Protocol for the development and validation of risk prediction equations to estimate absolute and conditional survival in patients with cancer

Julia Hippisley-Cox, Professor. Carol Coupland, Professor.

***Abstract*—This is a protocol to describe the development and validation of a set of prediction equations to quantify absolute survival for patients with different types of cancers taking account of other clinical factors available through routine linkage of cancer registry data to primary care electronic health records. We will also include estimates of conditional survival since it may be a more accurate measure of survival among those surviving the first year, especially when the initial prognosis is poor, such as with advanced stage colorectal cancer. Such estimates can be used to provide better information for patients and doctors to help inform treatment and other life decisions.**

***Index Terms*—cancer survival, primary care, predictive modelling, cancer registration, primary care databases**

1 Introduction

Traditional cancer survival estimates provide important information for guidelines, planning treatment, follow-up and ongoing surveillance of different types of cancer. *Relative survival* estimates are used to "cancel out" changes in competing causes of death so that changes in prevention and treatment strategies can be compared over time and between populations[1]. Relative estimates are usually based on analyses of cancer registries alone and presented as a series of tables taking account of one or two factors– such as the patient's age and sex or the stage of their cancer at diagnosis[2]. Whilst *relative survival* estimates (also known as *net survival* estimates) are useful for researchers and policy makers[3], they are less relevant for patients and doctors who tend to be interested in *absolute survival* for individuals which is able to simultaneously account for multiple patient characteristics. Consider an example of a 38 year old woman with stage 4, well differentiated, colorectal cancer for which she has had a hemi-colectomy. Based on published statistics[2], the 5 year *relative survival* for those aged 15-39 years is 66% whilst that for stage 4 disease is only 8%. However, she wishes to know her individualised 5 year *absolute survival* since she is taking aspirin[4], her tumour is well differentiated and has been resected, and given that she has survived the first year after diagnosis when many people with stage 4 cancer may have died[5]. She finds a US website which calculates 5 year survival for colorectal conditional on surviving the first year based on data from 1998-2000[5]. Whilst this takes account of 10 year age band, sex, ethnicity, stage and grade of cancer, it doesn't include other factors thought to affect survival such as cancer treatments, comorbidities or use of statins[6] or aspirin[4].

Therefore, we will derive and externally validate a set of prediction equations to quantify *absolute survival* for patients with different types of cancers taking account of other clinical factors available through routine linkage of cancer registry data to primary care electronic health records. We will also include estimates of *conditional survival* since it may be a more accurate measure of survival among those surviving the first year, especially when the initial prognosis is poor, such as with advanced stage colorectal cancer[5]. Such estimates can be used to provide better information for patients and doctors to help inform treatment and other life decisions[1].

2 Methods

2.1 *Study design and data source*

We will undertake an open cohort study to derive and validate the risk equations in a large population of primary care patients with cancer using the UK QResearch® database (version 41, www.qresearch.org). We will also carry out a second external validation using a cohort of patients not registered with QResearch practices but included on the national cancer registry. QResearch® is a continually updated patient level pseudonymised database with data extending back to 1989. It includes clinical and demographic data from over 1,200 general practices covering a population of > 22 million patients, collected in the course of routine healthcare. The primary care data includes demographic information, diagnoses, prescriptions, referrals, laboratory results and clinical values. Diagnoses are recorded using the Read code classification[7]. QResearch® has been used for a wide range of clinical research including the development and validation of risk prediction models[8-10] including those predicting risk of existing but as yet undiagnosed cancer[11-18] and those predicting risk of future cancers[19].

The primary care data on QResearch is linked at individual patient level to national cancer registry data supplied by Public

Submission Date 11 June 2016. Julia Hippisley-Cox is Professor of Clinical Epidemiology & General Practice at the University of Nottingham and Medical Director of ClinRisk Ltd (email: julia.hippisley-cox@nottingham.ac.uk). Carol Coupland is Professor of Medical Statistics in Primary Care at the University of Nottingham and consultant statistician for ClinRisk Ltd (email: carol.coupland@nottingham.ac.uk).

Health England (PHE cancer), hospital episode statistics (HES) supplied by the Health and Social Care Information Centre and the mortality register supplied by the Office for National Statistics (ONS Mortality). The cancer registry includes all cancers registered in England between 1990 and 2013 with follow-up for mortality until 31 Dec 2014. It includes information on the tumour diagnosis date, treatment, location, behavior, morphology, grade, stage, basis for diagnosis and whether the cancer diagnosis was only present on a death certificate.

ONS mortality data includes dates and causes of deaths in England coded using the international classification of diseases version 10 (ICD-10) classification.

HES includes inpatient admissions and outpatient appointments since 1998 including primary and secondary diagnoses coded using the ICD-10 classification and operative procedures coded using the Classification of Interventions and Procedures version 4 (OPCS-4)[20].

Patient records from each of the four data sources are linked using a project specific pseudonymised NHS number which is valid and complete for 99.8% of primary care patients, 99.9% for mortality and cancer registration records and 98% for hospital admissions records[1].

2.2 *Cohort selection*

We will include all QResearch® practices in England once they have been using their Egton Medical Information Systems (EMIS) computer system for at least a year during the study period. We will randomly allocate three quarters of these practices to the derivation dataset and the remaining quarter to a validation dataset. In both datasets, we will identify open cohorts of patients registered with eligible practices at any time between 01 January 1998 and 31st December 2013.

2.2.1 *Inclusion criteria*

We will include patients in the derivation and validation cohorts registered with QResearch® with a first recorded diagnosis of each type of cancer on the linked cancer registration data between 01.01.1998 and 31.12.2013. As in other studies[21-23], we will use ICD-10 codes to identify cases of cancer. We will restrict the analysis to patients aged 15-99 years who had a first diagnosis during the period of registration with the practice, ensuring that each patient had at least 12 months of registration with the practice prior to cancer diagnosis. We will use the date of cancer diagnosis from the cancer registry data as the index date for entry to the cohort and patients will be followed up until the earliest of the date of death or 31st December 2014 ensuring that all patients had the opportunity for at least 12 months of follow-up after diagnosis.

2.2.2 *Exclusion criteria*

Cases with previous cancer of a different type from the index cancer will not be excluded unless they died before the study period. As in other studies[21 22], we will exclude patients where the growth behaviour for the index cancer diagnosis was coded as benign and those where the diagnosis was made on the date of death or only recorded on the patients' death certificate since the date of diagnosis and duration of survival is unknown.

For the separate external validation cohort, we will identify patients aged 15-99 years with cancer recorded on cancer registry data who were not registered with QResearch® practices at the time of diagnosis, excluding those where the growth behaviour was coded as benign and those with a death certificate only diagnosis.

2.3 *Cancer types*

We will undertake separate analyses for the following types of cancer which represent the most commonly occurring cancers.

- Lung cancer
- Colorectal cancer
- Gastro-oesophageal cancer
- Pancreatic cancer
- Renal tract (kidney and bladder)
- Ovarian cancer
- Endometrial cancer
- Cervical cancer
- Breast cancer
- Prostate cancer
- Melanoma
- Brain cancer
- Haematological cancers (lymphoma, myeloma, leukaemia)
- Mesothelioma
- Sarcoma

2.4 *Outcomes*

Our main outcome of interest is all-cause mortality.

2.5 *Predictor variables*

We will examine the following predictor variables based on established factors thought to affect mortality following a diagnosis of cancer based on a review of the current literature. We will also include variables known to affect all-cause mortality so that the absolute risk estimates will be able to reflect these factors.

2.5.1 *Demographic variables*

- Age at diagnosis of cancer (continuous)[22]
- Sex (male or female) [22]
- Deprivation (Townsend deprivation score which is a continuous score ranging from -11 to +8 where higher values indicate higher levels of deprivation) [22]
- Ethnicity: 10 groups - white/not recorded; Indian; Pakistani; Bangladeshi; Other Asian; Black; Caribbean; Black African; Chinese; Other. Ethnicity was assigned using values recorded on GP records or PHE cancer registry records.
-

2.5.2 *Cancer specific variables from PHE cancer registry*

- Cancer location – classified using ICD-10
- Cancer stage[22 23] - classified using TNM version 7 classification[24] with 4 groups: local involvement only, extension to adjacent tissue, lymph node involvement, metastasis.
- Cancer grade – with 4 groups: well differentiated; moderately differentiated; poorly differentiated and undifferentiated.
- Calendar year of diagnosis – this is to account for advances in diagnoses and treatments over time
- Hormone status (oestrogen & progesterone), nodes excised, excision margin where recorded and relevant (e.g. breast cancer).

2.5.3 *Treatment variables from PHE cancer records linked to HES records*

Treatment variables will be identified according to procedural and treatment codes recorded on the patients' cancer registry record linked to hospital episode statistics information. Procedures will be identified as associated with the diagnosis of cancer if they occurred within 12 months of the date of diagnosis. The precise dates of treatment are unavailable.

- Surgical treatment (yes/no)
- Chemotherapy[25] (yes/no)
- Radiotherapy (yes/no)
- Hormone therapy[3] (yes/no)

2.5.4 *Predictors from linked GP records*

The following variables including recorded dates will be identified from the patients linked GP record if they were recorded prior to the diagnosis of cancer.

- Family history of cancer (yes/no)
- Cardiovascular disease (angina, myocardial infarction, stroke or TIA) (yes/no)
- Diabetes – three groups Type1; Type2; no diabetes
- Chronic renal disease (yes/no)
- Chronic obstructive airways disease (yes/no)
- Inflammatory bowel disease (Crohn's disease or ulcerative colitis) (yes/no)
- Prior cancer (yes/no)
- Venous thromboembolism (deep vein thrombosis or pulmonary embolus) (yes/no)
- Use of aspirin[4] (yes/no)
- Use of statins[6] (yes/no)
- Use of HRT in women (yes/no) (relevant for breast and gynae cancers)

We will identify clinical values from the patients' linked GP record. For body mass index, smoking and alcohol, we will use the most recent value prior to diagnosis. For blood tests, we will use the values closest to the diagnosis of cancer, selecting from those recorded within 12 months either side of the diagnosis date.

- Smoking status (non-smoker; ex-smoker; light (1-9 cigarettes/day); moderate (10-19/day); heavy (20+/day)
- Alcohol status (non-drinker, <1 unit/day; 1-2 units/day; 3+ units/day)
- Body mass index kg/m^2 (continuous).
- Anaemia[25] defined as haemoglobin level < 11g/dl (yes/no)
- Raised platelets defined as values above 480 * 10^9/L (yes/no)
- Abnormal LFTs defined as either GGT, ALT or bilirubin more than 3 times normal (yes/no)

2.6 *Descriptive statistics, survival rates and age standardisation*

We will calculate the observed survival and the relative net survival for patients from the date of diagnosis by age, sex, and calendar year to enable comparisons with other studies using similar datasets[21-23]. Relative survival is the ratio of the overall survival for a cohort of cancer patients to the expected survival in the general population matched by age, sex and calendar year[21]. We will use the *strs* program in Stata (version 14) with breakpoints set at 1, 5, and 10 years and the Ederer II method[26]. The lifetables will be obtained from the Office of National Statistics website[27] including background mortality until the end of 2014. In order to aid comparison with other studies[22], age-standardised relative survival will be calculated using a method and standard weights proposed by Corazziari et al 2004[28]. We will also calculate conditional relative survival for patients conditional on the patient surviving from 1 after diagnosis[21 29 30].

For cancers where stage is commonly recorded, we will undertake analyses to describe which factors are associated with late stage of cancer (i.e. stage 4) at diagnoses adjusting for confounders where appropriate. We will also describe trends in stage of diagnosis over time.

2.7 *Derivation of the predictive models*

We will derive risk prediction equations in the derivation cohort to predict absolute survival using established methods[8 10]. We will derive separate equations for men and women. Initially we will use complete case analyses to derive fractional polynomial terms[31] to model non-linear risk relationships with continuous variables if appropriate (age and body mass index). We will use the *mi impute* chained command in STATA to perform multiple imputation chained equations to replace missing values for continuous values (body mass index and Townsend score) and categorical variables (smoking status, alcohol status, cancer stage and cancer grade) and use these values in our main analyses[32-34]. We will carry out 5 imputations. We will include all potential predictor variables in the imputation model along with the outcome (mortality) and cumulative hazard. We will then use Cox's proportional hazards models, with robust standard errors to account for clustering of patients within general practices, to estimate the coefficients for each predictor variable for death using the fractional polynomial terms obtained from the complete case analyses. We will use Rubin's rules to combine the regression coefficients across the imputed datasets[35]. We will fit full models initially then retain variables if they had a hazard ratio of < 0.80 or > 1.20 (for binary variables) and are statistically significant at the 0.01 level. We will examine interactions between predictor variables and age and include these where they are significant, plausible and improve model fit. Model fit will be assessed by measuring the AIC and BIC values for each imputed set of data.

We will use the regression coefficients for each variable from the final models as weights which we will combine with the baseline survivor functions evaluated up to 10 years to derive risk equations over a period of 10 years of follow-up[36]. We will estimate the baseline survivor function based on zero values of centred continuous variables, with all binary predictor values set to zero. This will enable us to derive absolute risk estimates for each year of follow-up, with a specific focus on 1, 5 and 10 year risk estimates. We will calculate conditional survival estimates by dividing the absolute survival at each time point by the absolute survival estimates at one year and five years as described by Dickman[37]. In order to assess potential survival bias of cancer treatments (which could otherwise make them appear more effective), we will undertake an additional analysis of conditional survival at one year using the delayed entry approach so that the hazard ratios can be compared.

2.8 *Validation of the predictive models*

We will use multiple imputation in the both the QResearch® and PHE validation cohorts to replace missing values for stage and grade. We will use multiple imputation in the QResearch® to replace missing values for continuous variables, smoking status, alcohol status. We will carry out 5 imputations. We will apply the risk equations to the both the QResearch and the PHE validation cohorts. We will calculate measures of discrimination. We will calculate R^2 values (explained variation in time to diagnosis of outcome[38]), D statistics[39] (a measure of discrimination where higher values indicate better discrimination) and receiver operating characteristic statistic over 1, 5 and 10 years and combine these model performance measures across imputed datasets using Rubin's rules. We will assess calibration, comparing the mean predicted risks at 1, 5 and 10 years with the observed risk by tenth of predicted risk. The observed risks will be obtained using Kaplan-Meier estimates evaluated at 1, 5 and 10 years.

We will include all the eligible patients in each database to maximise power and generalisability. We will use STATA (version 14.1) for all analyses. We will adhere to the TRIPOD statement for reporting[40].

2.9 *Methodological considerations*

The statistical methods we will use to derive and validate these models are very similar to those for other risk prediction tools derived from the QResearch® database, the strengths and limitations of which have been discussed in detail[8 10]. In summary, key strengths include cohort 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[41]. The QResearch® database has linked cancer, hospital and mortality records for nearly all patients and is therefore likely to have picked up the majority of cases of cancer thereby minimising ascertainment bias. We will undertake two validations, one using a separate set of practices and patients contributing to QResearch® and the other using patients not registered with QResearch practices but included on the PHE cancer registry. Whilst we will have derived and validated the equations using UK datasets, the equations could be used internationally by using alternative deprivation scores relevant to the setting (which would need to be scaled to conform with the Townsend score). Local validation should be done to ensure good calibration and discrimination in the applicable population.

Limitations of our study will include the lack of formal adjudication of diagnoses, and potential for bias due to missing data which we have addressed using multiple imputation. Dates of cancer treatments (surgery, radiotherapy and chemotherapy) are not available on the PHE cancer registry dataset, other than that treatment were done within 12 months of the date of cancer diagnosis. This could lead to a survival bias, making such treatments appear to be more effective. We will undertake additional analyses to assess the potential extent of these biases. We will not provide definite comment on what threshold of absolute risk should be used to define a "high risk" group as that would require (a) consideration of the balance of risks and benefits for individuals and (b) cost-effectiveness analyses which are outside the scope of this study.

3 OTHER INFORMATION

3.1.1 *Acknowledgements*

We acknowledge the contribution of EMIS practices who contribute to the QResearch® and EMIS for expertise in establishing, developing and supporting the database. We also acknowledge the contribution of the Health and Social Care Information for supplying the hospital episodes data and the Office of National Statistics for supplying the mortality and Public Health England for supplying the cancer registration data.

3.1.2 *Approvals:*

The project is being reviewed in accordance with the QResearch® agreement with NRES Committee East Midlands - Derby [reference 03/4/021].

3.1.3 *Competing Interests*

JHC is professor of clinical epidemiology at the University of Nottingham and co-director of QResearch® – a not-for-profit organisation which is a joint partnership between the University of Nottingham and 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 equations within clinical computer systems to help improve patient care. CC is Professor of Medical Statistics at the University of Nottingham and a paid consultant statistician for ClinRisk Ltd. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations.

3.1.4 *Data Sharing*

The equations presented in this paper will be released as Open Source Software under the GNU lesser GPL v3. The open source software allows use without charge under the terms of the GNU lesser public license version 3. Closed source software can be licensed at a fee.

4 REFERENCES

1. Howlader N, Mariotto AB, Woloshin S, et al. Providing clinicians and patients with actual prognosis: cancer in the context of competing causes of death. J Natl Cancer Inst Monogr 2014;**2014**(49):255-64.
2. UK CR. Bowel Cancer Survival Statistics London2015 [Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/survival.
3. Eloranta S, Adolfsson J, Lambert PC, et al. How can we make cancer survival statistics more useful for patients and clinicians: an illustration using localized prostate cancer in Sweden. Cancer Causes Control 2013;**24**(3):505-15.
4. Bastiaannet E, Sampieri K, Dekkers OM, et al. Use of aspirin postdiagnosis improves survival for colon cancer patients. Br J Cancer 2012;**106**(9):1564-70.
5. Chang GJ, Hu CY, Eng C, et al. Practical application of a calculator for conditional survival in colon cancer. J Clin Oncol 2009;**27**(35):5938-43.
6. Cai H, Zhang G, Wang Z, et al. Relationship Between the Use of Statins and Patient Survival in Colorectal Cancer: A Systematic Review and Meta-Analysis. PLoS ONE 2015;**10**(6):e0126944.
7. Wikipedia. Readcodes 2015 [Available from: http://en.wikipedia.org/wiki/Read_code.
8. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ 2008:bmj.39609.449676.25.
9. 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-.
10. 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.
11. Hippisley-Cox J, Coupland C. Identifying patients with suspected gastro-oesophageal cancer in primary care: derivation and validation of an algorithm. British Journal of General Practice 2011;**61**(592):e707-14.
12. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. British Journal of General Practice 2011;**61**(592):e715-23.
13. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: derivation and validation of algorithm. BMJ 2012;**344**.
14. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. British Journal of General Practice 2012;**62**(594):e29-e37.
15. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. British Journal of General Practice 2012;**62**(594):e38-e45.
16. 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**(597):e251-60.
17. 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**(606):11-21.
18. 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**(606):1-10.
19. 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**(3).
20. Health and Social Care Information Centre. OPCS-4 Classification [Available from: http://systems.hscic.gov.uk/data/clinicalcoding/codingstandards/opcs4/.
21. Coleman MP, Forman D, Bryant H, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. Lancet 2011;**377**(9760):127-38.
22. McPhail S, Johnson S, Greenberg D, et al. Stage at diagnosis and early mortality from cancer in England. Br J Cancer 2015;**112**(s1):S108-S15.
23. Maringe C, Walters S, Rachet B, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-2007. Acta Oncol 2013;**52**:919-32.
24. Sobin LH, Gospodarowicz MK, Wittekind CH. *TNM classification of malignant tumours*. Oxford: Wiley-Blackwell, 2009.
25. ZACHARAKIS M, XYNOS ID, LAZARIS A, et al. Predictors of Survival in Stage IV Metastatic Colorectal Cancer. Anticancer Research 2010;**30**(2):653-60.
26. Dickman PW, Sloggett A, Hills M, et al. Regression models for relative survival. Stat Med 2004;**23**(1):51-64.
27. Statistics OoN. National Life Tables London2014 [Available from: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-365199.
28. Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. Eur J Cancer 2004;**40**:2307-16.
29. Merrill RM, Henson DE, Ries LA. Conditional survival estimates in 34,963 patients with invasive carcinoma of the colon. Dis Colon Rectum 1998;**41**(9):1097-106.
30. Yu XQ, Baade PD, O'Connell DL. Conditional survival of cancer patients: an Australian perspective. BMC Cancer 2012;**12**:460.
31. 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.
32. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. J Epidemiol Community Health 2007;**60**:979.
33. 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.
34. Schafer J, Graham J. Missing data: our view of the state of the art. Psychological Methods 2002;**7**:147 - 77.
35. Rubin DB. *Multiple Imputation for Non-response in Surveys*. New York: John Wiley, 1987.
36. Hosmer D, Lemeshow S, May S. *Applied Survival Analysis: Regression Modelling of Time to Event data*. US: Wiley, 2007.
37. Dickman PW. Conditional relative survival modelling 2015 [Available from: http://www.statalist.org/forums/forum/general-stata-discussion/general/1311691-conditional-relative-survival-modelling.
38. Royston P. Explained variation for survival models. Stata J 2006;**6**:1-14.
39. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;**23**:723-48.
40. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. Annals of Internal Medicine 2015;**162**(1):55-63.
41. Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. Health Statistics Quarterly 2004(21):5-14.