The AUGIS survival predictor: Prediction of long-term and conditional survival after esophagectomy using Random Survival Forests

First author: Saqib A Rahman. MRCS[1,5],

Co-authors: Robert C Walker MRCS[1], Nick Maynard FRCS[2], Nigel Trudgill MBBS[3], Tom Crosby FRCP[4], David A Cromwell PhD[5], Timothy J Underwood PhD[1], on behalf of the NOGCA project team and AUGIS

1. School of Cancer Sciences, Faculty of Medicine, University of Southampton
2. Oxford University Hospitals NHS Trust
3. Sandwell and West Birmingham Hospitals NHS Trust
4. Velindre Cancer Centre, Cardiff
5. Clinical Effectiveness Unit, Royal College of Surgeons of England

Corresponding Author
Professor T J Underwood
Email: tju@soton.ac.uk
Tel: +44 (0)2381206923
Fax: +44 (0)2381205152

Running Head
Augis-SURV: Prediction of survival after esophagectomy

Disclosures
The authors present no conflicts of interest.

Patient and public involvement
Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research. Patient consent for publication was not required.

Ethics approval
The study is exempt from UK National Research Ethics Committee approval as it involved secondary analysis of an existing dataset of anonymized data. The National esophago-Gastric (OG) Cancer Audit has approval for processing health care information under Section 251 (reference number: ECC 1-06 (c)/2011) for all National Health Service (NHS) patients diagnosed with OG cancer in England and Wales. Data for this study are based on patient-level information collected by the NHS, as part of the care and support of patients with cancer.

Acknowledgements and Funding
This study was undertaken as part of the work by the National esophago-Gastric (OG) Cancer Audit. The Audit is commissioned by the Healthcare Quality Improvement Partnership (HQIP) as part of the National Clinical Audit and Patient Outcomes Programme and funded by NHS England and the Welsh Government (www.hqip.org.uk/national-programmes). The authors had full independence from the Healthcare Quality Improvement Partnership. The aim of National esophago-Gastric Cancer Audit is to evaluate the care of patients with OG cancer in England and Wales, and support NHS providers to improve the quality of hospital care for these patients. More information can be found at: www.nogca.org.uk

SAR is supported by a Royal College of Surgeons of England Research Fellowship and a British Association of Surgical Oncologists Research Project Grant.

TJU is supported by a Cancer Research UK and Royal College of Surgeons of England Advanced Clinician Scientist Fellowship, ID:A23924.

This project has been supported by the Association of Upper Gastrointestinal Surgery (AUGIS), Heartburn Cancer UK, and The Royal College of Surgeons of England Surgical Specialty Lead Programme.

Disclaimer
Neither HQIP nor the funders had any involvement in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

NOGCA Project Team
Clincal Effectivness Unit: David Cromwell, Min Hae Park, Hussein Wahedally
Clinical leads: Nick Maynard, Tom Crosby, Nigel Trudgill
NHS Digital: Jane Gaskell and Rose Napper

Mini-Abstract

In this study we aimed to derive a prediction model for estimating overall survival after esophagectomy for cancer using a random survival forest methodology and 6399 patients. The technique provided excellent accuracy in characterising postoperative survival and provided significantly greater discrimination than using TNM status alone.

ABSTRACT

Objective

To develop a predictive model for overall survival after esophagectomy using pre/postoperative clinical data and machine learning.

Summary Background Data

For patients with esophageal cancer, accurately predicting long-term survival after esophagectomy is challenging. This study investigated survival prediction after esophagectomy using a Random Survival Forest (RSF) model derived from routine data from a large, well curated, national dataset.

Methods

Patients diagnosed with esophageal adenocarcinoma or squamous cell carcinoma between 2012 and 2018 in England and Wales who underwent an esophagectomy were included. Prediction models for overall survival were developed using the RSF method and Cox regression from 41 patient and disease characteristics. Calibration and discrimination (time dependent AUC) were validated internally using bootstrap resampling.

Results

The study analysed 6399 patients, with 2625 deaths during follow-up. Median follow-up was 41 months. Overall survival was 47.1% at 5 years. The final RSF model included 14 variables and had excellent discrimination with a 5-year tAUC of 83.9% (95%CI 82.6-84.9%), compared to 82.3% (95%CI 81.1-83.3%) for the Cox model. The most important variables were lymph

node involvement, pT stage, CRM involvement (tumour at <1mm from cut edge) and age. There was a wide range of survival estimates even within TNM staging groups, with quintiles of prediction within Stage 3b ranging from 12.2-44.7% survival at 5 years.

Conclusions

An RSF model for long-term survival after esophagectomy exhibited excellent discrimination and well calibrated predictions. At a patient level, it provides more accuracy than TNM staging alone and could help in the delivery of tailored treatment and follow-up.

INTRODUCTION

Esophagectomy for cancer is a highly morbid operation from which patients frequently take more than 18 months to recover. [1–3] Long-term prognosis for patients also remains poor, with 5-year survival estimated to be less than 50%. [4]

Currently, clinicians have a limited number of tools to identify patients with esophageal cancer who are likely to respond well to surgery and those who may not. TNM staging is widely used for patient stratification, but the classification is based on largely historic data (patients treated in 1980s to 2000s).[5] In addition, staging groups remain coarse, even with the introduction of post-neoadjuvant staging (i.e. ypTNM) in TNM 8.[6] Important characteristics that are readily available and routinely collected (such as circumferential resection margin) are not considered for the sake of simplicity, leading to a range of survival outcomes for patients within the same stage groups. This makes application at the patient level inaccurate.

The delivery of personalized long-term survival estimates after treatment for esophageal cancer is challenging. In addition to informing patients, reliable survival figures would enable the identification of high-risk individuals or groups in whom enhanced surveillance or treatment intensification (with traditional or novel agents such as immunotherapy) could be considered, or conversely patients where de-escalation would be the preferred option.

Prognostic models can address these limitations by combining multiple risk factors, although none have entered widespread use among surgeons or oncologists treating esophageal cancer. [7,8]Models based on Machine Learning (ML) techniques may produce more accurate predictions than models built using traditional statistical methods (e.g. logistic/cox regression). [9,10] In particular, Random Survival Forest (RSF) models have produced promising results[11–13] in various settings, and in esophageal cancer were used to derive the AJCC TNM 7th and 8th edition staging manuals,[5,6] and to

quantify the benefits of optimising treatment.[14] RSF is a machine-learning method that, when developed to predict survival, builds many decision trees with log-rank test based split points to identify different survival trajectories, with the predicted probability for an individual being derived as the average prediction across all of the trees.

The aim of this study was to derive and validate a prognostic model based on Random Survival Forest methods for long-term survival after esophagectomy for cancer, and to compare its performance to a model developed using a common statistical approach (Cox regression), using a population-based dataset from England and Wales.

METHODS

Study cohort

The study used a linked dataset prepared by the National esophago-Gastric Cancer Audit (NOGCA), a national clinical audit of patients undergoing treatment for cancer of the esophagus or stomach in England and Wales.[15] The audit was commissioned by the Healthcare Quality Improvement Partnership (HQIP) and funded by NHS England and the Welsh government. Patients were eligible for inclusion in the audit if they had a histological diagnosis of epithelial cancer, with the first patients being registered in April 2012. The audit collects a dataset that covers the care pathway from diagnosis to the end of initial treatment and links these patient records with information from other national health care datasets, including the National Cancer Registration and Analysis Service (NCRAS, see [15] for more details). Data collection was approved by the Confidentiality Advisory Group under section 251 of the NHS Act 2006.

Ethics approval

The study is exempt from UK National Research Ethics Committee approval as it involved secondary analysis of an existing dataset of anonymized data. The National esophago-Gastric (OG) Cancer Audit has approval for processing health care information under Section 251 (reference number: ECC 1-06 (c)/2011) for all National Health Service (NHS) patients diagnosed with OG cancer in England and Wales. Data for this study are based on patient-level information collected by the NHS, as part of the care and support of patients with cancer.

The study cohort included patients diagnosed with adenocarcinoma or squamous cell carcinoma of either the esophagus or gastro-esophageal junction (Siewert I – II) between 1 April 2012 and 31 March 2018 who underwent a planned curative esophagectomy. The study excluded patients who died in hospital prior to discharge, had confirmed metastatic disease on post-operative histology or had an inadequate lymphadenectomy (<15 lymph nodes)[16], in whom interpretation of lymph node

status would be biased. Supplementary Figures S.1 and S.2 details the patient exclusions and assumptions to derive the final sample size (n=6399).

The primary outcome was overall survival from the date of discharge following surgery. Survival was confirmed by linking the audit records with records from the Office for National Statistics (ONS) death register. Median duration of follow-up was 41 months (IQR 24-59).

Variable definition

The audit data contained 41 variables that were routinely measured in clinical practice, were beyond the control of the provider, had >50% completeness, and were clinically relevant to survival, listed in Table S.1. The dataset contained patient characteristics, disease information, details of treatment received, postoperative complications and tissue pathology. Circumferential resection margins were considered involved if there was tumour at <1mm from the cut edge and longitudinal resection margins were considered involved if tumour was found at the cut edge, in line with Royal College of Pathologists Guidelines.[17] In patients undergoing neoadjuvant therapy, treatment was specified as 'complete' if it was completed as prescribed or 'not complete' (due to disease progression, treatment toxicity, technical problems or patient choice). Malignant esophageal and gastric surgery is centralised in England and Wales and undertaken solely by dedicated teams. We therefore defined annual hospital volume as average number of major upper gastrointestinal resections (esophagectomy/major gastrectomy) per year, in line with NHS commissioning guidelines.[18] Staging was conducted using the 8th edition of the AJCC TNM staging manual.

Among the 41 variables considered for inclusion, five had missing values for more than 5% of patients: completion of neoadjuvant treatment (19.9%), return to theatre (15.8%), grade of differentiation (7.0%), cT stage (5.9%) and surgical approach (5.7%). Missing data was assumed to be

missing at random and was addressed using multiple imputation by chained equations (MICE) with 10 imputations.[19]

Model development

The study aimed to develop a model using a subset of variables so that data collection would be straightforward, and the model easy to use in clinical settings. To select core variables, we used permutation based Random Forest variable importance (VIMP)[11] with bootstrapped confidence intervals. Variables with a $p<0.01$ for VIMP greater than 0 were included in the final model (Table S.2). Pre-treatment histology (i.e. Adenocarcinoma or SCC) was also included to improve the face validity of the model. The final model was trained using 14 variables; Age, Gender, cT, cN, Site of Tumor, Pre-treatment histology, Neoadjuvant Treatment, Completion of neoadjuvant treatment, pT/ypT, number of positive lymph nodes, circumferential and longitudinal margin involvement, grade of differentiation and presence/absence of surgical complications.The RSF hyperparameters (i.e. number of trees, number of variables per tree and minimum node size) were optimIzed by grid search. Final predictions were combined across the imputed data.[20–22]

A cox regression model was also developed using the same set of variables. Not all relationships between survival and continuous variables were linear, and a square root transformation was adopted for positive lymph nodes, while age was included as a restricted cubic spline.

Assessment of model performance

Model performance was quantified by discrimination and calibration. Discrimination was assessed using the time dependent area under the receiver operator curve (tAUC)[23]. Here this represents the proportion of random pairs of patients (one alive at time point 't' and one dead before this) where the model gives the patient who is alive a higher probability of survival than the patient who is dead. It can be considered analogous to the standard AUC in a binary regression model, extended to

survival by weighting of censored patients,[24] and has advantages over the C-statistic measure of performance.[25] Assessment of calibration was conducted visually for five patient subgroups of increasing risk (i.e. patients were grouped by quintiles of predicted risk of mortality at 5 years). In addition, we calculated the integrated brier score.[26,27] A score closer to 0 indicates better accuracy of predictions.

Finally, the relative performance of the two models was compared using decision curve analysis(DCA).[28] This method is based on evaluating the 'net-benefit' of model predictions across of range of possible decision thresholds that reflect how a patient might weigh the risk of harm associated with a false positive result (compared with a true positive result). Models with a better performance have a greater net benefit across all thresholds of probability.

Data analysis was conducted in R 3.5.3.[29] The RSF was trained using the packages Ranger[30] and RF-SRC[31]. This study was conducted to comply with the AJCC prognostic model[32] and TRIPOD[33] criteria, a compliance checklist is provided in Table S.3. Complete R code to reproduce the analysis is available on request. More extensive methodology and instructions to perform external validation are provided in the supplementary materials. All performance metrics were validated internally by the 0.632 estimator[34] in 1000 replications of the bootstrap with replacement.

RESULTS

The study included 6399 patients with esophageal cancer who underwent an esophagectomy between April 2012 and March 2018. Table 1 summarizes the characteristics of patients and their treatment. The median age at diagnosis was 66 years and only 1 in 5 were women. Tumours were predominantly adenocarcinoma (87%) and about 3 in 10 were classified as GEJ-Siewert I-II. There were 2625 recorded deaths, and the median survival was 53 months. Survival at 1, 3 and 5 years was 83.7%, 57.1% and 47.1% respectively **(**Figure 1). Differences in survival stratified by stage according to if patients received neoadjuvant treatment (i.e. ypTNM) or surgery alone (i.e. pTNM) are shown and discussed in supplementary Figure S.9.

A total of 13 variables were identified as important to include in the final model in addition to histological diagnosis. The RSF variable importance measure indicated the number of lymph nodes as the most important single risk factor for worse prognosis followed by pT/ypT stage (see partial dependence plots, supplementary figures S.3/S.4).

Model performance: internal validation

The RSF model demonstrated excellent discrimination, with a bootstrapped tAUC at 60 months of 83.9% (95% CI 82.6-84.9%), which was similar at other time points (Figure S.6). This was better than the Cox regression model (coefficients of which are given in Table S.4), which had a bootstrapped tAUC of 82.3% (95%CI 81.1-83.3%) and TNM stage alone (tAUC 74.5%). Figure 2 shows the agreement between the RSF model predicted and observed survival times for patients grouped according to quintile of prediction and in both models, calibration was visually good throughout these groups. The integrated brier scores for the RSF model was superior to the cox regression at 0.136 (95%CI 0.134-0.138) and 0.141 (0.139-0.143), respectively. Decision curve analysis also showed a greater net benefit for the RSF over Cox regression model (Figure S.7) or using TNM alone (Figure S.8).

There was a broad range of predictions yielded even within p/ypTNM staging groups, with the lowest risk quintile of Stage 3b patients having a predicted 5-year survival of 44.7% compared to 12.2% for the highest risk quintile. Moreover, there is a subgroup of early stage disease (TNM stage 0-1), who would generally be considered to be cured, who had a relatively poorer survival of only 64.7% at 5 years (Figure 3) and overlap of quintiles between staging groups.

Figure 4 gives an overview of mean predicted 5-year survival for combinations of the most important variables (Lymph node status, T-stage, CRM involvement and age at diagnosis). Age at diagnosis is most influential with early stage (T0-2,N0-1) disease, however its importance diminishes with increasing T/N-stage. Examples of how the model might be used are given in the supplementary materials (Figure S.10)

DISCUSSION

Accurate predictions of long-term survival following surgery for esophageal cancer may help clinicians and patients. This study has demonstrated that an RSF model can discriminate between patients with different long-term prognoses using a small number of routinely collected variables. The model showed very good calibration and discrimination on internal validation, and exceeded that achieved using Cox regression analysis. The model is applicable to patients who have undergone a planned curative esophagectomy for adenocarcinoma or squamous cell carcinoma of the esophagus, who had an adequate lymphadenectomy and survived to discharge from hospital.

At present, information given to patients after surgery about their long-term survival is limited and is largely based on TNM staging. This can mean the information provided to patients can be vague, such as '50% survive to 5 years'. Decisions on whom to offer adjuvant treatment or consider for entry into trials may involve more criteria than TNM staging, but the relationship between these criteria and survival may be uncertain.

The model described here provides a more precise prediction of prognosis for an individual patient than TNM staging alone, and this will be valuable in postoperative discussions with patients. This increased accuracy has several benefits. In a research setting, it is key for establishing the efficacy of treatment. In clinical practice, it supports selecting the right patients for the right treatments, particularly with the emergence of novel therapies (e.g. Immune checkpoint inhibitors[35]). Further research on how best to communicate predicted survival to patients is required, even in early stage disease, desire for detail of prognosis is highly variable,[36] and the effective use of decision aids is challenging.[37]

The model compares favourably to those published previously. Cox proportional hazard models using a variety of predictors have reported C-index/tAUCs of between 0.61-0.70.[38–40] In comparison

the C-index of our model was 0.76 (0.75-0.78) and the 5-year tAUC 0.84 (0.83-0.85). It also is more broadly applicable and includes patients with all modalities of neoadjuvant treatment.

The variables found to be most influential are consistent with clinical experience and the findings from other studies, with lymph node involvement being widely recognized as the most influential determinant of long-term survival in esophageal cancer. Both clinical T stage and clinical N stage were found to be important independent of their pathological equivalents. There is some logic to support including both clinical and pathological variables, in that changing variables within patients may indicate the impact of neoadjuvant treatment (although this is limited by the relatively decreased accuracy of clinical staging). This is supported by recent studies which have shown that downstaging after neoadjuvant treatment improves absolute survival independent of the ypTNM stage.[41] Completion of neoadjuvant treatment was included which is biologically sensible, and important in the context of the increasing use of potentially more toxic regimens such as FLOT.[4]

The main strength of this study is the large sample size from a national population. The case ascertainment of esophagectomies exceeds 90% in the national audit, and the dataset was representative of patients within England and Wales who underwent curative surgery. Another strength is the linkage of audit records with ONS mortality data which enabled complete follow-up.

There are a number of limitations to the approaches taken in this study. Despite being more accurate than TNM staging at the individual patient level in the post-operative setting, no attempt has been made to develop a pre-treatment predictive model and cTNM remains the gold-standard in this domain. The NOGCA lacks several data items known to influence survival such as tumour regression grade (TRG)[42] and lympho-vascular invasion.[43] Additionally length of tumour[44] and BMI[45] could be considered, but were only available in recent years and therefore had too many missing

values. There was also no clear information on what adjuvant treatment this patient cohort had received in addition to their neoadjuvant/perioperative treatment.

Involvement of circumferential margin was defined according to Royal College of Pathologists (RCP) criteria, i.e. <1mm from cut edge is involved. Throughout much of the rest of the world the American College of Pathologist guidelines are used,[46] i.e. involved if tumour at cut edge. There is considerable debate about the most appropriate measure,[47] and this model requires validation if it is to be used with this definition. It was not possible to use T stage subdivided into 'a' or 'b' because not all patients were recorded with this information. Consequently, the analysis used the base T stage only. Patients treated solely with endoscopic techniques (mucosal resection or submucosal dissection) who did not require surgery were also excluded and it is not appropriate to use the model in this patient group.

The Esophageal Complications Consensus Group – ECCG[48] has recently specified and defined a core set of complications for esophagectomy which has been adopted worldwide.[49] The NOGCA relies on reporting from local cancer centres and pragmatically uses a limited set of complications with broader definitions. In this study the reported rate of complications was 40.0%, which is significantly less than the figures from the ECCG data (59%). This is likely to reflect the varying definitions and under reporting rather than a truly lower rate, which may explain the low overall importance of complications and absence of specific complications (e.g. anastomotic leak) in the model. The ECCG classification of complications has recently been adopted into the national Cancer Outcomes and Services Dataset (COSD) used in cancer registration within England, so more accurate complication data will be available in future iterations of the model.

CONCLUSIONS

Using a large, nation-wide, contemporaneous clinical dataset, this study has demonstrated the ability of a Random Survival Forest model to provide accurate predictions of long-term survival after surgery for esophageal cancer. A key benefit of the model is its performance in identifying patients with the same disease stage who have diverging 5-year survival. For example within Stage 3b, the largest group with 2023 patients, the model identifies a low risk quintile of patients with a predicted 5-year survival more than three times the highest risk quintile (44.7% vs 12.2%). These groups will likely benefit from different post-operative monitoring and/or treatment strategies. A similar pattern is seen with stage 4a disease (21.7% vs 6.1% 5-year survival), suggesting that there is a subgroup even in the most advanced (non-metastatic disease) who might be well served by targeted intervention.

The RSF model described in this paper is available at https://uoscancer.shinyapps.io/AugisSurv/ and could be a valuable prognostication tool for patients, surgeons and oncologists. In the future, it may also be useful to guide adjuvant treatment. External validation of this tool in other healthcare systems would be of benefit to confirm its performance.

References

1. Zhang Y, Yang X, Geng D, Duan Y, Fu J. The change of health-related quality of life after minimally invasive esophagectomy for esophageal cancer: a meta-analysis. *World J Surg Oncol*. 2018;16(97):1-13. doi:10.1186/s12957-018-1330-9
2. Derogar M, Lagergren P. Health-related quality of life among 5-year survivors of esophageal cancer surgery: A prospective population-based study. *J Clin Oncol*. 2012;30(4):413-418. doi:10.1200/JCO.2011.38.9791
3. Geeraerts M, Silva Corten LC, van Det M, et al. Insights in work rehabilitation after minimally invasive esophagectomy. *Surg Endosc*. 2019;33(10):3457-3463. doi:10.1007/s00464-018-06626-5
4. Al-Batran S-E, Homann N, Pauligk C, et al. Perioperative chemotherapy with fluorouracil plus leucovorin, oxaliplatin, and docetaxel versus fluorouracil or capecitabine plus cisplatin and epirubicin for locally advanced, resectable gastric or gastro-oesophageal junction adenocarcinoma (FLOT4): a randomised phase 3 trial. *Lancet*. 2019;393. doi:10.1016/S0140-6736(18)32557-1
5. Rice TW, Rusch VW, Ishwaran H, Blackstone EH. Cancer of the esophagus and esophagogastric junction: Data-driven staging for the seventh edition of the American Joint Committee on Cancer/International Union Against Cancer Cancer Staging Manuals. *Cancer*. 2010;116(16):3763-3773. doi:10.1002/cncr.25146
6. Rice TW, Patil DT, Blackstone EH. 8th edition AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: application to clinical practice. *Ann Cardiothorac Surg*. 2017;6(2):119-130. doi:10.21037/acs.2017.03.14
7. Gupta V, Coburn N, Kidane B, et al. Survival prediction tools for esophageal and gastroesophageal junction cancer: A systematic review. *J Thorac Cardiovasc Surg*. 2018;156(2):847-856. doi:10.1016/j.jtcvs.2018.03.146
8. van den Boorn HG, Engelhardt EG, van Kleef J, et al. Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. *PLoS One*. 2018;13(2):e0192310. doi:10.1371/journal.pone.0192310
9. Caruana R. An Empirical Comparison of Supervised Learning Algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh; 2006.
10. Fernández-Delgado M, Cernadas E, Barro S, Amorim D, Amorim Fernández-Delgado D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J Mach Learn Res*. 2014;15:3133-3181. doi:10.1016/j.csda.2008.10.033
11. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*.

2008;2(3):841-860. doi:10.1214/08-AOAS169

12. Hu C, Steingrimsson JA. Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. *J Biopharm Stat*. 2018;28(2):333-349. doi:10.1080/10543406.2017.1377730
13. Dietrich S, Floegel A, Troll M, et al. Random Survival Forest in practice: A method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol*. 2016;45(5):1406-1420. doi:10.1093/ije/dyw145
14. Rice TW, Lu M, Ishwaran H, Blackstone EH. Precision Surgical Therapy for Adenocarcinoma of the Esophagus and Esophagogastric Junction. *J Thorac Oncol*. 2019. doi:10.1016/j.jtho.2019.08.004
15. Cromwell D, Wahedally H, Park MH, et al. *National Oesophago-Gastric Cancer Audit 2019*.; 2019.
16. Allum WH, Blazeby JM, Griffin SM, Cunningham D, Jankowski JA, Wong R. Guidelines for the management of oesophageal and gastric cancer. *Gut*. 2011;60(11):1449-1472. doi:10.1136/gut.2010.228254
17. RCPath Cancer Services Working Group. *Dataset for the Histopathological Reporting of Oesophageal Carcinoma (2nd Edition)*.; 2013.
18. NHS England. *2013/14 NHS Oesophageal and Gastric Cancer Commissioining Guidelines*.; 2013.
19. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45(3):1-67. https://www.jstatsoft.org/v45/i03/.
20. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Med Res Methodol*. 2009;9(1):1-8. doi:10.1186/1471-2288-9-57
21. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical J*. 2015;57(4):614-632. doi:10.1002/bimj.201400004
22. Hosmer D, Lemeshow S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 1st ed. New York: John Wiley & Sons, Inc.; 1999.
23. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Med Res Methodol*. 2017;17(1):1-19. doi:10.1186/s12874-017-0332-6
24. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks.

Stat Med. 2013;32(30):5381-5397. doi:10.1002/sim.5958

25. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of 't'-year predicted risks. *Biostatistics*. 2018;20(2):347-357. doi:10.1093/biostatistics/kxy006
26. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18(1718):2529-2545. doi:10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.3.co;2-x
27. Kronek LP, Reddy A. Logical analysis of survival data: Prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*. 2008;24(16):248-253. doi:10.1093/bioinformatics/btn265
28. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565-574. doi:10.1177/0272989X06295361
29. R Core Team. R: A Language and Environment for Statistical Computing. 2019.
30. Wright MN, Ziegler A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J Stat Softw*. 2017;77(1):1-17. doi:10.18637/jss.v077.i01
31. Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.9.1. 2019. https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf. Accessed November 27, 2019.
32. Moons KGM, Weiser MR, Lu Y, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA Cancer J Clin*. 2016;66(5):370-374. doi:10.3322/caac.21339
33. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1):55-63. doi:doi:10.7326/M14-0697
34. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*. 1983;78(382):316-331. doi:10.1080/01621459.1983.10477973
35. Kato K, Cho BC, Takahashi M, et al. Nivolumab versus chemotherapy in patients with advanced oesophageal squamous cell carcinoma refractory or intolerant to previous chemotherapy (ATTRACTION-3): a multicentre, randomised, open-label, phase 3 trial. *Lancet Oncol*. 2019;20(11):1506-1517. doi:10.1016/S1470-2045(19)30626-6
36. Mühlbauer V, Berger-Höger B, Albrecht M, Mühlhauser I, Steckelberg A. Communicating prognosis to women with early breast cancer - Overview of prediction tools and the development and pilot testing of a decision aid. *BMC Health Serv Res*. 2019;19(1):1-15. doi:10.1186/s12913-019-3988-2

37. Kopecky KE, Urbach D, Schwarze ML. Editorial - Risk Calculators and Decision Aids Are Not Enough for Shared Decision Making. *JAMA Surg*. 2018:E1-E2. doi:10.1016/j
38. Gabriel E, Attwood K, Shah R, Nurkin S, Hochwald S, Kukar M. Novel Calculator to Estimate Overall Survival Benefit from Neoadjuvant Chemoradiation in Patients with Esophageal Adenocarcinoma. *J Am Coll Surg*. 2017;224(5):884-894.e1. doi:10.1016/j.jamcollsurg.2017.01.043
39. Goense L, van Rossum PSN, Xi M, et al. Preoperative Nomogram to Risk Stratify Patients for the Benefit of Trimodality Therapy in Esophageal Adenocarcinoma. *Ann Surg Oncol*. 2018;25(6):1598-1607. doi:10.1245/s10434-018-6435-4
40. Hagens ERC, Feenstra ML, Eshuis WJ, et al. Conditional survival after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. *Br J Surg*. February 2020. doi:10.1002/bjs.11476
41. Kamarajah SK, Navidi M, Wahed S, et al. Significance of Neoadjuvant Downstaging in Carcinoma of the Esophagus and Gastro-Esophageal Junction. *Ann Surg Oncol*. 2020. https://doi.org/10.1245/s10434-020-08358-0.
42. Noble F, Lloyd MA, Turkington R, et al. Multicentre cohort study to define and validate pathological assessment of response to neoadjuvant therapy in oesophagogastric adenocarcinoma. *Br J Surg*. 2017;104(13):1816-1828. doi:10.1002/bjs.10627
43. Tu CC, Hsu PK, Chien LI, et al. Prognostic histological factors in patients with esophageal squamous cell carcinoma after preoperative chemoradiation followed by surgery. *BMC Cancer*. 2017;17(1):1-9. doi:10.1186/s12885-017-3063-5
44. Zeybek A, Erdoğan A, Gülkesen KH, et al. Significance of tumor length as prognostic factor for esophageal cancer. *Int Surg*. 2013;98(3):234-240. doi:10.9738/INTSURG-D-13-00075.1
45. Gu WS, Fang WZ, Liu CY, et al. Prognostic significance of combined pretreatment body mass index (BMI) and BMI loss in patients with esophageal cancer. *Cancer Manag Res*. 2019;11:3029-3041. doi:10.2147/CMAR.S197820
46. College of American Pathologists (CAP). *Protocol for the Examination of Specimens From Patients With Carcinoma of the Esophagus*.; 2017. www.cap.org/cancerprotocols. Accessed October 2, 2018.
47. Quinn LM, Hollis AC, Hodson J, et al. Prognostic significance of circumferential resection margin involvement in patients receiving potentially curative treatment for oesophageal cancer. *Eur J Surg Oncol*. 2018;44(8):1268-1277. doi:10.1016/j.ejso.2018.05.017
48. Low DE, Alderson D, Cecconello I, et al. International consensus on standardization of data collection for complications associated with esophagectomy: Esophagectomy Complications

Consensus Group (ECCG). *Ann Surg*. 2015;262(2):286-294. doi:10.1097/SLA.0000000000001098

49. van der Werf LR, Busweiler LAD, van Sandick JW, van Berge Henegouwen MI, Wijnhoven BPL. Reporting National Outcomes After Esophagectomy and Gastrectomy According to the Esophageal Complications Consensus Group (ECCG). *Ann Surg*. January 2019. doi:10.1097/SLA.0000000000003210

FIGURE LEGENDS

Figure 1 Survival of patients who underwent an esophagectomy between April 2012 and March 2018, stratified by TNM stage

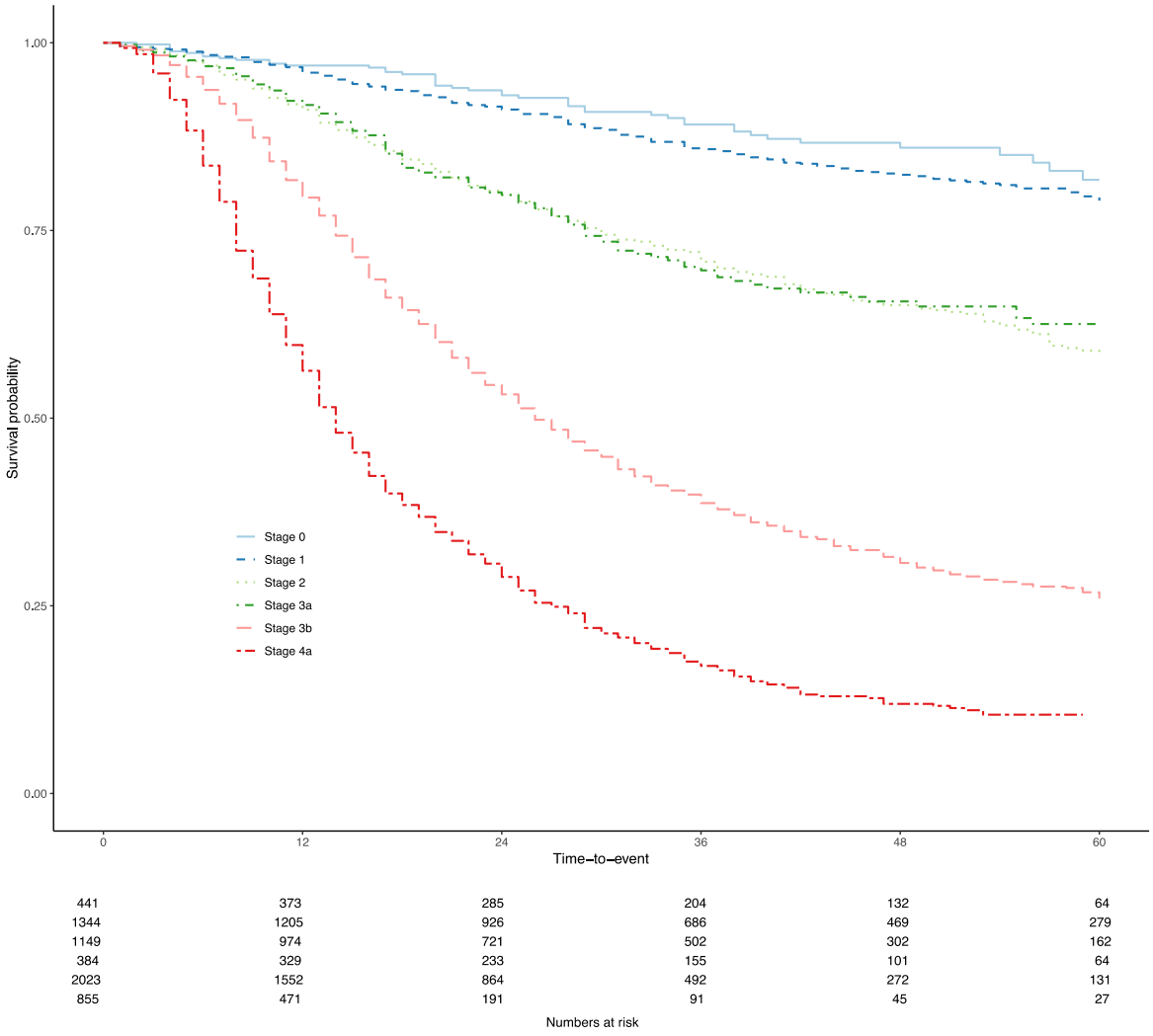


Figure 2 Calibration of predictions from RSF model. Patients grouped into quintiles according to predicted survival at 60 months post surgery

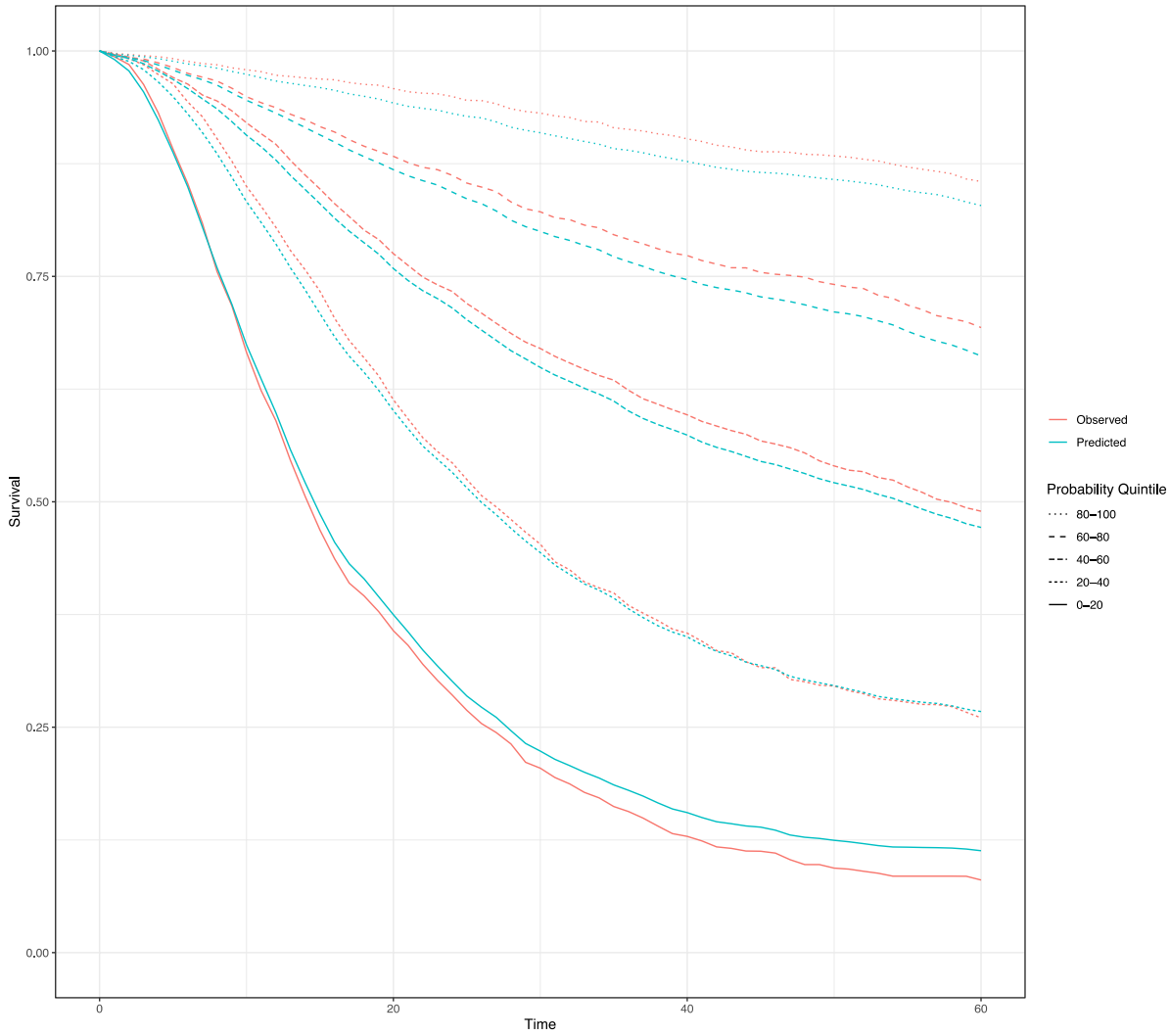


Figure 3 Range of predictions by p/ypTNM stage. (A) Stage 0-1, (B) Stage 2-3a, (C) Stage 3b, (D) Stage 4a. Patients were grouped into quintiles by predicted survival at 60 months, with the highest and lowest groups shown.

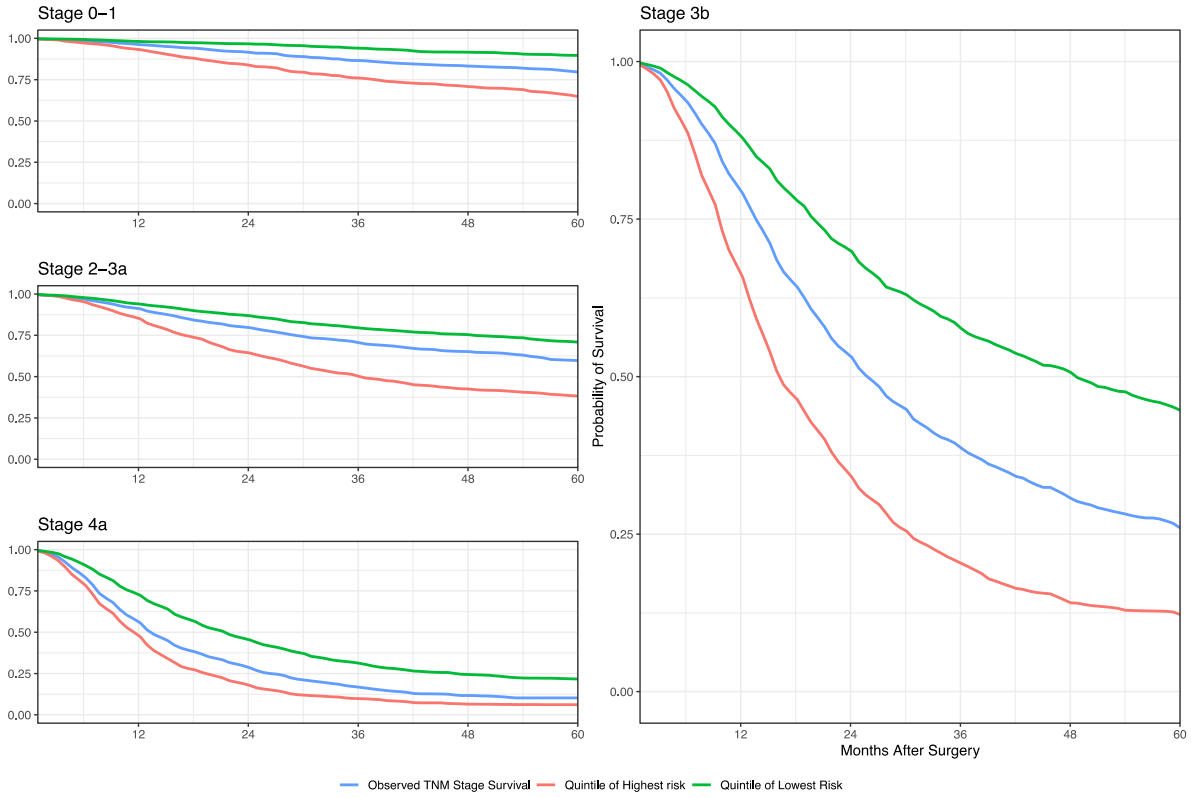


Figure 4 Predicted 5-year survival for a given combination of selected variables. Colors represent differing prognosis, with green more favourable and red less favourable.

			Lymph Node Status							
			N0		N1		N2		N3	
			CRM		CRM		CRM		CRM	
			-ve	+ve	-ve	+ve	-ve	+ve	-ve	+ve
p/ypT Stage	T0	Age 18-50	82.6%	65.2%	74.6%	54.6%	33.9%	31.7%	30.9%	25.6%
		Age 51-60	83.3%	65.5%	74.8%	54.7%	33.6%	31.6%	30.4%	25.5%
		Age 61-70	85.0%	66.2%	74.1%	54.3%	33.4%	29.9%	30.6%	23.7%
		Age 71+	83.0%	63.2%	70.4%	52.5%	32.8%	30.6%	29.4%	22.8%
	T1	Age 18-50	73.6%	64.5%	69.9%	55.2%	34.1%	32.0%	31.9%	26.3%
		Age 51-60	74.3%	64.8%	70.1%	55.2%	33.7%	31.9%	31.4%	26.2%
		Age 61-70	76.8%	65.9%	69.7%	54.9%	33.6%	30.1%	31.7%	24.3%
		Age 71+	76.6%	63.4%	66.7%	53.1%	32.9%	30.9%	30.4%	23.4%
	T2	Age 18-50	81.0%	67.3%	73.7%	56.4%	36.0%	32.6%	33.4%	26.6%
		Age 51-60	81.2%	67.4%	73.7%	56.4%	35.6%	32.5%	32.9%	26.5%
		Age 61-70	81.7%	67.6%	72.8%	56.0%	35.2%	30.5%	32.9%	24.4%
		Age 71+	75.6%	62.6%	69.4%	54.1%	34.0%	31.0%	31.1%	23.3%
	T3	Age 18-50	74.0%	58.4%	59.6%	42.5%	33.6%	27.3%	25.1%	20.6%
		Age 51-60	74.5%	58.4%	59.3%	42.3%	33.1%	27.1%	24.4%	20.4%
		Age 61-70	68.7%	57.8%	54.8%	41.1%	33.0%	24.6%	26.0%	18.0%
		Age 71+	62.3%	48.5%	50.3%	38.4%	30.0%	25.0%	23.5%	16.6%
	T4	Age 18-50	36.1%	22.0%	31.0%	18.3%	13.7%	13.1%	11.0%	11.0%
		Age 51-60	36.2%	22.1%	31.0%	18.3%	13.6%	13.1%	11.0%	11.0%
		Age 61-70	33.9%	21.3%	28.9%	17.5%	12.9%	12.3%	10.6%	10.3%
		Age 71+	31.4%	19.6%	26.8%	16.7%	11.9%	11.4%	9.6%	9.2%
Key: Survival at 5 years			100%	50%	0%	CRM = Circumferential Resection Margin				

Table 1 Background characteristics of patients who underwent an esophagectomy between April 2012 and March 2018

Characteristic		N = 6399	Median Survival
Sex	Male	5045 (78.8)	47
	Female	1354 (21.2)	72
Age	0-40	64 (1.0)	NR
	41-50	405 (6.3)	68
	51-60	1397 (21.8)	59
	61-70	2615 (40.9)	60
	71-80	1787 (27.9)	42
	81+	131 (2.0)	29
Site of Tumor	Upper/Mid Esophagus	792 (12.4)	61
	Lower Esophagus	3795 (59.3)	52
	GEJ (S1-2)	1812 (28.3)	53
Histopathology	Adenocarcinoma	5540 (86.6)	51
	SCC	859 (13.4)	68
cT	T0/is/1	467 (7.3)	NR
	T2	1294 (20.2)	67
	T3	3979 (62.2)	38
	T4	284 (4.4)	36
	Unknown	375 (5.9)	NR
cN	N0	2551 (39.9)	76
	N1	2547 (39.8)	41
	N2	938 (14.7)	32
	N3	159 (2.5)	28
	Unknown	204 (3.2)	47
cM	M0	6151 (96.1)	54
	M1	44 (0.7)	26
	Unknown	204 (3.2)	47
ASA	1	892 (13.9)	64
	2	3745 (58.5)	53
	3	1726 (27.0)	42
	4	36 (0.6)	43
Approach	Open	3357 (52.5)	51
	Hybrid	1931 (30.2)	51
	MIO	748 (11.7)	NR
	Unknown	363 (5.7)	41
Anastomotic Leak	No	5923 (92.6)	54
	Yes	445 (7.0)	40
	Unknown	31 (0.5)	NR
Any Complication	No	3810 (59.5)	55
	Yes	2558 (40.0)	49
	Unknown	31 (0.5)	
Involved Longitudinal Margin	No	6188 (96.7)	55
	Yes	211 (3.3)	19
Involved Circumferential Margin	No	4617 (72.2)	77
	Yes	1534 (24.0)	21
	Unknown	248 (3.9)	57
pT/ypT	T0/is	524 (8.2)	NR
	T1	1201 (18.8)	NR
	T2	836 (13.1)	NR
	T3	3549 (55.5)	30
	T4	289 (4.5)	13
Lymph Nodes Examined		26 [15-130]	
pN/ypN	N0	2994 (46.8)	NR
	N1	1414 (22.1)	46
	N2	1133 (17.7)	22
	N3	858 (13.4)	14
Grade	G1 (Well)	226 (3.5)	NR
	G2 (Moderate)	2331 (36.4)	60
	G3/4 (Poor/Anaplastic)	2697 (42.1)	38
	GX (Unable to determine)	695 (10.9)	66
	Unknown	450 (7.0)	72
NAT	Chemotherapy	3976 (62.1)	42
	Chemoradiotherapy	450 (7.0)	NR
	None	1973 (30.8)	66
Completion of NAT	Completed	2981 (46.6)	51
	Not Completed	282 (4.4)	32
	Not Applicable	1861 (29.1)	67
	Unknown	1275 (19.9)	42
Annual Hospital Volume	1 to 30	504 (7.9)	59
	31 to 60	3351 (52.4)	55
	>60	2544 (39.8)	48

Data given as absolute number (percentage) and median [Range] KEY: NAT = Neoadjuvant Treatment, NR= median survival not reached. MIO = Minimally invasive esophagectomy

Supplementary Materials

Supplementary Methods

Missing Data

2817 cases out of 6399 had missing data (44.0%), with a total of 5015 missing data points (5015/270143, 1.9%). Variables considered that had >5% missing data were outcome of neoadjuvant treatment (22.3%), return to theatre (15.8%), differentiation grade (7.0%) and surgical approach (open/minimally invasive/hybrid, 5.7%). Circumferential resection margin, Clinical/pathological T/N/M stage, longitudinal resection margin, length of stay and survival time all had missing data in 1-5% of cases. All other variables had <1% missing data.

In this study missing data was assumed to be at random (MAR) and was handled using multiple imputation by chained equations (MICE),(1) with 10 iterations across 10 imputed datasets. MICE imputes iteratively, across variables, one at a time using a variety of methods according to the class of the data. In general, a type of regression model is generated with the missing value as the dependent variable and the known variables as the independent variables. Extension to the survival setting has been described previously(2,3). The optimal number of imputed datasets has been suggested to be 100 multiplied by the fraction of incomplete cases(4). Here, the number of imputed datasets was limited to 10 for computational reasons and 10 iterations were performed as suggested(1). All 41 candidate covariates (Table 1) were included in the imputation model, which should increase the quality of its specification. We elected to include cases with missing survival data (i.e. the outcome) as this have been shown to improve the imputation model(5) and is generally recommended(6). For the Cox-Proportional hazards method, final pooled models were generated. For the RSF, pooled survival probabilities were created as described below.

Training and Validation Datasets

The performance of the model was reported as the mean of 1000 bootstrap replications assessed using the 0.632 estimator. For each of the 1000 bootstrap resamples, an RSF was trained on a random sample of the whole dataset with replacement, meaning that approximately 63.2% of cases were selected randomly and some of these cases are then replicated to give a sample size of 6399 for model training. Having trained the model, its performance was then assessed on the samples not used for training (testing set) and on those samples used for model training (training set).

RSF Derivation

To derive the RSF, hyperparameters were optimised to mimimise prediction error (1 – c index) in the random forest Out-of-Bag (OOB) samples. Each tree in the random forest is trained on 2/3 of the cases with replacement (the in-bag samples) and then tested on the remaining 1/3 (the OOB samples), and hence the OOB error averaged across all trees is equivalent to the simple bootstrap with repetitions equal to the number of trees in the forest. Hyperparameters optimised were number of trees, number of variables in each decision tree and minimum node size.

This process was repeated for each imputed dataset to yield 10 separate models. It is not possible to directly combine RSF models generated on multiply imputed data as would be the case for Cox or Logistic Models, due to the absence of coefficients with this technique. To address this, predictions were directly combined(7,8). The standard error of prediction for each unique death time was calculated using the predictions from individual decision trees in the forest (i.e. 200-400 samples), allowing the predictions from each dataset to be combined using Rubin's rules after a complementary log-log transformation(9).

Assessment of Discrimination

In comparison to the binary outcome setting, there is no clear consensus on the assessment of discrimination in survival models. The Harell's C-statistic in a binary outcome model is equivalent to the area under the receiver operator characteristic (ROC) curve or AUC. It can be defined as the proportion of pairs of cases where one has an outcome and the other does not, that the model correctly orders their predictions (i.e. the case

with an outcome scores higher than the case without). In a survival setting, confusingly, Harrell's C-statistic is not equivalent to this, instead representing the proportion of random pairs of cases where the case which lives longer is given a higher probability of survival at a specified time point.

Although a reasonable diagnostic measure of a model, it differs crucially from the binary outcome setting with which most users are familiar and may be misinterpreted in that way. It has also been criticised for considering only 'usable pairs' (i.e. uncensored pairs of cases) and is hence heavily dependent on the censoring distribution of the study dataset. An alternative method of assessment that has increasing popularity due to its ready interpretability and better utilisation of censored data is the time-dependent area under receiver operating characteristic curve (tAUC).

To calculate the tAUC, firstly we must consider censoring. If we treat the censored patients as missing (i.e. ignore censoring), then a biased comparison of cases and controls is conducted. There are multiple ways of addressing this issue, and here we have used the Inverse Probability of Censoring Weighting (IPCW) , a standard technique for survival analysis to appropriately weight the cases, as described by Blanche et al(10). In short, surviving cases are weighted according to the inverse of the probability of being censored at each time point so that surviving cases are proportionally over-represented in the final comparison.

Secondly, we must decide how to account for time. In this study we used the Cumulative sensitivity/dynamic specificity (C/D) definitions (there are two alternative definitions, Incident sensitivity and dynamic specificity – I/D and Incident sensitivity and static specificity – I/S). Here, for each measured time point (i.e. death time), all patients are classified into being either cases (i.e. dead at that time point) or controls (i.e. alive at that time point). The cumulative sensitivity at time point 't' is the proportion of cases who are dead at time 't' who are given a probability of dying above a set cut-off point 'c'. The dynamic specificity is the proportion of cases who are alive at time 't' who are given a probability of dying below 'c'(11). Cases are weighted according to IPCW to yield a final sensitivity and specificity and calculated across a range of cut-points to derive the time dependent receiver operator curve (tROC), from which the tAUC can be calculated. The performance across all time points (integrated tAUC, iAUC) can also be calculated, and was 82.4% in the RSF. There was preservation of performance across time points (Figure S.6).

Decision Curve Analysis

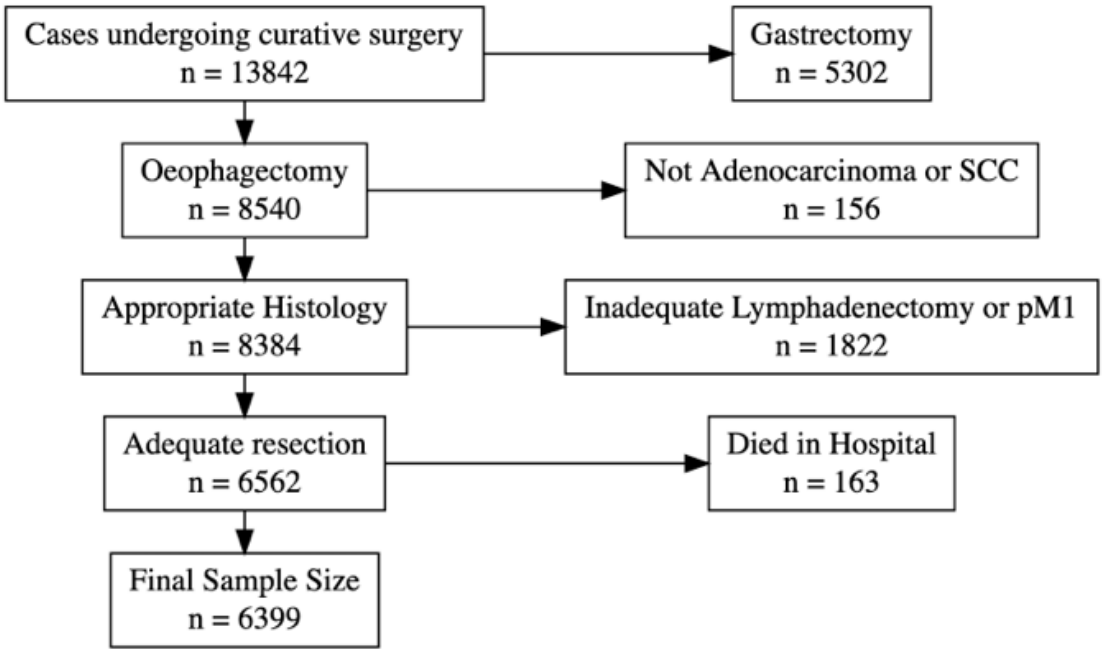
Predicted probabilities for each model were compared using decision curves. These were first developed in 2006 by Vickers et al.(12) and designed to provide more information on the clinical utility of a model than arbitrary statistics, such as the C-index. Decision curve analysis (DCA) calculates the 'net benefit' of decisions made across a range of probability thresholds by applying validation data to the formula:

$$NetBenefit = \frac{TP}{n} - \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right)$$

Where TP is the number of true positives, FP is the number of false positives, n is the number of patients and p_t is the chosen probability threshold. Plotting net benefit across all probability thresholds yields the decision curve. The principle is to assess the cost/benefit of basing a decision to intervene at a defined probability (i.e. p_t) of an event (e.g. disease present/absent), and has been previously extended to censored data(13). Having generated the curve, it is possible to see the range of probabilities of disease in which the model has utility for basing treatment decisions above that of treating all patients or none.

Furthermore, different clinical prediction models can be compared by the same technique. The DCA is important in that it can show that a model that both discriminates and calibrates well, is still useless in clinical practice and should be discarded due to the distribution of predicted probabilities(14). DCA can also be used to show if a particular model would be beneficial based on individual patients expectations of acceptable risk(15). In a survival setting, decision curves are generated for individual survival times, which we conducted annually between 1 and 5 years (Figure S.7). The net-benefit of the RSF model was greater than for the Cox regression predictions across all threshold probabilities. This pattern is preserved at all time periods. Importantly, the RSF model also has a net benefit over the use of mortality as a function of pathological TNM stage alone (Figure S.8).

Figure S.1 Study Flow Diagram



Factor levels in Clinical T Stage were collapsed so that T0,Tis and T1 were a single category due to low number of patients in these categories and similar observed survival (as shown below in *Figure S.2*). TX was treated as missing.

Figure S.2 Observed survival for 6399 patients after oesophagectomy, stratified by cT Stage

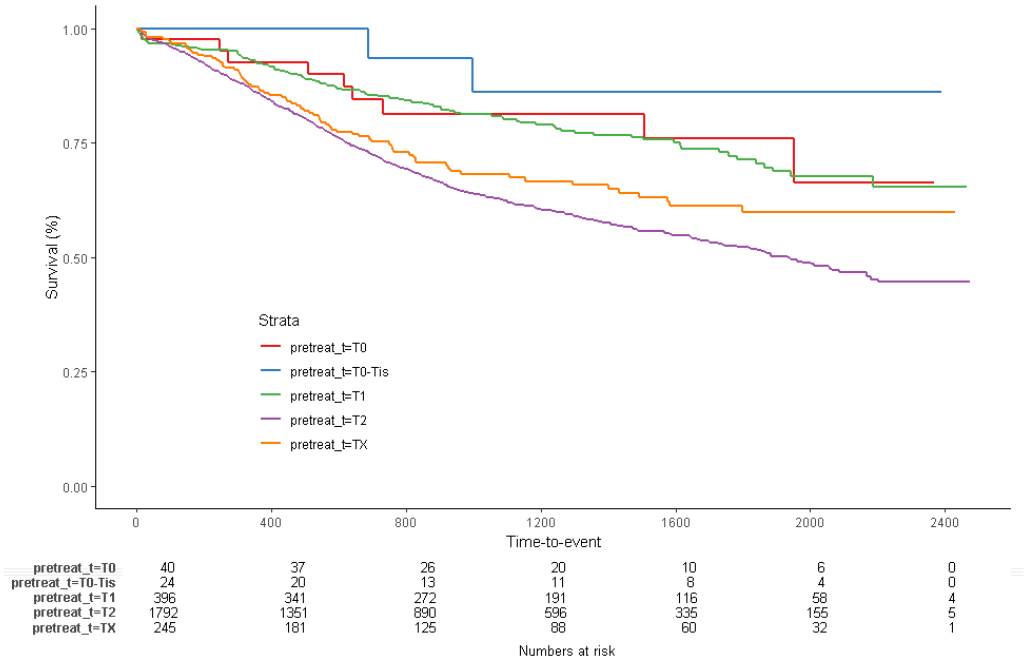


Table S.1 Candidate variables considered for inclusion

Preoperative	Operative	Post-Operative
Gender	Any Surgical Complication	Involved Longitudinal Margin
Age	Anastomotic Leak	Involved Circumferential Margin
Site of Tumour	Chyle Leak	Number of Positive Lymph Nodes
Histopathology	Bleeding Complication	pT/ypT
cT	Cardiac Complication	Differentiation Grade (Worst)
cN	Acute Kidney Injury	
cM	Pneumonia	
Neoadjuvant Treatment Modality	ARDS	
Completion of Neoadjuvant Treatment	Pulmonary Embolism	
Surgical Approach	Pleural Effusion	
Type of Operation	Any Respiratory Complication	
Cardiovascular Comorbidity	Infective Complication	
COPD	Return to Theatre	
Chronic Kidney Disease	Number of Operations	
Chronic Liver Disease		
Diabetes Mellitus		
Psychiatric Illness		
Cerebrovascular Disease		
Barrett's Oesophagus		
Performance Status (PS)		
ASA Grade		
Annual Hospital Volume		

Table S.2 Full Model VIMP, ordered by magnitude

	VIMP	LCI	UCI
Total Positive Lymph Nodes*	10.85	9.76	12.00
pT/ypT*	3.57	2.90	4.19
Involved Circumferential Margin*	1.69	1.41	2.00
Age*	0.39	0.28	0.49
cT*	0.32	0.19	0.45
cN*	0.22	0.11	0.34
Neoadjuvant Treatment*	0.15	0.08	0.22
Completion of Neoadjuvant Treatment*	0.13	0.06	0.22
Differentiation Grade (Worst)*	0.09	0.02	0.15
Involved Longitudinal Margin*	0.07	0.03	0.10
Gender*	0.06	0.01	0.12
Site of Tumour*	0.06	0.02	0.10
Any Complication*	0.05	0.00	0.11
Approach	0.03	-0.02	0.08
ASA	0.02	-0.04	0.08
Return to Theatre	0.02	-0.02	0.06
Procedure	0.02	-0.01	0.05
Histopathology*	0.02	-0.01	0.04
Anastomotic Leak	0.01	0.00	0.03
Hospital Volume	0.01	-0.02	0.05
COPD	0.01	-0.02	0.03
Barrett's	0.01	-0.01	0.02
Respiratory Complication	0.01	-0.02	0.03
ARDS	0.01	0.00	0.01
DM	0.00	-0.01	0.02
Chyle Leak	0.00	-0.01	0.02
IHD	0.00	-0.03	0.04
Pneumonia	0.00	-0.01	0.02
Cardiac Complication	0.00	-0.02	0.02
cM	0.00	0.00	0.00
CVD	0.00	-0.01	0.01
Number of Procedures	0.00	-0.02	0.02
Renal Complication	0.00	0.00	0.00
Pulmonary Embolism	0.00	0.00	0.00
Chronic Liver disease	0.00	0.00	0.00
Bleeding Complication	0.00	0.00	0.00
Pleural Effusion	0.00	-0.01	0.01
Psychiatric Illness	0.00	-0.01	0.00
CKD	0.00	0.00	0.00
Infective Complication	0.00	-0.01	0.00
Performance Status	0.00	-0.05	0.05

*Included in final model

Table S.3 AJCC Prognostic Model Criteria

Inclusion Criteria	Checklist
The probability of overall survival, disease-specific survival (DSS), or disease-specific mortality (DSM) must be the outcome predicted	*Predicts Overall Survival*
The model should address a clinically relevant question	*Yes*
At face value, the model should include the relevant predictors, or explain why something relevant was not included	*Yes*
The model validation study should specify precisely which patients were used to evaluate the model and the validation dataset's inclusion/exclusion criteria	*See Patient Flow Diagram*
The model should be assessed for generalizability and external validation	*Internal validation conducted*
The model should have a well-defined prognostic time zero	*Time of Surgery*
All predictors must be known at time zero and sufficiently defined for someone else to use	*Yes*
Sufficient detail must be available to implement the model OR the author must allow free access to the model.	*Yes*
A measure of discrimination must have been reported	*Yes*
Calibration in the small must be assessed (from the external validation data set) and provided	*Yes*
The model should be validated over a time frame and practice setting that is relevant to contemporary patients with disease	*Yes*
It should be clear which initial treatment(s), if any, were applied, and with what frequency	*Yes*
The development and/or the validation of the prediction model must appear as a peer-reviewed journal article	*Yes*
Exclusion Criteria	
A substantial proportion of patients had essentially no follow up, either missing entirely or very short censored follow up, in the validation dataset	*No*
No information on number of missing values in validation dataset	*Provided*
The number of events in the validation dataset is small (<100)	*No*

Variable Importance Measures

Partial dependence plots can be used to visualise the relative importance of different variables. The partial dependence function at specified variable values (e.g. Total Positive Lymph nodes = 5 or Neoadjuvant Treatment = Chemoradiotherapy) is calculated as the average prediction of the RSF if all cases in the training set were changed to have this value. It therefore allows visualisation of the average marginal effect of individual variables and how changing them effects overall predictions, although does not assess how individual variables interact. Figure S.3 and Figure S.4 show this function for the variables included in the final model, ordered by importance. The differential survival can be seen to be far more important for the first 4 variables compared to the others.

The interpretation of this graph is complex. We note that cT and cN appear to have little effect on survival based on these measures. In reality we know they are important, and the graph reflects that the p/yp staging values are also included in the model.

Figure S.3 Partial Dependence Plots (1). (A) pN/ypN, (B)pT/ypT, (C) Circumferential Margin, (D) Age at diagnosis, (E) cT, (F) cN

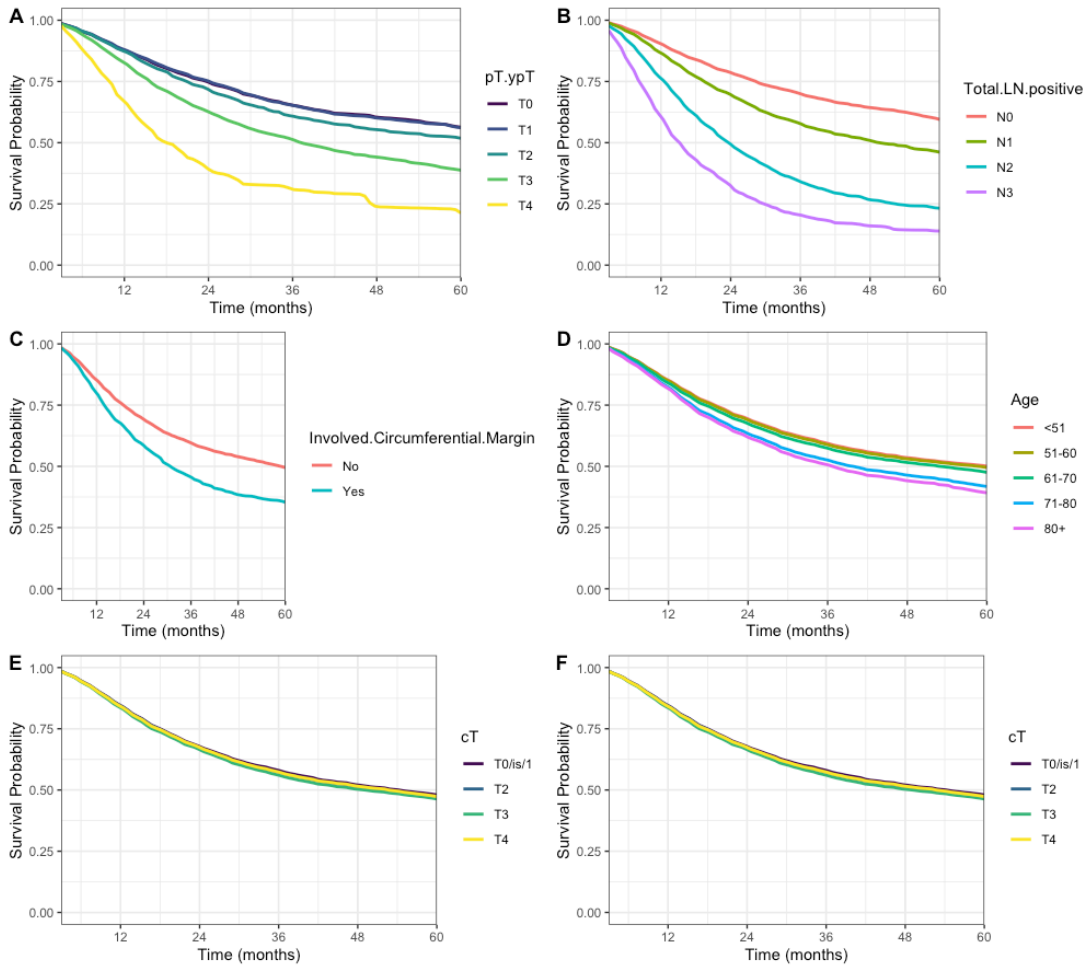
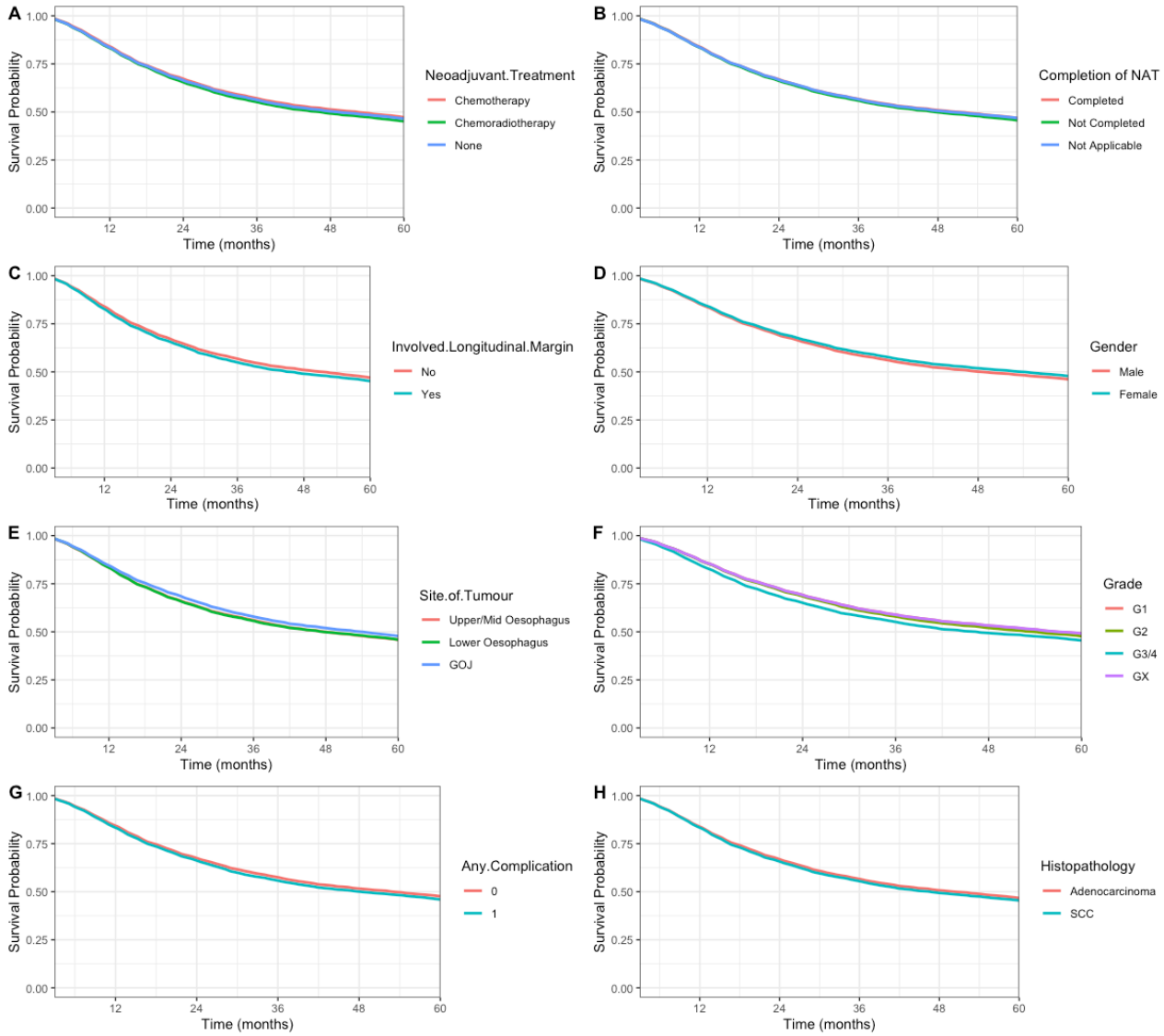


Figure S.4 Partial Dependence Plots(2). (A) Neoadjuvant Treatment Modality, (B) Completion of Neoadjuvant Treatment, (C) Grade of Differentiation, (D) Longitudinal Resection Margin,(E) Gender,(F) Site of Tumour,(G) Surgical Complications, (H) Histological Diagnosis



Cox Proportional Hazards Model (CPH)

A CPH was fitted to the variables used for the RSF model with no interactions or transformations. The proportional hazards assumption was assessed using Schoenfeld residuals. To assess for non-linearity of variables, the Martingale residuals of continuous variables separately entered into a null Cox model (Age and Number of Positive Lymph nodes) were calculated and a loess smoother applied. Both of these variables showed a clear non-linear form. A square root transformation improved the fit of positive lymph nodes but standard transformation (power or logarithmic) did not results in an improvement for age (more details available on request). A restricted cubic spline was therefore applied to age with 4 knots (ages 49, 62, 69 and 78). Table S.4 gives the final model coefficients (excluding Age, where the spline is represented in Figure S.5).

We examined the gain in performance for adding interaction terms between key variables. In brief we evaluated whether there was evidence of interactions between:

1. Age and pT stage
2. pT stage and number of positive nodes
3. Age and number of positive nodes
4. Histology, pT and number of positive nodes

These demonstrated only a marginal effect on performance of the model (C-index 0.751 without interactions, 0.749-0.752 with interactions) and were therefore not included. A benefit of this approach is that the cox regression model gives the reader information about the basic strengths of the relationships between the explanatory variables and survival.

Table S.4 Final CPH Coefficients

		Regression Coefficients			Proportional Hazards	
		Beta	Hazard Ratio (95% CI)	P value	Chi2	P value
Site of Tumour	Mid/Upper Oesophagus		1			
	Lower Oesophagus	-0.007	0.99 (0.87 - 1.14)	**0.916**	1.49	0.222
	GOJ	-0.152	0.86 (0.74 - 0.99)	**0.042***	0.24	0.623
Gender	Male		1			
	Female	-0.205	0.81 (0.73 - 0.91)	**<0.001***	0.12	0.733
Histopathology	Adenocarcinoma		1			
	SCC	0.216	1.23 (1.07 - 1.42)	**0.004***	3.85	0.05
cT	T0/1/is		1			
	T2	0.186	1.2 (0.94 - 1.55)	0.1443	0.34	0.559
	T3	0.213	1.24 (0.96 - 1.59)	0.0948	0.47	0.491
	T4	0.2	1.22 (0.9 - 1.66)	0.2019	0.23	0.628
cN	N0		1			
	N1	0.136	1.15 (1.04 - 1.26)	**0.006***	0.76	0.384
	N2	0.144	1.15 (1.02 - 1.31)	**0.022***	0.2	0.656
	N3	-0.015	0.98 (0.78 - 1.25)	0.899	0.24	0.624
Neoadjuvant Treatment	None		1			
	CRT - Completed	0.165	1.18 (0.92 - 1.51)	0.190	0.8	0.372
	CRT - Not Completed	0.888	2.43 (1.01 - 5.86)	**0.048***	0.09	0.765
	CT - Completed	-0.013	0.99 (0.88 - 1.11)	0.836	0	0.967
	CT - Not Completed	0.11	1.12 (0.93 - 1.34)	0.237	0.01	0.934
Any Complication	No		1			
	Yes	0.12	1.13 (1.04 - 1.22)	**0.004***	0.07	0.784
Square root Total LN Positive		0.453	1.57 (1.52 - 1.63)	**<0.001***	0.12	0.728
pT/ypT	T0		1			
	T1	0.205	1.23 (0.93 - 1.63)	0.1531	1.11	0.291
	T2	0.512	1.67 (1.27 - 2.19)	**<0.001***	0.83	0.363
	T3	0.93	2.53 (1.97 - 3.25)	**<0.001***	1.2	0.272
	T4	1.442	4.23 (3.17 - 5.64)	**<0.001***	2.16	0.141
Involved CRM	No		1			
	Yes	0.309	1.36 (1.24 - 1.49)	**<0.001***	1.32	0.251
Involved LRM	No		1			
	Yes	0.289	1.33 (1.11 - 1.61)	**0.003***	0.42	0.515
Differentiation Grade	Well		1			
	Moderate	0.315	1.37 (1.05 - 1.79)	**0.021***	0.3	0.584
	Poor/Anaplastic	0.453	1.57 (1.21 - 2.05)	**<0.001***	5.41	**0.020***
	GX	0.28	1.32 (0.99 - 1.76)	0.055	2.32	0.128

* = p<0.05

Figure S.5 Hazard of increasing age over time fitted with a restricted cubic spline. Knots are placed at age 49, 62, 69 and 78.

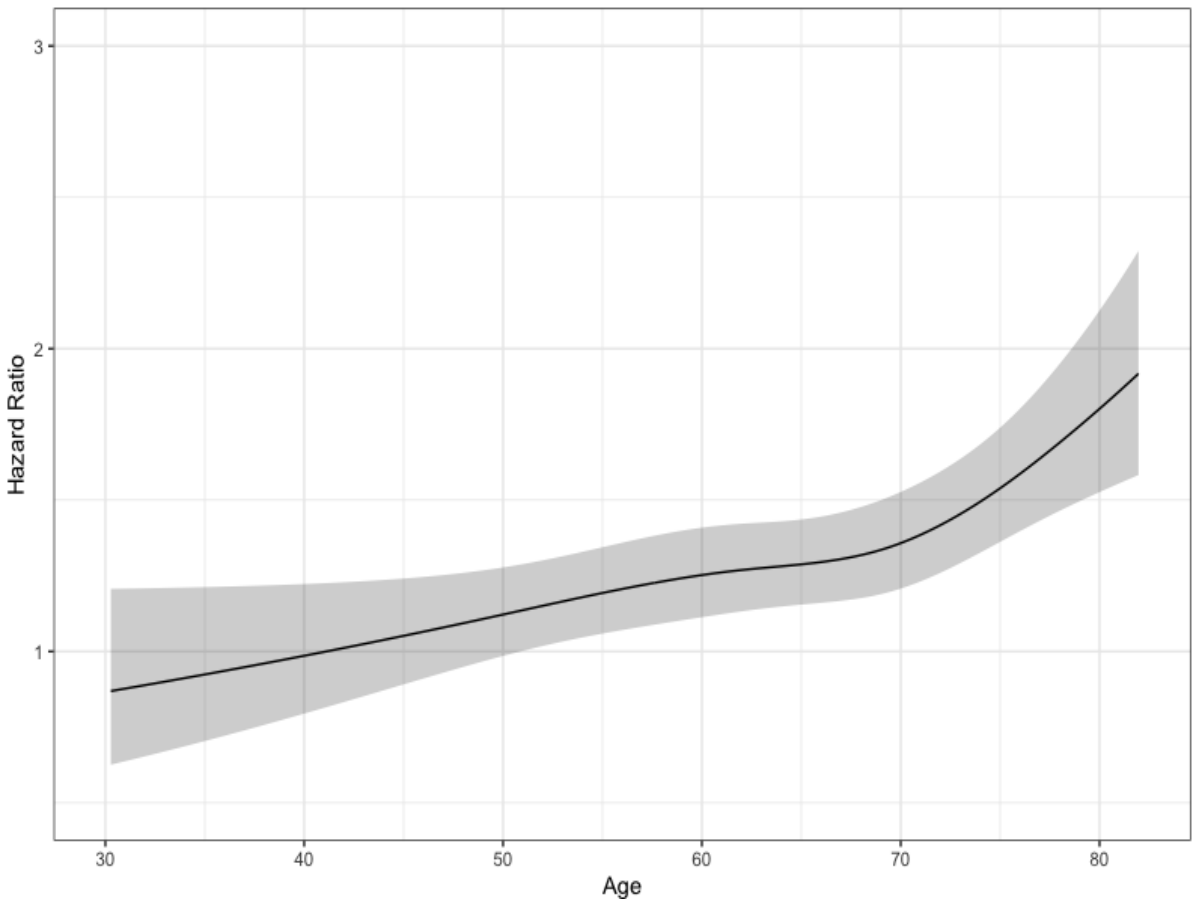


Figure S.6 Random Forest Model AUC across time points. Dashed lines represent 95% confidence intervals

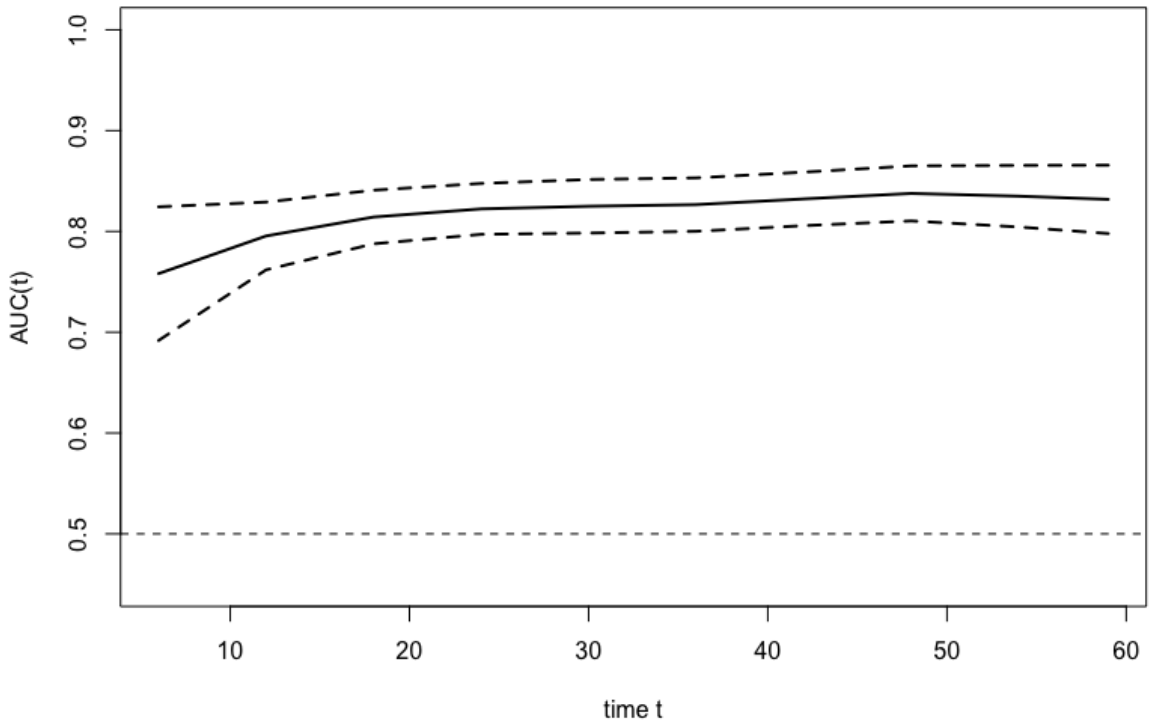


Figure S.7 Net benefit of the RSF model compared with the CPH model

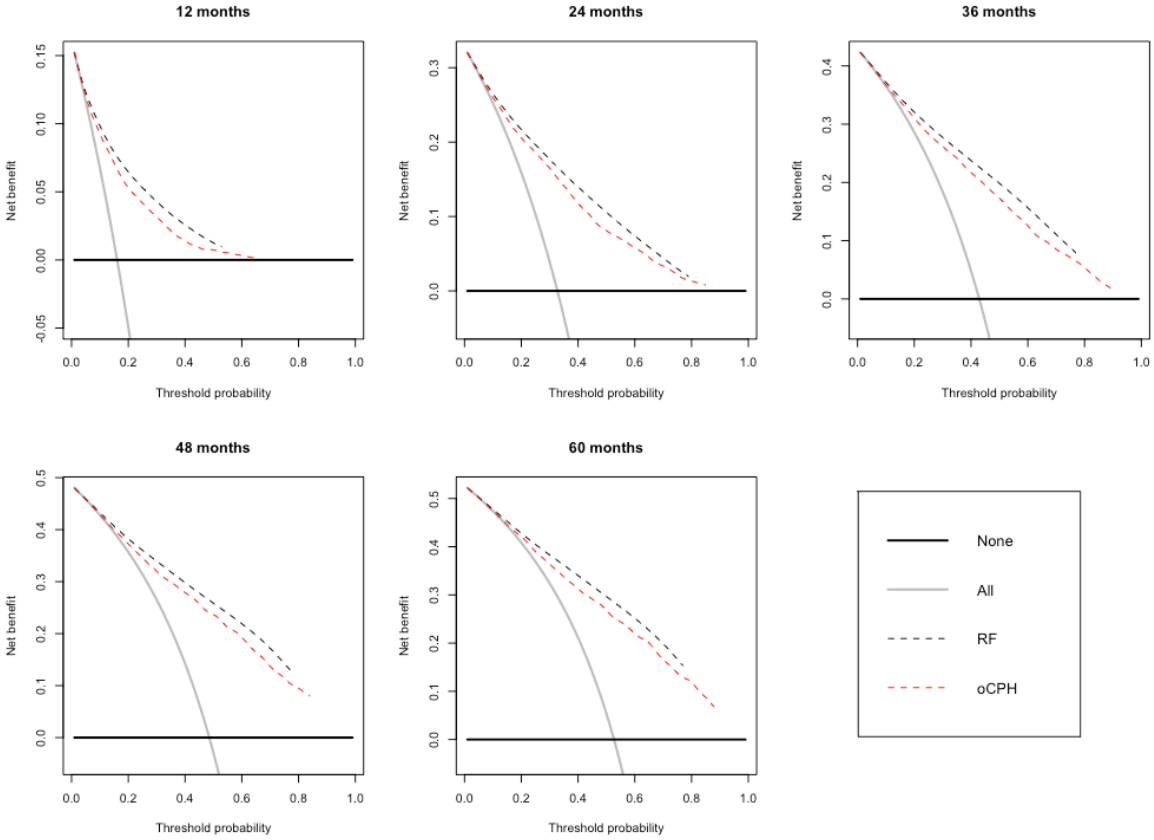


Figure S.8 Net benefit of the RSF model compared with TNM stage

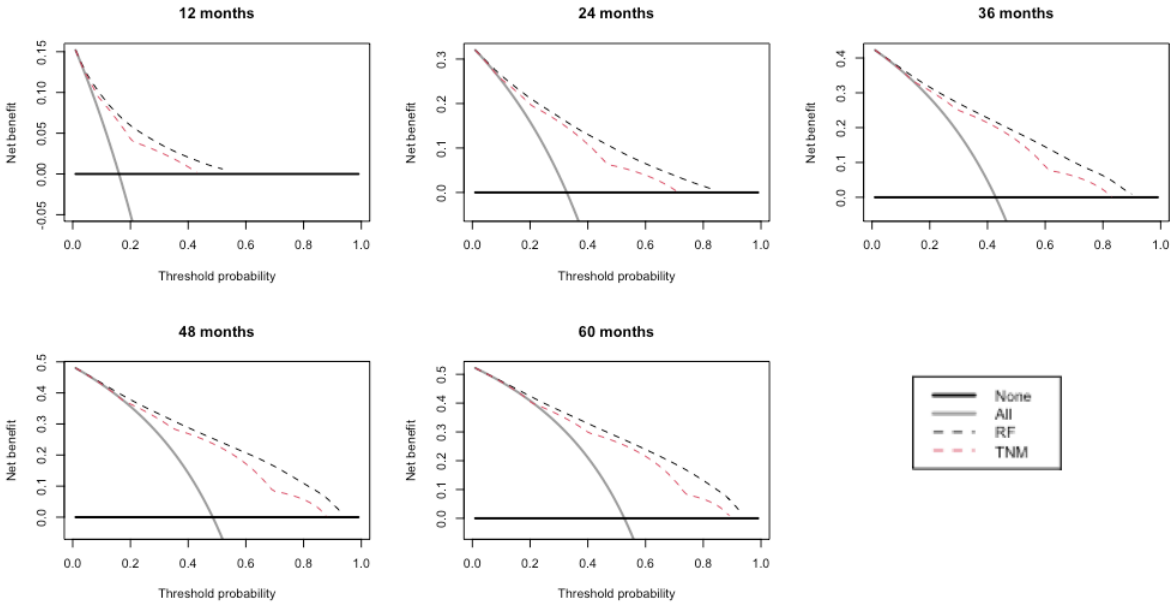
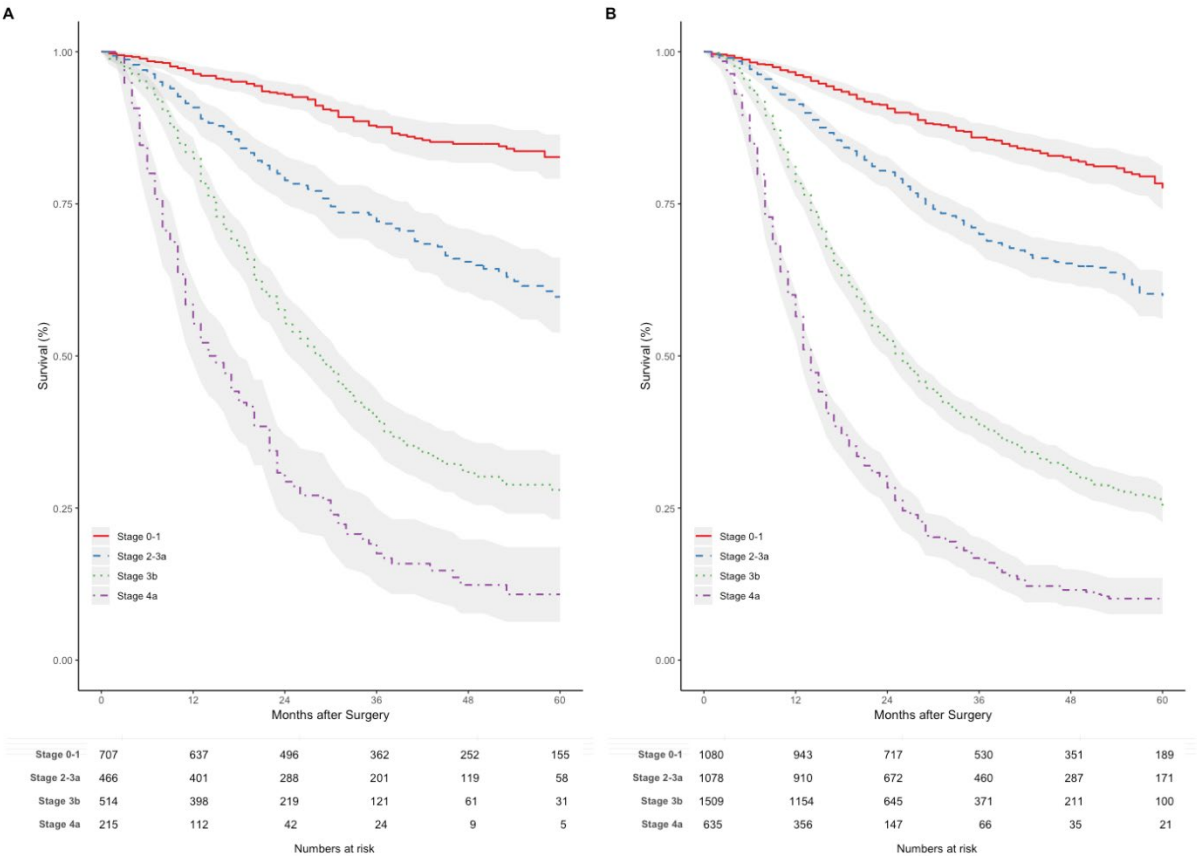


Figure S.9 Survival stratified by (A) pTNM stage and (B) ypTNM stage



Interestingly, we saw very little difference in outcomes between p and yp staging groups in comparison to that demonstrated in the 8th edition TNM data. It is not clear why this discrepancy is seen – but it may reflect a larger proportion of cases in the Worldwide Esophageal Cancer Collaboration (WECC) undergoing preoperative chemoradiotherapy, compared with a large majority of patients in our dataset who instead underwent perioperative chemotherapy. While it is true that p and yp staging differ it is probably also true that yp staging differs between modalities of treatment, and a smaller discrepancy between surgery-alone and chemotherapy treated patients would not be surprising.

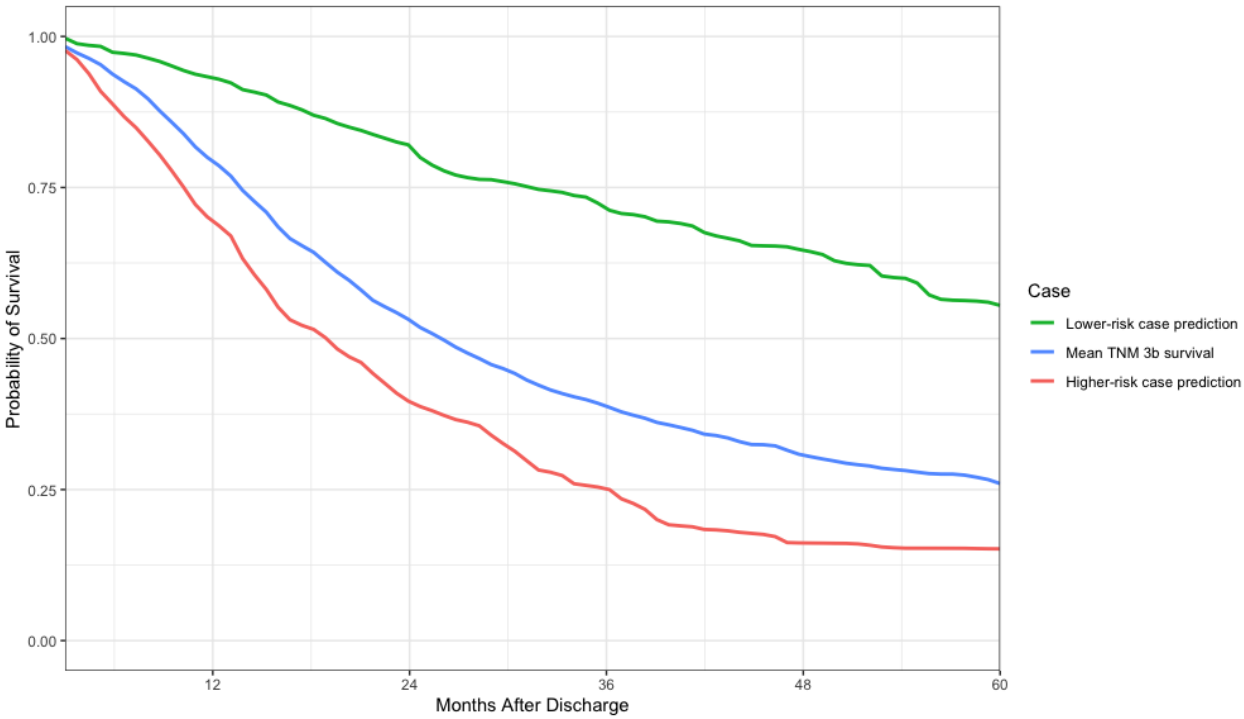
Examples of use

To illustrate how the model provides discrete predictions, two example cases are described below:

Case 1: 60-year-old male patient with adenocarcinoma of the oesophagus, who undergoes neoadjuvant chemotherapy before an oesophagectomy. His post-operative pathology reveals a well differentiated T3N1 (1/30) M0 tumour with no circumferential or longitudinal margin involvement. The ypTNM Stage is 3b.

Case 2: 60-year-old male patient with adenocarcinoma of the oesophagus, who undergoes neoadjuvant chemoradiotherapy which he fails to complete before an oesophagectomy. His post-operative pathology reveals a poorly differentiated T3N2 (6/30) M0 tumour with circumferential margin involvement. He also suffers from a post-operative complication. The ypTNM Stage is also 3b.

Figure S.10 Example Case Predicted Survival



A marked difference can easily be seen between these two cases, with a predicted 3-year survival of 55.5% and 15.1% respectively for two cases with identical ypTNM staging. The mean survival observed for TNM3b patients is 26.0%.

External Validation Instructions

A basic knowledge of R is required to conduct the external validation. As the model does not generate coefficients, access to the model itself is required.

First, download the file packet from the web application in the *'Model Details'* tab. This contains the models themselves and the manner in which dummy coding was conducted.

An example blank dataframe is also included showing the structure in which data must be presented to the model. Care should be taken to match the variables/names/factor-levels in this file. If the model fails to generate predictions, it is probably due to a discrepancy here.

Then access and download the R script from github:

https://github.com/saqibrahmanUGI/AUGIS-Surv

Running this script will firstly install and load the needed R packages, then batch generate predictions, calculate the tAUC and C-index, plot annual calibration curves and plot quintiles of prediction against observed KM estimates.

Supplementary References

1. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.
2. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681–694.
3. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28:1982–1998.
4. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–399.
5. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–1101.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;339(7713):157–160.
7. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. BMC Med Res Methodol. 2009;9(1):1–8.
8. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biometrical J. 2015 Jul;57(4):614–632.
9. Hosmer D, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time to Event Data. 1st ed. New York: John Wiley & Sons, Inc.; 1999.
10. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013 Dec 30;32(30):5381–5397.
11. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: Current methods and applications. BMC Med Res Methodol. 2017;17(1):1–19.
12. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–574.
13. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:1–17.
14. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. Med Decis Mak. 2008;28(1):146–149.
15. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Vol. 35, European Heart Journal. Oxford University Press; 2014. p. 1925–1931.