European Journal of Surgical Oncology xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Surgical Oncology

journal homepage: www.ejso.com

Machine learning to predict curative multidisciplinary team treatment decisions in oesophageal cancer

Navamayooran Thavanesan [a,*], Indu Bodala [b], Zoë Walters [a], Sarvapali Ramchurn [b], Timothy J. Underwood [a,1], Ganesh Vigneswaran [a,1]

[a] School of Cancer Sciences, Faculty of Medicine, University of Southampton, UK
[b] School of Electronics and Computer Science, University of Southampton, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Oesophageal cancer multidisciplinary team

ABSTRACT

Background: Rising workflow pressures within the oesophageal cancer (OC) multidisciplinary team (MDT) can lead to variability in decision-making, and health inequality. Machine learning (ML) offers a potential automated data-driven approach to address inconsistency and standardize care. The aim of this experimental pilot study was to develop ML models able to predict curative OC MDT treatment decisions and determine the relative importance of underlying decision-critical variables.

Methods: Retrospective complete-case analysis of oesophagectomy patients ± neoadjuvant chemotherapy (NACT) or chemoradiotherapy (NACRT) between 2010 and 2020. Established ML algorithms (Multinomial Logistic regression (MLR), Random Forests (RF), Extreme Gradient Boosting (XGB)) and Decision Tree (DT) were used to train models predicting OC MDT treatment decisions: surgery (S), NACT + S or NACRT + S. Performance metrics included Area Under the Curve (AUC), Accuracy, Kappa, LogLoss, F1 and Precision -Recall AUC. Variable importance was calculated for each model.

Results: We identified 399 cases with a male-to-female ratio of 3.6:1 and median age of 66.1yrs (range 32–83). MLR outperformed RF, XGB and DT across performance metrics (mean AUC of 0.793 [±0.045] vs 0.757 [±0.068], 0.740 [±0.042], and 0.709 [±0.021] respectively). Variable importance analysis identified age as a major factor in the decision to offer surgery alone or NACT + S across models ($p < 0.05$).

Conclusions: ML techniques can use limited feature-sets to predict curative UGI MDT treatment decisions. Explainable Artificial Intelligence methods provide insight into decision-critical variables, highlighting underlying subconscious biases in cancer care decision-making. Such models may allow prioritization of caseload, improve efficiency, and offer data-driven decision-assistance to MDTs in the future.

1. Introduction

Oesophageal cancer (OC) is a devastating condition. Despite improving survival rates, it remains 7th in worldwide incidence and the 7th most common cause of cancer death [1,2]. Treatment decisions for OC cancer patients in the UK are managed by multidisciplinary teams (MDT) integrating healthcare expertise for shared decision-making [3]. Decisions are driven by tumour features (size, location, spread), as well as patient factors (fitness for surgery, co-morbidities and demographics), which may impact tolerability of therapy [4]. OC treatment decisions thus carry implications for patient quality of life [5]. OC MDTs however have been shown to reduce the incidence of open-and-close surgeries, reduce operative mortality, increase rates of completed staging and are an independent positive predictor for survival in OC [3,6–8].

MDTs are inherently informed by individual experience, perception and bias. Additionally, multiple clinical and human factors such as case complexity, increasing caseload, individual clinician preference or even seniority can lead to unexplained variability or suboptimal decision-making [9,10]. One Danish study reported clinical impact in as many as 60% of test cases on subsequent management because of MDT disagreement [11].

Predictive modelling to assist decision-making for OC patients has demonstrated excellent results when predicting survival post-surgery in OC patients [12,13]. These studies have generally accessed both pre-

* Corresponding author. South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, SO16 6YD, UK.
E-mail address: N.Thavanesan@soton.ac.uk (N. Thavanesan).
[1] These authors are Joint Last Author for this manuscript.

https://doi.org/10.1016/j.ejso.2023.106986
Received 26 April 2023; Received in revised form 22 June 2023; Accepted 11 July 2023
Available online 13 July 2023

and post-operative data to train such models. At the point of first diagnosis however, the MDT must act on a relatively restricted pool of information, a scenario in which Machine Learning (ML) modelling techniques may offer significant benefit especially if able to pair MDT decisions with data-driven evaluation [14,15]. Accurate predictive models would provide for consistent clinical assistive decision tools (CADT) capable of standardizing such decisions, improving efficiency, and positively impacting healthcare equality.

The aim of this pilot study was to explore whether an accurate ML model for predicting which curative patients will receive neoadjuvant chemotherapy (NACT), neoadjuvant chemoradiotherapy (NACRT) or proceed straight to surgery could be created using a limited pool of variables available to a single-centre OC MDT at the time of deciding a patient's final curative treatment pathway. Secondary aims included comparison of ML algorithmic performance and investigation of variable importance in order to provide model explainability within OC decision-making.

2. Methods

This study was a retrospective complete-case analysis of potentially curative oesophageal cancer patients at a single tertiary referral centre (University Hospital Southampton) under the ethical approval of IRAS 233065.

2.1. Study cohort

All patients who underwent an oesophagectomy for OAC or OSCC from 2010 to 2020 were identified from a prospectively maintained oesophagectomy database. This proof-of-principle pilot study focussed on curative patients because reliable high-quality data was available for this cohort. Treatment decisions at our institution were made as per National Institute for Clinical Excellence (NICE) guidelines [16]. Patients underwent either NACT or NACRT (prior to surgery) or proceeded directly to surgery. Variables consistently available to the MDT prior to a final treatment decision were included within the models (Table 1). This is more reflective of "real world" scenarios where the quality and quantity of such data can often vary. Clinical staging was assessed on baseline imaging (Computer Tomography (CT) and/or Positron Emission Tomography (PET)) and tissue biopsies in accordance with the American Joint Committee on Cancer (AJCC) Tumour-Node-Metastasis (TNM) staging system.

Table 1
Patient demographics and model predictor variables by sub-group (sub-group comparison of continuous variables by Kruskal-Wallis analysis and categorical variables by Chi-Squared test of independence).

Pre-treatment variables	NACT (N = 172) (%)	NACRT (N = 127) (%)	Surgery (N = 100) (%)	Total (N = 399) (%)	P Value
Gender					0.016*
Male	146 (84.9%)	91 (71.7%)	75 (75%)	312 (78.2%)	
Female	26 (15.1%)	36 (28.3%)	25 (25%)	87 (21.8%)	
Median Age in years (Range)	65.1 (32.4–81.8)	65.9 (40.5–79.0)	72.6 (33.7–83)	66.1 (32.4–83.00)	<0.001
Performance status					<0.001***
0	87 (50.6%)	83 (65.3%)	33 (33%)	203 (50.9%)	
1	80 (46.5%)	41 (32.3%)	56 (56%)	177 (44.3%)	
2	5 (2.9%)	3 (2.4%)	11 (11%)	19 (4.8%)	
3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
ASA grade					0.017*
1	10 (5.8%)	9 (7.1%)	7 (7%)	26 (6.5%)	
2	107 (62.2%)	89 (70.1%)	49 (49%)	245 (61.4%)	
3	55 (32.0%)	29 (22.8%)	44 (44%)	128 (32.1%)	
4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
cT stage					<0.001***
0	1 (0.6%)	0 (0%)	8 (8%)	9 (2.3%)	
1	0 (0%)	0 (0%)	6 (6%)	6 (1.5%)	
2	30 (17.4%)	24 (18.9%)	46 (46%)	100 (25.1%)	
3	124 (72.1%)	91 (71.7%)	38 (38%)	253 (63.4%)	
4	17 (9.9%)	12 (9.4%)	2 (2%)	31 (7.7%)	
cN stage					<0.001***
0	34 (19.8%)	28 (22.0%)	55 (55%)	117 (29.3%)	
1	120 (69.8%)	83 (65.4%)	40 (40%)	243 (60.9%)	
2	18 (10.4%)	16 (12.6%)	4 (4%)	38 (9.5%)	
3	0 (0%)	0 (0%)	1 (1%)	1 (0.3%)	
Tumour location					<0.001***
Oesophagus	36 (20.9%)	62 (48.8%)	25 (25%)	123 (30.8%)	
GOJ	136 (79.1%)	65 (51.2%)	75 (75%)	276 (69.2%)	
Tumour Histology					<0.001***
Adenocarcinoma	159 (92.4%)	83 (65.4%)	91 (91%)	333 (83.5%)	
Squamous Cell	13 (7.6%)	44 (34.6%)	9 (9%)	66 (16.5%)	
Co-morbidities					
History of MI (MI)	9 (5.2%)	6 (4.7%)	9 (9%)	24 (6.0%)	0.344
Chronic heart failure (CHF)	1 (0.6%)	0 (0%)	2 (2%)	3 (0.8%)	0.211
Chronic pulmonary disease (CPD)	25 (14.5%)	14 (11.0%)	19 (19%)	58 (14.5%)	0.239
Connective tissue disease	2 (1.2%)	5 (3.9%)	1 (1%)	8 (2.0%)	0.170
Peripheral vascular disease (PVD)	2 (1.2%)	0 (0%)	4 (4%)	6 (1.5%)	0.043*
Cerebrovascular disease (CVD)	6 (3.6%)	3 (2.4%)	8 (8%)	17 (4.3%)	0.091
History of Peptic Ulcer Disease (XPUD)	6 (3.6%)	2 (1.6%)	5 (5%)	17 (4.3%)	0.344
Uncomplicated diabetes (DM uncomp)	17 (9.9%)	13 (10.2%)	16 (16%)	46 (11.5%)	0.269
Complicated diabetes (DM comp)	0 (0%)	0 (0%)	1 (1%)	1 (0.3%)	0.223
Leukaemia	0 (0%)	0 (0%)	3 (3%)	3 (0.8%)	0.011*
Lymphoma	1 (0.6%)	1 (0.8%)	3 (3%)	5 (1.3%)	0.191
Mild liver disease	0 (0%)	0 (0%)	0 (0%)	2 (0.5%)	0.265

2.2. Model development

2.2.1. Data preparation and analysis

Data analyses were conducted using RStudio (Version 4.1.2) with relevant packages described where first used. The choice of final treatment pathway was assigned as the outcome variable: Surgery (S), (NACT + S), or (NACRT + S). Cases with missing data were removed for the purposes of complete-case analysis. The final dataset contained a total of 399 complete cases (Table 1).

2.2.2. Machine learning algorithms

Four established ML algorithms were selected and implemented via the "caret" package; Multinomial Logistic Regression (MLR) [17], Random Forests (RF) [18], Extreme Gradient Boost (XGB) [19] and Decision Tree (DT) analysis [20]. The MLR model was trained using the "nnet" package extension with L2 regularisation. The RF model was trained using the "randomForest" package extension. The XGB model was trained using the "xgboost" package extension. Decision Trees were trained using the "rpart" package. This provided diversity of ML techniques (regression-based, tree-based and ensemble).

2.2.3. Validation and model performance

All models were developed using nested cross-validation (CV) and optimised for accuracy. A 5 × 10 configuration was chosen (10-fold CV within the inner loop with 5-fold outer loop). The ROC values for the best model from each outer fold (N = 5) were then averaged to generate a mean Area Under the Receiver Operator Characteristic curve (AUROC) in a one-versus-others approach. This provided a more accurate estimate of overall model generalisability at differing probability thresholds. Each ROC curve was plotted with confidence intervals of 1x Standard Error of the Mean (SEM). Mean out-of-sample predictive performance was also compared between algorithms for balanced accuracy, mean AUC, Kappa, Log Loss, F1 and precision-recall AUC (PRAUC) using the resamples () function (caret package).

2.2.4. Variable importance analysis

Variable importance was derived for each algorithm to examine, quantify and rank overall importance a given feature provided to the final models. This provided insight into variables contributing most significantly to current OC MDT treatment decisions. Variable importance was calculated using the varImp () function (caret package) for MLR, RF and DT, and the xgb.importance () function (xgboost package) for the XGBoost model. Absolute values were scaled (0–100) to allow comparison between algorithms.

2.2.5. Inter-algorithmic and inter-class predictive performance

For meaningful statistical comparison of AUROCs produced for each algorithm all algorithms were further re-trained total of 10 times, (now producing a total of 50 "outer-fold" models). In each repeat the set-seed was randomized, and the resulting 50 AUROCs were analysed using the Kruskal – Wallis test coupled with the Pairwise Wilcoxon Rank Sum Test where appropriate (p values were adjusted using the Benjamini-Hochberg correction, ($p < 0.05$ was deemed significant)). This allowed robust comparison of differences in predictive performance across algorithms for a specific outcome class as well as a comparison of all outcome classes from a given algorithm.

3. Results

3.1. Cohort demographics

A total of 436 cases were identified, with 5 cases excluded for missing data (Complicated Diabetes (N = 2), cN stage (N = 2) and Tumour location (N = 1)) and 32 cases excluded for ineligible histology. This produced a final cohort of 399 cases.

3.2. Algorithm performance

Predictive performance for each algorithm was assessed on mean-model performance and individualised outcome-class prediction. All algorithms produced models which performed above random chance (AUROC = 0.5). At class-level, all algorithms performed best when predicting patients likely to be offered surgery (MLR 0.865, RF 0.859, XGB 0.805, DT 0.802). All algorithms perform less confidently in predicting NACRT + S (MLR 0.772, RF 0.699, XGB 0.696, DT 0.651) and NACT + S (MLR 0.704, RF 0.651, XGB 0.644, DT 0.704). Individual ROC curves for each algorithm are illustrated in Fig. 1 (additional ROC curves for models trained solely on adenocarcinoma are in Supplemental Fig. 1).

3.3. Comparison of algorithms

Repeated, nested-CV was used to assess for statistical differences in AUROC between algorithms (Supplemental Table 1). MLR outperformed RF and XGB on Kruskal-Wallis analysis when predicting NACT + S ($P = <0.001$) and NACRT + S ($P < 0.001$) but comparably with DT (Pairwise Wilcoxon Rank Sum test, $P = 0.143$). MLR also outperformed XGB and DT, and comparably to RF when predicting surgery (Pairwise Wilcoxon Rank Sum test, $P = 0.001$, $P < 0.001$ and $P = 0.134$ respectively). On mean-model out-of-sample predictive performance MLR performed best across all performance metrics (Table 2). RF and XGB performed comparably on balanced accuracy (0.679 vs 0.698 respectively), mean AUC (0.757 vs 0.740), mean F1 (0.575 vs 0.607), mean PRAUC (0.560 vs 0.544) and mean kappa (0.352 vs 0.386). XGB was outperformed by MLR, RF and DT on mean LogLoss (1.360 vs 0.833, 0.942 and 1.146 respectively).

3.4. Inter-class performance

Statistical difference between outcome-class prediction was assessed for each algorithm to determine if overall model performance was weighted towards a given treatment decision. A significant difference was demonstrated on Kruskal-Wallis and Pairwise Wilcoxon Rank Sum test for all classes (Supplemental Table 2).

3.5. Variable importance

Variable importance analysis highlighted factors critical to model formation (Fig. 2). The MLR model highlighted cT stage as most important, but with more salience attributed to co-morbidities such as connective tissue disease, lymphoma, leukaemia, and liver disease. Within tree-based models (RF, XGB and DT) the single most influential variable was age (scaled importance = 100%). DT analysis delineated an age cut-off of 77yrs as key within the decision-making pathway (Supplemental Fig. 2). Across models, factors such as tumour histology, tumour location, cT stage, cN stage, and performance status remained important contributors to the final models (this was consistent even when trained solely on adenocarcinoma patients).

3.6. Role of age in predicting treatment decisions

As age emerged as the most important variable in RF, XGB and DT models, all algorithms were retrained without age to assess its overall significance by examining the effect its removal produced on mean-model AUROC (Fig. 3).

Differences in AUROC for all algorithms ± age were then compared statistically (Kruskal-Wallis test, P values provided in Fig. 3). Across all algorithms, the removal of age produced a significant drop in mean AUROC when predicting a surgery treatment decision (MLR 0.858 vs 0.835 ($P = 0.017$), RF 0.846 vs 0.785 ($P < 0.001$), XGB 0.828 vs 0.781 ($P < 0.001$)), DT 0.747 vs 0.682 ($P < 0.001$). This was again seen in the decision to offer NACT + S for RF and XGB models (RF 0.676 vs 0.647 (P

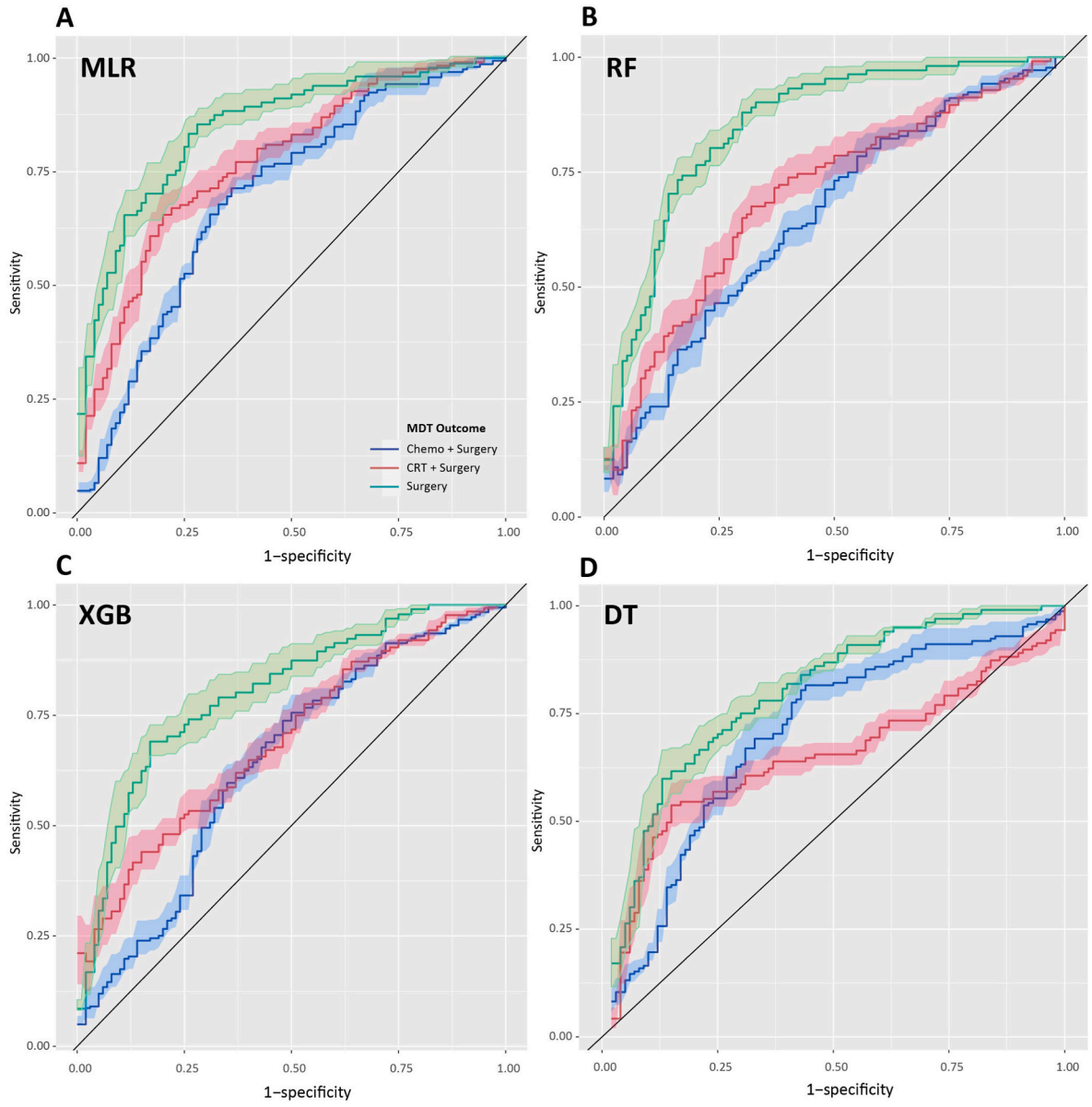

Fig. 1. ROC curve for averaged nested, cross-validated model performance given with ± 1x standard error of the mean (SEM), A = Multinomial Logistic Regression, B = Random Forests, C = Extreme Gradient Boost and D = Decision Tree. AUROC = Area under Receiver Operator Characteristic.

Table 2
Mean performance metrics by algorithm (best performance metric in bold). Abbreviations – sd = Standard Deviation, AUC = Area Under Curve, PRAUC = Precision Recall AUC.

	Mean Balanced Accuracy (±sd)	Mean AUC (±sd)	Mean Kappa (±sd)	Mean LogLoss (±sd)	Mean F1 (±sd)	Mean PRAUC (±sd)
MLR	**0.718 ± 0.066**	**0.793 ± 0.045**	**0.428 ± 0.127**	**0.833 ± 0.080**	**0.624 ± 0.083**	**0.594 ± 0.066**
RF	0.679 ± 0.075	0.757 ± 0.068	0.352 ± 0.155	0.942 ± 0.160	0.575 ± 0.101	0.560 ± 0.073
XGB	0.698 ± 0.050	0.740 ± 0.042	0.386 ± 0.101	1.360 ± 0.235	0.607 ± 0.062	0.544 ± 0.052
DT	0.676 ± 0.027	0.709 ± 0.021	0.347 ± 0.012	1.146 ± 0.110	0.564 ± 0.038	0.365 ± 0.025

= 0.005), XGB 0.666 vs 0.619 (P < 0.001)) with a non-significant drop noted for MLR (0.710 vs 0.692, P = 0.065) and DT models (0.688 vs 0.670, P = 0.212). Removing age did not impact prediction of NACRT + S regardless of algorithm (MLR 0.778 vs 0.774 (P = 0.710), RF 0.714 vs 0.711 (P = 0.679), XGB 0.710 vs 0.707 (P = 0.767)), DT 0.647 vs 0.687 (P = 0.002). ROC plots for each algorithm and outcome class are provided in Supplemental Fig. 3. This pattern continued to hold when models were limited to adenocarcinoma patients with significant drops in AUC seen in both NACT + S (P values: MLR 0.034, RF 0.003, XGB 0.004, DT < 0.001) and Surgery prediction (P values: MLR 0.025, RF < 0.001, XGB <0.001, DT < 0.001) while CRT remains largely unaffected (P values: MLR 0.389, RF 0.393, XGB 0.577, DT 0.033).

4. Discussion

We have demonstrated feasibility for ML models to predict curative OC MDT treatment decisions with limited feature-sets. Importantly, these algorithms are computationally inexpensive as any real-world clinical assistive decision tool (CADT) needs to operate within current

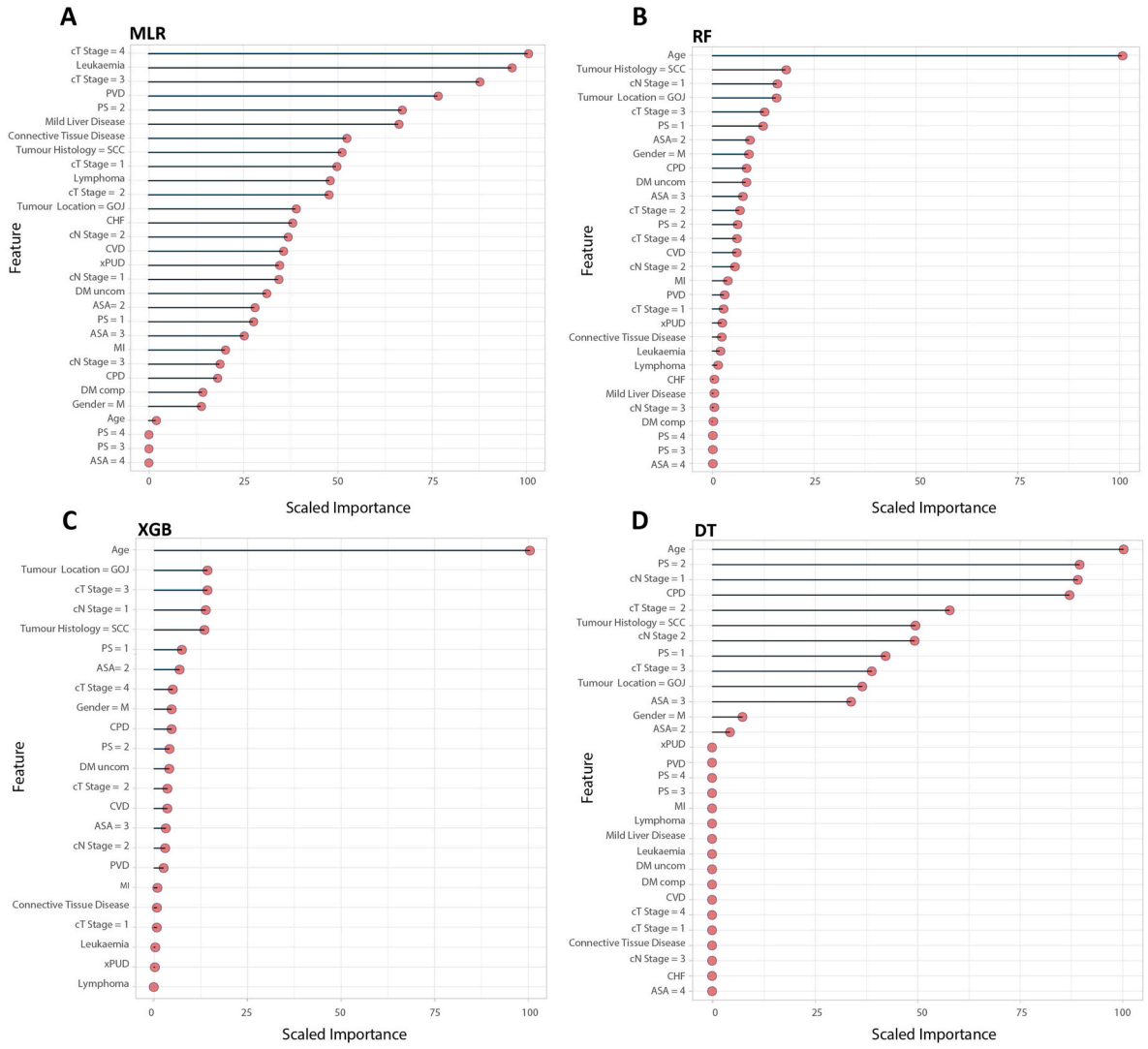

Fig. 2. Scaled Variable Importance plots for nested cross-validated models. "Y" indicates the presence of the stated co-morbidity. A = Multinomial Logistic Regression, B = Random Forests, C = Extreme Gradient Boost and D = Decision Tree. (Abbreviations: PVD = Peripheral Vascular Disease, SCC = Squamous Cell Carcinoma, GOJ = Gastroesophageal Junction, CHF = Congestive Heart Failure, CVD = Cerebrovascular Disease, xPUD = History of Peptic Ulcer Disease, DM uncom = Uncomplicated Diabetes Mellitus, DM comp = Diabetes Mellitus with complications, Gender (M = Male/F = Female), PS = Performance Status, ASA = American Society of Anaesthesiologist score, CPD = Chronic Pulmonary Disease, MI = History of Myocardial Infarction).

electronic healthcare infrastructure. While MLR performed best, all models demonstrated good AUROCs and were confident discriminating between patients recommended surgery versus those offered NAT across a mixed histology cohort (while this remained so when trained on adenocarcinoma alone, the best performances were achieved with the full cohort indicating a machine-preference for learning from both subtypes). While performance was attenuated when predicting a specific NAT subtype, all algorithms performed well above random chance. Variable importance analysis offered insight into the critical variables underpinning these models, identifying age to be most significant to all tree-based models, and to a lesser extent, with MLR. When age was removed from the feature-set, all algorithms suffered a reduction in predictive performance for surgery or NACT + S though the decision to offer NACRT + S appeared unaffected by age. DT analysis highlighted an age cut-off of 77 years to be significant with those older, more likely to proceed to surgery.

The consistency in ROC curves across algorithms, irrespective of design likely reflects an underlying pattern within the OC patient cohort itself and is readily observed in the prediction of NACT versus NACRT. Evidence for the survival benefit of NAT in locally advanced OC is well established [21–24]. The superior NAT modality (for adenocarcinoma) remains unknown. Recent 3-year follow-up data from the NeoAegis trial remains equivocal on survival outcomes despite a higher incidence of patients with a good primary tumour response to treatment (TRG 1–3) in the NACRT arm [25]. It is reasonable to infer that while clinical equipoise remains within the field, these ML models mirror a similar uncertainty within the MDT. The benefit of explainable ML approaches is therefore in offering valuable insight into both the human decision-making at play as well as areas of uncertainty which may propagate inconsistent decisions within the MDT.

The contribution of individual variables to our OC MDT ML models is a key aspect of this study. It has been postulated previously that some factors (biases) inherent to MDT decisions may not be consistently or explicitly reflected in that decision-making and by extension into current models [26]. Significant importance was unsurprisingly assigned to T-stage, N-stage, performance status, tumour histology and tumour location in all models. Co-morbidities such as chronic pulmonary disease and diabetes ranked higher within tree-based models, while haematological cancers, connective tissue disease and liver dysfunction were more relevant to regression models. This demonstrates how

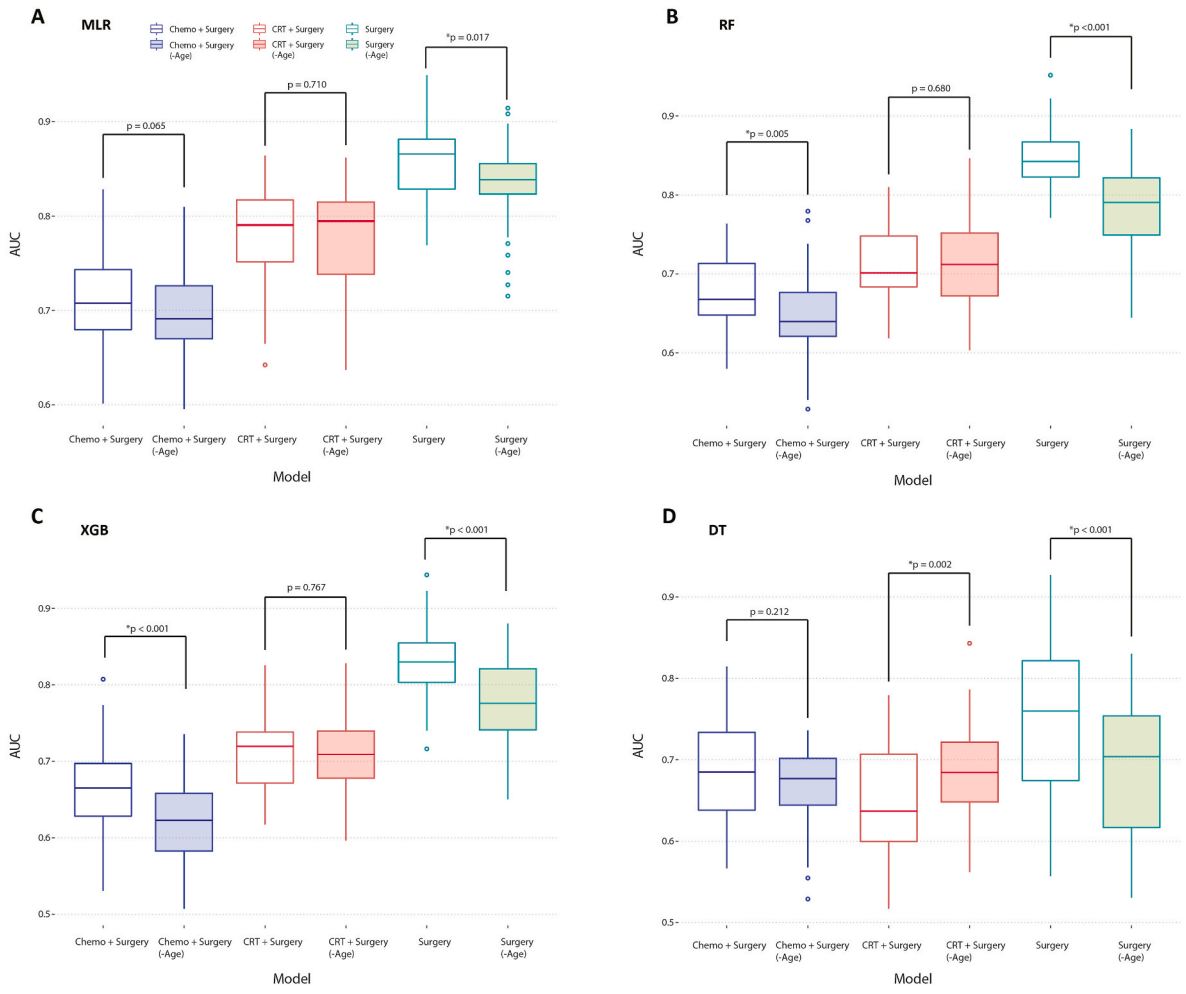

Fig. 3. Boxplot comparison of mean model AUCs for MLR (A), RF (B), XGB (C) and DT (D) models with and without Age. Significant P values denoted with and asterisk.

incorporating co-morbidities into models can reflect intuitive human decision-making. Most interesting proved the importance contributed by age in RF, XGB and DT models where its removal provoked a significant drop in performance when predicting surgery and NACT + S. Historically, clinician bias in cancer management for elderly patients led to the UK Department of Health initiative in 2012 to drive personalised treatment decisions based on physiological age over chronological age [27,28]. Within our cohort a higher median age was seen in patients offered surgery versus any NAT, and DT analysis suggests an important cut-off at 77 years. This may be explained by the well-recognised risk of deconditioning frail patients after NAT and potentially rendering them unfit for surgery [29]. A single attempt may be their only chance at cure which NAT may compromise. It is less apparent why CRT prediction was unperturbed by age and may reflect the broadly held opinion that pre-operative CRT (CROSS-style) for OC is less toxic and less debilitating versus modern chemotherapy regimens (e.g., FLOT). While median age in both NACT + S and NACRT + S groups were comparable, a higher proportion of NACRT + S patients presented with robust performance status scores when compared with NACT + S patients. In the context of an already physiologically fitter cohort, chronological age may prove less influential in their resilience for multimodal NAT. While it is tempting to assume chronological age is not an automatic blockade to aggressive treatments, ML lets us challenge such pre-conceived notions by highlighting hidden patterns within MDT decision-data. In characterising these patterns, we learn about potential subconscious biases in decision-making and address any inequality that may result.

Acceptability and explainability of CADTs is a major consideration in the integration AI-based tools within healthcare where regulatory approval will almost certainly hinge upon explainable and interpretable solutions [30]. This is problematic for deep-learning platforms which are inherently "black-box" solutions [31]. MLR performed best in this study and is the most explainable. Decision-trees are also members of explainable AI (XAI) approaches, however, once the model training involves many hundreds of trees (RF and XGB-models) explainability becomes challenging, requiring post-hoc explainability methods [32]. Simple visual analysis of the scaled variable importance plots in Fig. 2 might lead treating clinicians towards a tree-based model, as the ordering of listed variables fits the intuitive assessment of patients made on a day-to-day basis in the clinic. However, as MLR outperformed tree-based models it also highlights the pragmatic need to balance performance against ease of explainability and acceptability to the end user.

The long-term clinical implications of this study are most likely to relate to health economy (via streamlining of future MDTs which may increase caseload efficiency and staffing costs) and health equality (by standardizing decision-making for cases with comparable demographics and disease staging). At present nuanced treatment decisions such as surgical approach are influenced by tumour characteristics combined with surgeon preference and experience. Observational evidence for minimally invasive surgery favoured improved rates of post-operative pneumonia and recovery times although formal trials such as the Traditionally Invasive versus Minimally invasive Oesophagectomy (TIME) and MIRO trials showed equivalence in survival benefits

compared to open resection [33–37]. Early Indications from the Randomised Oesophagectomy: Minimally Invasive or Open (ROMIO) study [38] also appear to reiterate comparable recovery and complication rates although a formal report is awaited. While robotic oesophagectomy offers greater surgeon ergonomics and stereoscopic visualisation, a growing evidence base for reduced pulmonary complications must be offset against longer operative time and resource-costs for otherwise comparable patient outcomes [39,40]. In all scenarios such treatment decisions are driven heavily by perceived post-operative outcomes over pre-treatment clinicopathological characteristics. Modelling such decisions at a pre-treatment time-point thus poses significant challenges such as sensitive surgeon-specific data on operative experience and preference which in turn risks its own ethical concerns. In the interim, broader treatment recommendations by a CADT however remains feasible and preserve MDT nuance.

4.1. Study limitations and future directions

There are natural limitations to this pilot study. Despite a cohort encompassing approximately 10 years within a tertiary referral centre, our final dataset comprised 399 patients. By utilising supervised-learning techniques which tolerate smaller datasets in conjunction with nested cross-validation we attenuated the generalisability error within our models. The predictor variables selected were, by design, limited to those the MDT could reasonably consider at the time of a final treatment decision, with limited granularity in this pilot study. However, these models do not presently incorporate visual data (radiological and histopathological imaging), nor key social/human factors (the last of which, previous studies have found inconsistent in MDT environments) [9,10]. The authors additionally recognise that OC management underwent shifts in oncological practice over the study period, however this was primarily focussed on specific adjunctive therapeutic regimens, and changes in surgical approaches as opposed to specific indications for a given treatment category. While it is also likely that clinician preferences and human factors are relevant to these decisions, such data is not routinely recorded and a more simplified proof-of-concept was pursued in this instance to ensure model feasibility.

Nevertheless, we have shown that ML models can use even limited feature-sets to produce good predictive models offering proof-of-principle of ML-based CADTs. This offers future potential for applying semi-automated tools to improve workload and efficiency. Such tools may run in parallel with MDTs to provide data-driven recommendations for complex patients, provide a means to sense-check decisions and offer assessments unaffected by natural variation over time in MDT attendees.

Future models will need to integrate variables such as lifestyle risk factors, BMI, shifts in oncological practice (e.g., NACT regimens or TNM classification updates) and even the geographical distribution of patients relative to chemotherapy and chemoradiotherapy units. Features can be expanded to include more detailed tumour geography, tumour size, tumour differentiation, and molecular classification of histological subtypes while outcome classes may also include choice of chemotherapy regimens, newer immunotherapies, as well as palliative interventions. Incorporating both imaging data and social variables into more sophisticated ‘hybrid’ models that more accurately reflect everyday practice is likely to be crucial for trustworthiness by patients and clinicians alike.

5. Conclusions

We have demonstrated ML – based predictive models trained on pre-treatment clinicopathological variables can predict curative oesophageal cancer MDT treatment decisions with good accuracy. We have shown that age plays a key role, especially when moving straight to surgery. The application of ML techniques has not yet been widely applied to oesophageal cancer MDTs despite some success in other clinical specialties [41–44]. ML tools have the potential to transform OC MDT workflow and efficiency with future research recommended towards integrated multimodal input datasets and focussed attention towards explainable XAI solutions thereby increasing trustworthiness and routine clinical use.

CRediT author statement

N.Thavanesan Conceptualization, Methodology, Data Curation, Investigation, Software, Formal analysis, Visualisation, Writing - Original draft.

I.Bodala Methodology, Validation, Writing – Review and Editing.

Z. Walters Conceptualization, Writing – Review and Editing.

S. Ramchurn Resources, Writing – Review and Editing.

T. Underwood Supervision, Project Administration, Resources, Writing - Review and Editing.

G. Vigneswaran Supervision, Project Administration, Methodology, Software, Writing – Review and Editing.

Funding support acknowledgement

NT receives a joint studentship from the Institute For Life Sciences (University of Southampton) and University Hospital Southampton.

Data availability

Anonymised study data available on request.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejso.2023.106986.

References

[1] Cancer research UK. Cancer Mortality for common cancers. https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared#heading-Zero; 2022.
[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca - Cancer J Clin 2018;68:394–424. https://doi.org/10.3322/caac.21492.
[3] Freeman RK, Van Woerkom JM, Vyverberg A, Ascioti AJ. The effect of a multidisciplinary thoracic malignancy conference on the treatment of patients with esophageal cancer. Ann Thorac Surg 2011;92:1239–43. https://doi.org/10.1016/j.athoracsur.2011.05.057.
[4] Depypere L, Thomas M, Moons J, Coosemans W, Lerut T, Prenen H, et al. Analysis of patients scheduled for neoadjuvant therapy followed by surgery for esophageal cancer, who never made it to esophagectomy. World J Surg Oncol 2019;17:1–9. https://doi.org/10.1186/s12957-019-1630-8.
[5] Al-Batran S-E, Ajani JA. Impact of chemotherapy on quality of life in patients with metastatic esophagogastric cancer. Cancer 2010;116:2511–8. https://doi.org/10.1002/cncr.25064.
[6] Calman K, Hine D. A policy framework for commissioning cancer services. 1995.
[7] Stephens MR, Lewis WG, Brewster AE, Lord I, Blackshaw GRJC, Hodzovic I, et al. Multidisciplinary team management is associated with improved outcomes after surgery for esophageal cancer. Dis Esophagus 2006;19:164–71. https://doi.org/10.1111/j.1442-2050.2006.00559.x.
[8] Van Hagen P, Spaander MCW, Van Der Gaast A, Van Rij CM, Tilanus HW, Van Lanschot Jjb, et al. Impact of a multidisciplinary tumour board meeting for upper-GI malignancies on clinical decision making: a prospective cohort study. Int J Clin Oncol 2013;18:214–9. https://doi.org/10.1007/s10147-011-0362-8.
[9] Lamb BW, Brown KF, Nagpal K, Vincent C, Green JSA, Sevdalis N. Quality of care management decisions by multidisciplinary cancer teams: a systematic review. Ann Surg Oncol 2011;18:2116–25. https://doi.org/10.1245/s10434-011-1675-6.
[10] Lamb BW, Sevdalis N, Arora S, Pinto A, Vincent C, Green JSA. Teamwork and team decision-making at multidisciplinary cancer conferences: barriers, facilitators, and opportunities for improvement. World J Surg 2011;35:1970. https://doi.org/10.1007/s00268-011-1152-1. –6.
[11] Achiam MP, Nordsmark M, Ladekarl M, Olsen A, Loft A, Garbyal RS, et al. Clinically decisive (dis)agreement in multidisciplinary team assessment of

esophageal squamous cell carcinoma; a prospective, national, multicenter study. Acta Oncol (Madr) 2021;60:1091–9. https://doi.org/10.1080/0284186X.2021.1937308.
[12] Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA, et al. The AUGIS survival predictor. Ann Surg; 2021. https://doi.org/10.1097/sla.0000000000004794. Publish Ah.
[13] Gong X, Zheng B, Xu G, Chen H, Chen C. Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer, vol. 13; 2021. p. 6240–51. https://doi.org/10.21037/jtd-21-1107.
[14] Tian Y, Liu X, Wang Z, Cao S, Liu Z, Ji Q, et al. Concordance between watson for oncology and a multidisciplinary clinical decision-making team for gastric cancer and the prognostic implications: retrospective study. J Med Internet Res 2020;22:1–11. https://doi.org/10.2196/14122.
[15] Thavanesan N, Vigneswaran G, Bodala I, Underwood TJ. The oesophageal cancer multidisciplinary team: can machine learning assist decision-making? J Gastrointest Surg 2023. https://doi.org/10.1007/s11605-022-05575-8.
[16] National Institute for Health and Care Excellence. Oesophago-gastric cancer: assessment and management in adults NICE guideline. 2018.
[17] Venables WN, Ripley BD. Modern applied statistics with S. fourth ed. Springer; 2002.
[18] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[19] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[20] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Routledge; 2017. https://doi.org/10.1201/9781315139470.
[21] Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, et al. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer from the departments of medicine, vol. 355. D.; 2006.
[22] Girling DJ, Bancewicz J, Clark PI, Smith DB, Donnelly RJ, Fayers PM, et al. Surgical resection with or without preoperative chemotherapy in oesophageal cancer: a randomised controlled trial. Lancet 2002;359:1727–33. https://doi.org/10.1016/S0140-6736(02)08651-8.
[23] Al-Batran SE, Homann N, Pauligk C, Goetze TO, Meiler J, Kasper S, 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 2/3 trial. Lancet 2019;393:1948–57. https://doi.org/10.1016/S0140-6736(18)32557-1.
[24] Shapiro J, van Lanschot Jjb, Hulshof MCCM, van Hagen P, van Berge Henegouwen MI, Wijnhoven BPL, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol 2015;16:1090–8. https://doi.org/10.1016/S1470-2045(15)00040-6.
[25] Reynolds JV, Preston SR, O'Neill B, Lowery MA, Baeksgaard L, Crosby T, et al. Neo-AEGIS (neoadjuvant trial in adenocarcinoma of the Esophagus and Esophago-gastric junction international study): preliminary results of phase III RCT of CROSS versus perioperative chemotherapy (modified MAGIC or FLOT protocol). (NCT01726452). J Clin Oncol 2021;39:4004. https://doi.org/10.1200/JCO.2021.39.15_suppl.4004.
[26] Evans L, Liu Y, Donovan B, Kwan T, Byth K, Harnett P. Improving Cancer MDT performance in Western Sydney – three years' experience. BMC Health Serv Res 2021;21:1–9. https://doi.org/10.1186/s12913-021-06203-y.
[27] National Cancer Equality Initiative/Pharmaceutical Oncology Initiative. The impact of patient age on clinical decision-making in oncology. 2012.
[28] Ahamat N. Access all ages: assessing the impact of age on access to surgical treatment. Bull Roy Coll Surg Engl 2012;94. https://doi.org/10.1308/147363512x13448516926748. 300–300.
[29] Depypere L, Thomas M, Moons J, Coosemans W, Lerut T, Prenen H, et al. Analysis of patients scheduled for neoadjuvant therapy followed by surgery for esophageal cancer, who never made it to esophagectomy. World J Surg Oncol 2019;17. https://doi.org/10.1186/s12957-019-1630-8.
[30] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum Comput Interact 2019;3:1–24. https://doi.org/10.1145/3359206.
[31] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?, vols. 1–28; 2017.
[32] Chen V, Li J, Kim JS, Plumb G, Talwalkar A. Interpretable machine learning. Queue 2021;19:28–56. https://doi.org/10.1145/3511299.
[33] Brierley RC, Gaunt D, Metcalfe C, Blazeby JM, Blencowe NS, Jepson M, et al. Laparoscopically assisted versus open oesophagectomy for patients with oesophageal cancer—the Randomised Oesophagectomy: minimally Invasive or Open (ROMIO) study: protocol for a randomised controlled trial (RCT). BMJ Open 2019;9:e030907. https://doi.org/10.1136/bmjopen-2019-030907.
[34] Straatman J, Van Der Wielen N, Cuesta MA, Daams F, Roig Garcia J, Bonavina L, et al. Minimally invasive versus open esophageal resection. Ann Surg 2017;266:232–6. https://doi.org/10.1097/SLA.0000000000002171.
[35] Nuytens F, Dabakuyo-Yonli TS, Meunier B, Gagnière J, Collet D, D'Journo XB, et al. Five-year survival outcomes of hybrid minimally invasive esophagectomy in esophageal cancer: results of the MIRO randomized clinical trial. JAMA Surg 2021;156:323–32. https://doi.org/10.1001/jamasurg.2020.7081.
[36] Tsujimoto H, Takahata R, Nomura S, Yaguchi Y, Kumano I, Matsumoto Y, et al. Video-assisted thoracoscopic surgery for esophageal cancer attenuates postoperative systemic responses and pulmonary complications. Surgery 2012;151:667–73. https://doi.org/10.1016/j.surg.2011.12.006.
[37] Nafteux P, Moons J, Coosemans W, Decaluwé H, Decker G, De Leyn P, et al. Minimally invasive oesophagectomy: a valuable alternative to open oesophagectomy for the treatment of early oesophageal and gastro-oesophageal junction carcinoma. Eur J Cardio Thorac Surg 2011;40:1455–63. https://doi.org/10.1016/j.ejcts.2011.01.086. discussion 1463-4.
[38] Blazeby JM. Minimally invasive or open oesophagectomy for localized oesophageal cancer: results of the ROMIO phase 3 randomized controlled trial. J Clin Oncol 2021;39:e16057–e16057. https://doi.org/10.1200/JCO.2021.39.15_suppl.e16057.
[39] Mederos MA, De Virgilio MJ, Shenoy R, Ye L, Toste PA, Mak SS, et al. Comparison of clinical outcomes of robot-assisted, video-assisted, and open esophagectomy for esophageal cancer: a systematic review and meta-analysis. JAMA Netw Open 2021. https://doi.org/10.1001/jamanetworkopen.2021.29228.
[40] Washington K, Watkins JR, Jay J, Jeyarajah DR. Oncologic resection in laparoscopic versus robotic transhiatal esophagectomy. J Soc Laparoendosc Surg 2019;23. https://doi.org/10.4293/JSLS.2019.00017.
[41] Lin FPY, Pokorny A, Teng C, Dear R, Epstein RJ. Computational prediction of multidisciplinary team decision-making for adjuvant breast cancer drug therapies: a machine learning approach. BMC Cancer 2016;16:1–10. https://doi.org/10.1186/s12885-016-2972-z.
[42] Diller GP, Kempny A, Babu-Narayan SV, Henrichs M, Brida M, Uebing A, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. Eur Heart J 2019;40:1069–77. https://doi.org/10.1093/eurheartj/ehy915.
[43] Wang Z, Sun J, Sun Y, Gu Y, Xu Y, Zhao B, et al. Machine learning algorithm guiding local treatment decisions to reduce pain for lung cancer patients with bone metastases, a prospective cohort study. Pain Ther 2021;10:619–33. https://doi.org/10.1007/s40122-021-00251-2.
[44] Andrew TW, Hamnett N, Roy I, Garioch J, Nobes J, Moncrieff MD. Machine-learning algorithm to predict multidisciplinary team treatment recommendations in the management of basal cell carcinoma. Br J Cancer 2021:1–7. https://doi.org/10.1038/s41416-021-01506-7.