THE LANCET
Oncology

Supplementary appendix

Supplement to: Maringe C, Spicer J, Morris M, et al. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *Lancet Oncol* 2020; published online July 20. http://dx.doi.org/10.1016/S1470-2045(20)30388-0.

The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study

Dr. Camille Maringe PhD, Prof. James Spicer MD PhD, Dr Melanie Morris PhD, Prof. Arnie Purushotham MD, Prof. Ellen Nolte MPH PhD, Prof. Richard Sullivan MD PhD, Prof. Bernard Rachet MD PhD, Dr Ajay Aggarwal MD PhD.

Technical appendix

1. Re-allocation of patients to new referral pathway

Each patient had equal probabilities to be reallocated to each of the emergency referral routes. We randomly generated their probabilities and selected those patients with random values below the thresholds detailed in the paper, and necessary to maintain proportions of patients re-allocated to the emergency referral pathway in keeping with the original distributions seen in pre-pandemic cohorts (see proportions in Table 1).

2. Estimating net survival and deaths due to cancer

Baseline, pre-pandemic, levels of cancer-specific survival were assessed through multivariable excess hazard models. We use the *strcs* package in Stata[1]. *strcs* implements a two-step method that incorporates both analytical and numerical integration to estimate the cumulative hazard function required for the log-likelihood function. Flexible parametric survival models are fit using maximum likelihood estimation.

The main assumption of excess hazard models is that the overall mortality of the cohort of patients (λ) is the sum of two forces of mortality: the excess mortality hazard (λ_E), assumed to be the mortality hazard directly or indirectly due to cancer, and the expected or other causes mortality hazard, which is considered to be well approximated by the general population mortality hazard (λ_P).

$$\lambda(t,\boldsymbol{x}) = \lambda_E\ (t,\boldsymbol{x}) + \lambda_P\ (a+t, y+t, \boldsymbol{z}),$$

The cancer mortality hazard, λ_E, at time t for given patient's covariates $\boldsymbol{x}$, such as age at diagnosis (a), deprivation levels, and referral pathway, is what we need to estimate. The following model is fitted:

$$\lambda_E\ (t,\boldsymbol{x}) = \lambda_0\ (t) * exp(\beta * \boldsymbol{x})$$

λ_E (i.e. hazard of death due to cancer) was modelled as a function of age at diagnosis (a), deprivation (d), and mode of presentation (p) as follows: non-linear effects of age at diagnosis (restricted cubic splines, a_1 and a_2) and time-dependent effects of each variable were allowed, as well as interactions between age at diagnosis and deprivation and between age at diagnosis and mode of presentation. The excess hazard at the reference value of all covariables, the baseline hazard, $\lambda_0\ (t)$, was modelled using polynomials of follow-up time defined in three contiguous time intervals (restricted cubic splines with 3 degrees of freedom) and smoothly joined at the intervals' boundaries.

$$\begin{aligned}\lambda_E\,(t,\boldsymbol{x}) = \lambda_0\,(t) * exp(\beta_{a,1}(t) * a_1 + \beta_{a,2}(t) * a_2 \\ + \sum_{j=2}^{5}(\beta_{d,j}(t) * I_{d=j} + \gamma_{a,1,j}(t) * a_1 I_{d=j} + \gamma_{a,2,j}(t) * a_2 I_{d=j}) \\ + \sum_{k=2}^{P}(\beta_{p,k}(t) * I_{p=k} + \alpha_{a,1,k}(t) * a_1 I_{p=k} + \alpha_{a,2,k}(t) * a_2 I_{p=k}))\end{aligned}$$

$\beta_{a,1}$ and $\beta_{a,2}$ are the effects of each component of age, $\beta_{d,j}$ are the effects of each deprivation quintile j, $j = 2, \dots 5$, $\beta_{p,k}$ are the effects of each mode of presentation, and $\gamma_{a,1,j}$, $\gamma_{a,2,j}$, $\alpha_{a,1,k}$, and $\alpha_{a,2,k}$ are the interactive effects on the excess hazard of death. Each effect is allowed to vary with follow-up time t. The best-fitting forms of effects were selected using a hierarchical model selection algorithm designed by Royston and Sauerbrei (mfpigen),[2,3] combined with the Akaike Criteria (AIC).[4] The effects selected are presented in the Table below.

When analysing population-based data, the measure of interest, excess mortality due to cancer, is conventionally retrieved by removing the impact of competing risks of death, i.e. the deaths from causes other than the cancer of interest. These competing risks, derived from general population life tables defined by sex, single years of age, calendar years, deprivation quintile, and Government Office Regions ($\boldsymbol{z}$), were assigned to each patient at their date of last known vital status.

	Main effects				Interactions	
	Age at diagnosis	*Referral pathway*	*Deprivation*	*Sex*	*Age* referral*	*Age * deprivation*
Breast	Non linear, non proportional	Categorical, non proportional	Categorical		Included	Included
Colorectum	Non linear, non proportional	Categorical, non proportional	Categorical	Proportional	Included	
Lung	Linear, non proportional	Categorical, non proportional	Categorical	Proportional	Included	Included, non proportional
Oesophagus	Non linear, non proportional	Categorical, non proportional	Categorical	Non proportional	Included	

Effects selected for each excess hazard model

The final model selected for each cancer was fitted on the pre-pandemic cohorts of patients. The estimated coefficients associated with the effects of each variable and the parameters corresponding to the baseline excess hazard were retained. These inform the prediction of excess hazard of death due to cancer for each patient i at selected times t, $\lambda_{E,i}(t)$. Such predictions were made for each patient in the setting of the observed pre-pandemic cohorts in addition to the three scenarios A-C. From the individual excess hazards, we derived the following quantities:

<u>Cohort net survival</u>: the survival of the cohort of cancer patients, assuming patients can only die of their cancer. $S_{N,i}$ is the individual net survival, and S_N is the cohort net survival, such that:

$$S_{N,i}(t, x_i) = \exp\left(-\int_0^t \lambda_{E,i}(u, x_i)du\right)$$

$$S_N(t) = \frac{1}{N}\sum_{i=1}^{N} S_{N,i}(t, x_i)$$

Crude probability of cancer death: this is the probability of cancer-related death for each patient, $CPD_{C,i}$, or on average in the cohort, in the presence of competing risks of deaths.

$$CPD_{C,i}(t, x_i) = \int_0^t S_{O,i}(u^-|x_i) * \lambda_{E,i}(u, x_i)du$$

$S_{O,i}(u^-)$ represent individual overall survival of patient i, estimated just before time u. These were derived from multivariable hazard models, adjusting for the effects of age at diagnosis, deprivation, sex and referral pathway on the overall (all-cause) hazard of death. We performed model selection identical to that explained for the excess hazard models.

Number of deaths due to cancer at time t, D_C: these are directly derived from the individual crude probabilities of death estimated at time t.

$$D_C(t, x_i) = \sum_{i=1}^{N} CPD_{C,i}(t|x_i)$$

Number of years of life expectancy lost due to cancer: this is the total number of years of life expectancy lost due to cancer-related mortality for the cohort of cancer patients. $LEL_C(a, b, x_i)$ defines the number of years of life expectancy lost due to deaths due to cancer between years a and b.

$$LEL_C(0, t, x_i) = \sum_{i=1}^{N} (CPD_{C,i}(0, t|x_i) - CPD_{C,i}(t, \infty|x_i)) * \int_t^\infty S_i{}^*(u|z_i)du$$

$e_{x,i}(t) = \int_t^\infty S_i{}^*(u|z_i)du$ is the life expectancy of patient i at time t.

Each of these quantities were compared between the pre-pandemic setting and the 3 scenarios explored up to 5 years following diagnosis. The differences provided an estimated decrease in net survival, additional number of deaths due to cancer and additional numbers of years of life expectancy lost due to cancer, namely:

$$DiffD_C{}^X(t, x_i) = D_C{}^X(t, x_i) - D_C{}^{PP}(t, x_i)$$

Whereby $D_C{}^X$ is the number of deaths due to cancer in Scenario X (X=A, B, or C) and $D_C{}^{PP}$ is the number of deaths due to cancer in the pre-pandemic period, and

$$DiffLEL_C{}^X(0, t, x_i) = LEL_C{}^X(0, t, x_i) - LEL_C{}^{PP}(0, t, x_i)$$

We make the conservative assumption that $CPD_{C,i}(t, \infty|x_i)$ are equivalent in the pre-pandemic cohort and the cohort in each scenario, leading to:

$$DiffLEL_C{}^X(0, t, x_i) = \sum_{i=1}^{N} (CPD_{C,i}{}^X(0, t|x_i) - CPD_{C,i}{}^{PP}(0, t|x_i)) * \int_t^\infty S_i{}^*(u|z_i)du$$

For the later, only the figures at t=5 years were calculated and presented.

We provide the point estimates and their 95% CI around the estimations of CPD_C, D_C, and LEL_C based on bootstrap samples.[5,6]

1. Bower H, Crowther MJ and Lambert PC. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. *Stata Journal* 2016; 16: 989-1012.

2. Royston P, Sauerbrei W. (2008) Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. Series in probability and statistics. John Wiley & Sons.

3. Maringe, C., Belot, A., Rubio, F.J., Rachet, B. Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology. BMC Med Res Methodol 19, 210 (2019).

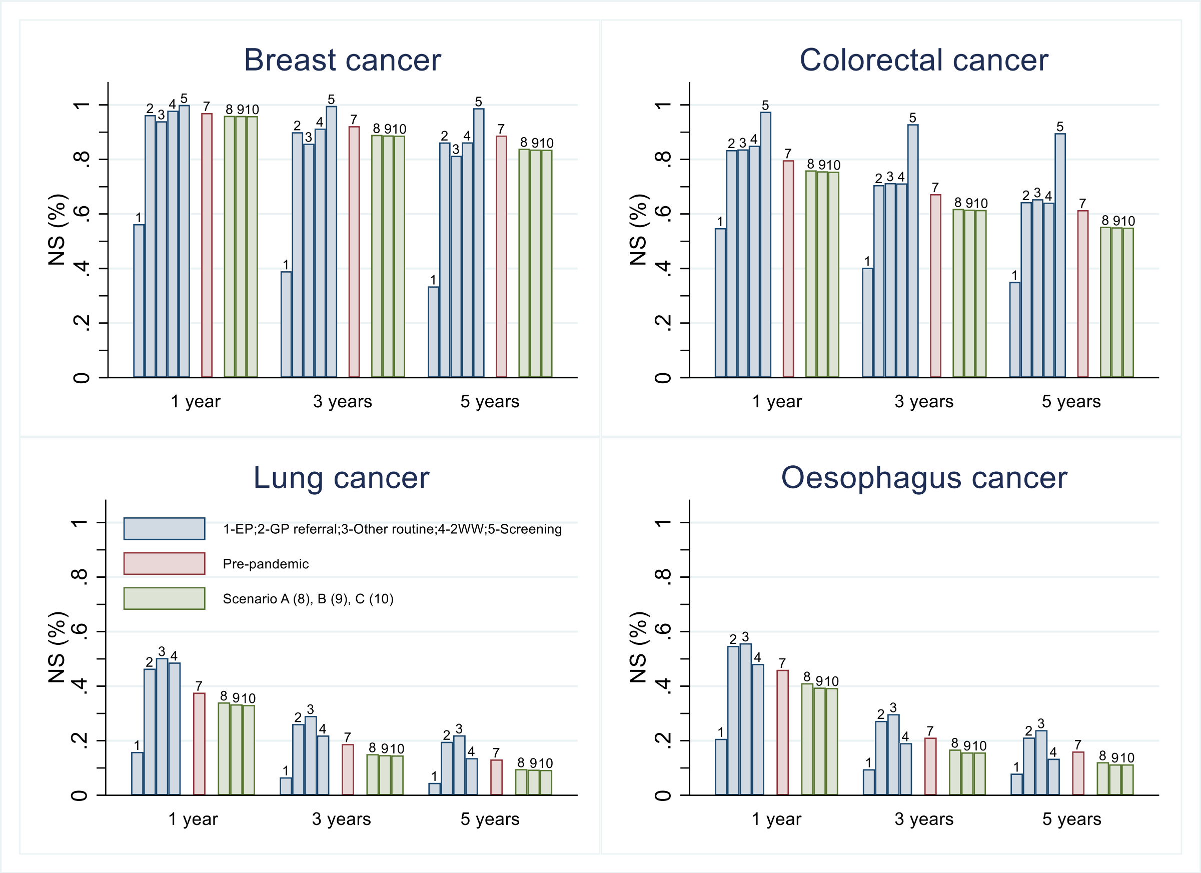
4. Maringe C, Belot A, Rachet B, Prediction of cancer survival for cohorts of patients most recently diagnosed using multi-model inference. *In press with Statistical Methods in Medical Research*.

5. Efron, B., Rogosa, D., & Tibshirani, R. (2004). Resampling methods of estimation. In N.J. Smelser, & P.B. Baltes (Eds.). International Encyclopedia of the Social & Behavioral Sciences (pp. 13216–13220). New York, NY: Elsevier.

6. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 2000;19(9);1141-6

Appendix Figure 1

1-, 3- and 5-year net survival, by referral pathway and overall, pre-pandemic and by scenario A-C


Breast cancer
Colorectal cancer
Lung cancer
Oesophagus cancer
NS (%)
1 year
3 years
5 years
1-EP;2-GP referral;3-Other routine;4-2WW;5-Screening
Pre-pandemic
Scenario A (8), B (9), C (10)