-smokers. Currently there is no easy to use widely available calculator, such as that described in this paper, which will allow a patient to quantify their absolute risk of getting different types of cancer taking account of their age, sex, family history and other risk factors such as alcohol consumption and smoking status. While identifying effective interventions to reduce alcohol intake or increase smoking cessation remains a challenge, there is evidence[96] that physician advice has some effect on smoking cessation rates albeit small and that brief alcohol interventions in primary care populations can reduce alcohol consumption.[97] It is possible that providing patients with information on how these factors influence their personal cancer risk might have an additional impact on smoking and alcohol consumption. There is also some evidence to support the use of biomedical risk assessment feedback such as 'lung age' to increase rates of smoking cessation[98] which suggests that patients may respond to information presented in an accessible format.

As another example, the doctor and patient could also review treatments such as the use of HRT in a woman at higher risk of breast cancer due to other factors such as family history. However, if a patient presents with potential symptoms of cancer such as appetite loss or haematuria the existing QCancer scores we developed previously[3–10] would be more suitable for assessing current cancer risk and informing decisions regarding further investigation and referral.

CONCLUSION

We have developed and validated a new set of risk prediction models which quantify the absolute risk of 11 common cancers in men and women. They can be used to identify patients at high risk of cancers for prevention or further assessment. Following external validation and cost-effectiveness assessments, the algorithms could be integrated into GP clinical computer systems and used to identify high-risk patients for prevention and screening.

Twitter Follow Julia Hippisley-Cox at @juliahcox

Acknowledgements The authors acknowledge the contribution of EMIS practices who contribute to the QResearch and EMIS for expertise in establishing, developing and supporting the database.

Contributors JH-C initiated the study, undertook the literature review, data extraction, data manipulation and primary data analysis and wrote the first draft of the paper. CC contributed to the design, analysis, interpretation and drafting of the paper. JH-C is the guarantor. Both authors agreed the final version of the paper.

Competing interests JHC is professor of clinical epidemiology at the University of Nottingham and codirector of QResearch—a not-for-profit organisation which is a joint partnership between the University of Nottingham and Egton Medical Information Systems (leading commercial supplier of IT for 60% of general practices in the UK). JHC is also a paid director of ClinRisk Ltd which produces open and closed source software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help improve patient care. CC 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 coauthors and not of any affiliated bodies or organisations.

Ethics approval The project was reviewed in accordance with the QResearch agreement with Trent Multi-Centre Ethics Committee [reference 03/4/021].

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The algorithms presented in this paper will be released as Open Source Software under the GNU lesser GPL v3. The open

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;8:773–83.
2. Department of Health. *The Cancer Reform Strategy*. In: Health Do, ed. London: Department of Health, 2007.
3. 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;61:e707–14.
4. 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;61:e715–23.
5. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. *BMJ* 2012;344.
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;62:e29–37.
7. 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;62:e38–45.
8. 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;62:e251–60.
9. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2013;63:11–21.
10. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2013;63:1–10.
11. Office for National Statistics. Ten most common cancers among males and females. 2014. http://www.ons.gov.uk/ons/rel/vsob1/cancer-statistics-registrations–england–series-mb1-/no–43–2012/info-most-common-cancers.html
12. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013;3:e003482.
13. Hippisley-Cox J. Validity and completeness of the NHS Number in primary and secondary care electronic data in England 1991–2013. 2013;1. Hippisley-Cox J. Validity and completeness of the NHS number in primary and secondary care: electronic data in England 1991–2013. http://eprints.nottingham.ac.uk/3153/ (accessed Jun 2013).
14. Hippisley-Cox J, Vinogradova Y, Coupland C, *et al*. Risk of malignancy in patients with schizophrenia or bipolar disorder: nested case-control study. *Arch Gen Psychiatry* 2007;64:1368–76.
15. Hippisley-Cox J, Coupland C. Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. *BMJ* 2010;340:c2197.
16. Hippisley-Cox J, Coupland C. Individualising the risks of statins in men and women in England and Wales: population-based cohort study. *Heart* 2010;96:939–47.
17. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to cyclooxygenase-2 inhibitors and risk of cancer: nested case-control studies. *Br J Cancer* 2011;105:452–9.
18. Cancer Research UK. Cancer Research Website. 2014. http://www.cancerresearchuk.org/ (accessed 30th Jul 2014).
19. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
20. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.
21. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34–9.
22. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584.
23. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002;7:147–77.
24. Group TAM. Academic Medicine: problems and solutions. *BMJ* 1989;298:573–9.
25. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007;60:979.
26. Moons KGM, Donders RART, Stijnen T, *et al*. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006;59:1092.
27. Rubin DB. *Multiple imputation for non-response in surveys*. New York: John Wiley, 1987.
28. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964–74.
29. Hosmer D, Lemeshow S. *Applied logistic regression*. New York: John Wiley & Sons, Inc., 1989.
30. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* 2014;4:e005809.
31. Royston P. Explained variation for survival models. *Stata J* 2006;6:1–14.
32. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723–48.
33. Tsoi KK, Pau CY, Wu WK, *et al*. Cigarette smoking and the risk of colorectal cancer: a meta-analysis of prospective cohort studies. *Clin Gastroenterol Hepatol* 2009;7:682–8.e1–5.
34. Freedman ND, Abnet CC, Leitzmann MF, *et al*. A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. *Am J Epidemiol* 2007;165:1424–33.
35. La Torre G, Chiaradia G, Gianfagna F, *et al*. Smoking status and gastric cancer risk: an updated meta-analysis of case-control studies published in the past ten years. *Tumori* 2009;95:13–22.
36. Doll R, Peto R, Wheatley K, *et al*. Mortality in relation to smoking:40 years' observations on male British doctors. *BMJ* 1994;309:901–11.
37. Parkin DM, Boyd L, Walker LC. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br J Cancer* 2011;105(Suppl 2):S77–81.
38. Bosetti C, Lucenteforte E, Silverman DT, *et al*. Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). *Ann Oncol* 2012;23:1880–8.
39. Zou L, Zhong R, Shen N, *et al*. Non-linear dose-response relationship between cigarette smoking and pancreatic cancer risk: evidence from a meta-analysis of 42 observational studies. *Eur J Cancer* 2014;50:193–203.
40. Viswanathan AN, Feskanich D, De Vivo I, *et al*. Smoking and the risk of endometrial cancer: results from the Nurses' Health Study. *Int J Cancer* 2005;114:996–1001.
41. Allen NE, Beral V, Casabonne D, *et al*. Moderate alcohol intake and cancer incidence in women. *J Natl Cancer Inst* 2009;101:296–305.
42. Hamajima N, Hirose K, Tajima K, *et al*. Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br J Cancer* 2002;87:1234–45.
43. Weikert C, Dietrich T, Boeing H, *et al*. Lifetime and baseline alcohol intake and risk of cancer of the upper aero-digestive tract in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Int J Cancer* 2009;125:406–12.
44. Johns LE, Houlston RS. A systematic review and meta-analysis of familial prostate cancer risk. *BJU Int* 2003;91:789–94.
45. Pharoah PD, Day NE, Duffy S, *et al*. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int J Cancer* 1997;71:800–9.
46. Collaborative Group on Hormonal Factors in Breast C. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 2001;358:1389–99.
47. Fearnhead NS, Wilding JL, Bodmer WF. Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br Med Bull* 2002;64:27–43.
48. Butterworth AS, Higgins JP, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. *Eur J Cancer* 2006;42:216–27.
49. Cote ML, Liu M, Bonassi S, *et al*. Increased risk of lung cancer in individuals with a family history of the disease: a pooled analysis

from the International Lung Cancer Consortium. *Eur J Cancer* 2012;48:1957–68.
50. Gayther SA, Pharoah PD. The inherited genetics of ovarian and endometrial cancer. *Curr Opin Genet Dev* 2010;20:231–8.
51. Granstrom C, Sundquist J, Hemminki K. Population attributable fractions for ovarian cancer in Swedish women by morphological type. *Br J Cancer* 2008;98:199–205.
52. Cancer CGoHFiB. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet* 1996;347:1713–27.
53. Million Women Study Collaborators. Breast cancer and hormone replacement therapy in the Million Women Study. *Lancet* 2003;362:419–27.
54. Chlebowski RT, Manson JE, Anderson GL, *et al*. Estrogen plus progestin and breast cancer incidence and mortality in the Women's Health Initiative Observational Study. *J Natl Cancer Inst* 2013;105:526–35.
55. Havrilesky LJ, Gierisch JM, Moorman PG, *et al*. Oral contraceptive use for the primary prevention of ovarian cancer. *Evid Rep Technol Assess (Full Rep)* 2013(212):1–514.
56. Salehi F, Dunfield L, Phillips KP, *et al*. Risk factors for ovarian cancer: an overview with emphasis on hormonal factors. *J Toxicol Environ Health B Crit Rev* 2008;11:301–21.
57. Lutgens MW, van Oijen MG, van der Heijden GJ, *et al*. Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies. *Inflamm Bowel Dis* 2013;19:789–99.
58. Castano-Milla C, Chaparro M, Gisbert JP. Systematic review with meta-analysis: the declining risk of colorectal cancer in ulcerative colitis. *Aliment Pharmacol Ther* 2014;39:645–59.
59. Hvid-Jensen F, Pedersen L, Drewes AM, *et al*. Incidence of adenocarcinoma among patients with Barrett's esophagus. *N Engl J Med* 2011;365:1375–83.
60. Duell EJ, Lucenteforte E, Olson SH, *et al*. Pancreatitis and pancreatic cancer risk: a pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4). *Ann Oncol* 2012;23:2964–70.
61. Zhou WB, Xue DQ, Liu XA, *et al*. The influence of family history and histological stratification on breast cancer risk in women with benign breast disease: a meta-analysis. *J Cancer Res Clin Oncol* 2011;137:1053–60.
62. Torrey EF. Prostate cancer and schizophrenia. *Urology* 2006;68:1280–3.
63. Hardiman P, Pillay OC, Atiomo W. Polycystic ovary syndrome and endometrial carcinoma. *Lancet* 2003;361:1810–12.
64. Jiang Y, Ben Q, Shen H, *et al*. Diabetes mellitus and incidence and mortality of colorectal cancer: a systematic review and meta-analysis of cohort studies. *Eur J Epidemiol* 2011;26:863–76.
65. Kramer HU, Schottker B, Raum E, *et al*. Type 2 diabetes mellitus and colorectal cancer: meta-analysis on sex-specific differences. *Eur J Cancer* 2012;48:1269–82.
66. Luo W, Cao Y, Liao C, *et al*. Diabetes mellitus and the incidence and mortality of colorectal cancer: a meta-analysis of twenty four cohort studies. *Colorectal Dis* 2011; doi:10.1111/j.1463-1318.2011.02875.x.
67. Wu L, Yu C, Jiang H, *et al*. Diabetes mellitus and the occurrence of colorectal cancer: an updated meta-analysis of cohort studies. *Diabetes Technol Ther* 2013;15:419–27.
68. Shimoyama S. Diabetes mellitus carries a risk of gastric cancer: a meta-analysis. *WJG* 2013;19:6902–10.
69. Starup-Linde J, Karlstad O, Eriksen SA, *et al*. CARING (CAncer Risk and INsulin analoGues): the association of diabetes mellitus and cancer risk with focus on possible determinants—a systematic review and a meta-analysis. *Curr Drug Saf* 2013;8:296–332.
70. Ben Q, Xu M, Ning X, *et al*. Diabetes mellitus and risk of pancreatic cancer: a meta-analysis of cohort studies. *Eur J Cancer* 2011;47:1928–37.
71. Zhang ZH, Su PY, Hao JH, *et al*. The role of preexisting diabetes mellitus on incidence and mortality of endometrial cancer: a meta-analysis of prospective cohort studies. *Int J Gynecol Cancer* 2013;23:294–303.
72. Zhang F, Yang Y, Skrip L, *et al*. Diabetes mellitus and risk of prostate cancer: an updated meta-analysis based on 12 case-control and 25 cohort studies. *Acta Diabetol* 2012;49(Suppl 1):S235–46.
73. Xu H, Mao SH, Ding GX, *et al*. Diabetes mellitus reduces prostate cancer risk—no function of age at diagnosis or duration of disease. *Asian Pac J Cancer Prev* 2013;14:441–7.
74. UK CR. Cancer incidence and survival by major ethnic group in England, 2002–2006. Secondary Cancer incidence and survival by major ethnic group in England, 2002–2006. 2009. http://publications.cancerresearchuk.org/cancerstats/statssurvival/incidence-survival-ethnicity.html.html
75. Ben-Shlomo Y, Evans S, Ibrahim F, *et al*. The risk of prostate cancer amongst black men in the United Kingdom: the PROCESS cohort study. *Eur Urol* 2008;53:99–105.
76. Youlden DR, Baade PD. The relative risk of second primary cancers in Queensland, Australia: a retrospective cohort study. *BMC Cancer* 2011;11:83.
77. Jegu J, Colonna M, Daubisse-Marliac L, *et al*. The effect of patient characteristics on second primary cancer risk in France. *BMC Cancer* 2014;14:94.
78. Dores GM, Metayer C, Curtis RE, *et al*. Second malignant neoplasms among long-term survivors of Hodgkin's disease: a population-based evaluation over 25 years. *J Clin Oncol* 2002;20:3484–94.
79. National Clinical Guideline Centre. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. *London* 2014:286.
80. 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.
81. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012;344:e3427.
82. 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;11:49.
83. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. *BMJ* 2011;343:d4656.
84. 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.
85. Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 2011;342:d3651.
86. Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing type 2 diabetes. *Diabet Med* 2011;28:599–607.
87. Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. *Health Stat Q* 2004;21:5–14.
88. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012;344:e4181.
89. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;340:c2442.
90. Wacholder S, Hartge P, Prentice R, *et al*. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* 2010;362:986–93.
91. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 2015;347:78–81.
92. Hippisley-Cox J, Coupland C. QRISK2–2014 Annual Update Information, 2014:5.
93. Hippisley-Cox J. QDiabetes 2013 Annual Update Information Secondary QDiabetes 2013 Annual Update Information 2013. http://qdiabetes.org/QDiabetes-2013-Annual-Update-Information.pdf
94. Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ* 2013;346:f2573.
95. Hippisley-Cox J, Coupland C. Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores. *BMJ* 2014;349:g4606.
96. Stead LF, Buitrago D, Preciado N, *et al*. Physician advice for smoking cessation. *Cochrane Database Syst Rev* 2013;5:CD000165.
97. Kaner EF, Beyer F, Dickinson HO, *et al*. Effectiveness of brief alcohol interventions in primary care populations. *Cochrane Database Syst Rev* 2007(2):CD004148.
98. Parkes G, Greenhalgh T, Griffin M, *et al*. Effect on smoking quit rate of telling patients their lung age: the Step2quit randomised controlled trial. *BMJ* 2008;336:598–600.

BMJ Open

Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study

Julia Hippisley-Cox and Carol Coupland

BMJ Open 2015 5:
doi: 10.1136/bmjopen-2015-007825