University of Southampton

Faculty of Medicine

School of cancer sciences

The application of machine learning to the multidisciplinary assessment and management of oesophageal cancer

by

Navamayooran Thavanesan, BM BCh Hons BA (Oxon) MA (Oxon) MRCS

ORCID ID 0000-0002-7127-9606

Thesis for the degree of Doctor of Philosophy in Cancer Sciences

April 2026

DOI: https://doi.org/10.5258/SOTON/PG/T246

University of Southampton

Abstract

Faculty of Medicine
School of Cancer Sciences
Doctor of Philosophy

The application of machine learning to the multidisciplinary assessment and management of oesophageal cancer

by
Navamayooran Thavanesan

BACKGROUND & AIMS: Rising workflow pressures within oesophageal cancer multidisciplinary teams (MDT) can potentiate inconsistent decision-making, decision-fatigue, and even health inequality. Machine learning can automate portions of this workflow to alleviate caseload and standardize care provided, provided these techniques can address regulatory needs such as safety, accuracy, and transparency to improve clinical translatability. The aim of this research was to develop machine learning models to predict oesophageal cancer MDT treatment decisions and do so in an explainable fashion.

METHODS: Historic MDT decisions from University Hospitals Southampton (UHS) trained established ML algorithms (Multinomial Logistic regression (MLR), Random Forests (RF), Extreme Gradient Boosting (XGB), Decision Tree (DT), and Random Survival Forests (RSF)) to perform treatment classification and prognostication tasks. Classification models (MLR, RF, XGB, DT) predicted specific curative and palliative treatment plans, while palliative patients also had their estimated survival predicted when associated with a specific treatment. Classification models were assessed primarily on Area Under the Curve (AUC), while survival forecasts were assessed primarily by calibration curve. All UHS models were externally validated using data from Oxford University Hospitals (OUH). To integrate responsible innovation, transparency and explainability within this research, select eXplainable AI techniques (variable importance and partial dependence analyses) were also employed to examine how individual variables influenced predictions. The final user interface for interacting with the models was also guided using Responsible Research and Innovation (RRI) principles.

RESULTS: UHS models were trained from a total cohort of 953 cases and validated on 978 OUH cases. Model performance generalised regardless of algorithms and between treatment centres. XGB performed best for the primary classification model (mean AUC 0.909±0.044) whereas MLR demonstrated the best generalisability between centres (0.894±0.056). XGB performed best on palliative treatment classification both locally and externally (0.815±0.081, 0.742±0.064 respectively). The palliative survival model calibrated best in the first 12 months post-diagnosis for both cohorts. During this work, XAI techniques identified age as a significant influence on treatment allocation. Partial dependence analysis narrowed down the precise age at which probabilities for treatments shifted, as approximately 77 years.

CONCLUSION: Within this thesis I have shown that ML techniques can successfully model oesophageal cancer MDT treatment decisions and in a select subset: survival. These models support decision-making early within the MDT pathway. High-performing AI-based decision-support for the OC MDT is technically possible when combined with eXplainable AI methods to provide transparency for regulators as well as driving insight into potential biases within MDT-based decision-making. While future models might benefit from integration of raw imaging data and novel molecular markers, ML can synergize with current MDT frameworks. In future, this can evolve to prioritizing caseload, accelerating decision-making, and providing data-driven support for counselling patients in clinic when discussing treatment plans.

University of Southampton

Lay Abstract

The treatment of cancers of the oesophagus (food pipe) and stomach are managed by a specialist team of many different healthcare professionals called a multidisciplinary team (MDT). The MDT make the difficult and complex decisions, which can have significant implications for patient outcomes and quality of life. MDTs since the 1990s have led to major improvements in the management of cancer patients. However, the rising workload puts severe pressures on the MDT creating inconsistencies in decision-making and health inequality.

The purpose of this research was to use a branch of Artificial Intelligence (AI) called Machine Learning (ML) to develop computer models capable of replicating the human decision-making process of the Upper Gastrointestinal Surgery MDT. Machine Learning is an evolving branch of AI capable of learning complex patterns and relationships within large-scale data. Mirroring current human decision-making can provide the foundations for an assistive decision-support platform to use in collaboration with MDTs. This can help ease workload and provide data-driven recommendations for more complex situations to assist clinicians.

In this work I trained several easy to access, off-the-shelf machine learning software to learn and produce models for predicting treatment plans for new oesophageal cancer patients. Additional models were also trained to forecast survival in a subset of these patients. The design of these algorithms ranged from those which produce mathematical equations (Multinomial logistic regression), to more flow-chart style models (Decision Trees) and ensemble models which combine and average out hundreds of mini models (Random Forests, eXtreme Gradient Boost). To allow a user to understanding the reasoning behind predictions I also explored the use of eXplainable AI (XAI), a field aimed at improving transparency within AI models. Finally, I have endeavoured to consider principles of Responsible Research and Innovation within the user boundaries designed to house these models.

Within this work I trained prediction models on 953 patients managed in the Southampton area and tested the models on a further 978 patients managed in a different geographical area (Oxford) to determine how well they truly functioned when faced with new data. The results of this work found that Southampton models predict cancer treatment very well and continued to do so when predicting treatment for Oxford patients. This remained true when the process was reversed, and Oxford models were tested on Southampton patients. Additionally, survival models trained on incurable patients also performed well when estimating a patient's survival in the immediate 12 months after diagnosis. Investigating the transparency of these models highlighted key insights reflective of the broader MDT: showing that the probability of a specific treatment can change with the patient's age or the level of a patient's physical fitness. Finally, by engaging with the needs of key stakeholders such as clinicians, their patients and patient representatives, has provided guidance on how tools using these models should develop and evolve in the future.

During this research I have shown that it is possible to make predictive models that can recommend cancer treatment pathways and even estimate the likely survival for patients who are incurable. These models are easy to implement within the NHS, are transparent with techniques designed to explain predictions and while there are several new biological markers, tests and treatments on the horizon which will need to be incorporated into future versions, this research establishes proof of principle.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

Table of Supplemental Tables

Table of Supplemental Tables

Table of Supplemental Figures

Table of Supplemental Figures

Research Thesis: Declaration of Authorship

Print name: Navamayooran Thavanesan

Title of thesis: The application of machine learning to the multidisciplinary assessment and management of oesophageal cancer

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    - ***Thavanesan N,*** *Vigneswaran G, Bodala I, Underwood TJ. The Oesophageal Cancer Multidisciplinary Team: can machine learning assist decision-making?* Journal of Gastrointestinal Surgery. 2023, **27**(4): 807-822. PMID: 36689150
    - ***Thavanesan N****, Bodala I, Walters ZS, Ramchurn S, Underwood TJ, Vigneswaran G. Machine learning to predict curative multidisciplinary team treatment decisions in oesophageal cancer*. European Journal of Surgical Oncology. 2023, **49** (11):106986. PMID 37463827
    - ***Thavanesan N,*** *Farahi A, Parfitt C, Belkhatir Z, Azim T, Perez Vallejos E, Walters Z, Ramchurn S, Underwood T, Vigneswaran G. Insights from Explainable AI in Oesophageal Cancer Team Decisions*. Computers in Biology and Medicine. 2024, **180**: 108978. PMID: 39106674
    - ***Webb C, Thavanesan N,*** *Naiseh M, Dewar-Haggart R, Underwood T, Vigneswaran G. Health Care Professionals' perceptions of machine learning based clinical decision support systems for oesophageal cancer management*. Computers in Biology and Medicine. **2025,** 200: 111373. PMID 41380440
    - ***Thavanesan N,*** *Naiseh M, Terol M, Rahman SA, Hill SL, Parfitt C, Walters Z, Ramchurn S, Markar S, Owen S, Maynard N, Azim T, Belkatir Z. Perez Vallejos E, McCord M, Underwood TJ, Vigneswaran G. Oesophageal Cancer Multi-Disciplinary Tool: A co-designed, externally validated, machine learning tool for oesophageal cancer decision making. Lancet: eClinicalMedicine, Vol* ***89****, 103527*

Signature: ..... Date: 15/04/2026

Acknowledgements

I have to thank several people in my life without whom this degree simply would not have been possible. Firstly, I wish to thank Professor Underwood and Professor Walters for their supervision, insight, guidance and trust in me to achieve what would otherwise have felt impossible. They along with the Innovation for Translation Research Group have been invaluable in helping me see the "story" throughout this research and kept me from drifting off course.

Secondly, I have several people to thank who have made my early journey into machine learning and coding possible – to Alex Korneliou for patiently teaching me the very basics of the R language from another country, Dr. Marc Thomas for helping me understand some of the fundamentals of Machine Learning, and Miguel Terol who started as a mentor in both R and Shiny Apps (also all the way from another country) but through this experience became a close friend, collaborator and in whom I am forever indebted.

Much of my work has been closely supervised by Associate Professor Ganesh Vigneswaran, who again started off as a mentor but is now also a close friend and collaborator. His guidance and support have been instrumental in allowing me to achieve my aims for this research, always challenging me to push harder and strive for more.

Finally, I need to thank my family, my late mother, father and sister, all of whose sacrifices throughout my life led me to this moment, especially my mother, who never got to see me take this challenge on, but who has been in my heart throughout the journey. My family in-law who has also been an enormous source of support and encouragement and above all my wife Sophia whose unwavering strength and faith in me even when I doubted myself through everything, made this all possible.

Definitions and Abbreviations

Definitions and Glossary

Black-Box models	Models which are complex in nature and inherently non-transparent when attempting to follow or infer the underlying logic applied to a given prediction
Co-morbidity	the simultaneous presence of two or more clinical conditions within a patient
CROSS trial	A landmark clinical trial published in 2012 which demonstrated the survival benefit of providing neoadjuvant chemoradiotherapy prior to surgery for oesophageal cancer patients
Curability	Curability refers to the potential to achieve a complete cure for a patient with cancer and initiate remission. This may be through medication or surgical means (or a combination thereof).
Decision Support Tool	Also referred to as Decision Support System or Assistive Decision Tool. These are digital systems designed to provided support towards clinical decision making for clinicians to utilise.
Explainability	Explainability refers to the ability to extract meaningful insight into the internal logic of AI models and in doing so, better understand how they reach a given prediction. This is primarily a post-hoc process and happens after the model is developed.
Feature sets	The pool of variables used in a given model
FLOT4 trial	A landmark clinical trial published in 2017 which demonstrated the efficacy of the neoadjuvant chemotherapy regimen of 5-Fluoro Uracil (5FU), Oxaloplatin, Leucovorin and Docetaxel over the previous gold-standard chemotherapy regimen of the day for OC
Interpretability	An inherent property of an AI model which allows immediate understanding of its internal logic towards a given prediction task. Unlike explainable models, this is observable at all times.

Multidisciplinary Teams ... This is a clinical framework for shared decision-making in cancer care defined by the presence of multiple separate domain experts at the time of determining a course of treatments. MDTs typically comprise; cancer surgeons, oncologists, radiologists, pathologists, specialist nurses, palliative care physicians, administrative and clerical team members as well as many other allied health care professions.

Multimodal therapy..... the use of multiple oncological strategies for a patient's cancer care. Within Oesophageal Cancer this specifically relates to the use of neoadjuvant therapies as well as formal surgical resection

National Health Service.... The national health care system of the United Kingdom

Neoadjuvant therapy..... the provision of cancer treatments (typically chemotherapy, radiotherapy, hormone therapy or immunotherapy) prior to formal surgical resection of a tumour to downstage (shrink or improve the size and invasion of) the cancer.

NICE..... The National Institute for Health and Care Excellence is an executive non-departmental public body within the UK Department of Health and Social Care. It is responsible for assessing the cost-effectiveness of new and emerging treatments across healthcare as well as providing clinical guidance on the national management of a large array of health conditions.

Oesophagectomy ..... the surgical removal of part of the oesophagus and typically a portion of the proximal stomach.

Operability..... The assessment of a patient's disease condition and pre-treatment fitness that that speaks to their ability to tolerate surgical intervention

Palliative therapies ..... Palliative treatments are assigned to alleviate symptoms arising from a health condition. In the context of cancer care, this is commonly associated with patients whose cancer cannot be cured.

Performance Status ..... the Eastern Cooperative Oncology Group (ECOG) performance status is a clinical grading scale from 0-5 (0 = Fully active, able to carry on all pre-disease performance without restriction, 5 = dead)

	which has been traditionally used within surgical oncology to evaluate a patient's physical activity levels as a barometer of physiological reserve and fitness for therapeutic interventions.
Polypharmacy.....	The simultaneous use of multiple medications
Region of interest.....	Areas within an image (radiological or histological) which are under investigation. Within radiomics studies this typically relates to an anatomical feature on a scan being analysed.
Resectability.....	The potential to surgically remove a cancerous tumour safely without causing additional collateral damage.
Staging	Disease staging is the process of clearly defining a patient's cancer burden. This includes the severity of the primary cancer as well as any evidence of local or distant spreads. Prior to treatment initiation this is referred to as clinical staging and is primarily based on imaging. Following neoadjuvant/adjuvant therapies and surgery this is then referred to as pathological staging.
Tertiary referral unit.....	A specialist clinical centre or hospital which has particular expertise in managing a specific clinical condition
Tumour regression grade..	Tumour regression grade (most typically the Mandard scoring system) is a classification system which defines the histopathological response seen within cancerous tissue to neoadjuvant therapies.
Unimodal therapy	within Oesophageal cancer this typically relates to forgoing neoadjuvant therapy and proceeding directly to surgery

Abbreviations 8

Chapter 1 Introduction

Oesophageal cancer (OC) is the 14th commonest cancer in the UK and the 7th commonest cause of cancer death (1). At presentation only 39% of patients are eligible for curative treatment, while less than 15% are likely to remain alive at 5 years (2,3). The adenocarcinoma subtype has seen a 400% increase over two decades and is now more prevalent than squamous cell carcinoma in some world regions including North America, Northern Europe and Oceania (4). This is in part due to the increased prevalence of gastro-oesophageal reflux and Barrett's oesophagus combined with higher pick-up rates through screening and Barrett's surveillance.

The gold-standard management of oesophageal cancer remains curative resection however eligibility is heavily dependent on disease stage at presentation. Patients presenting with evidence of nodal disease will also require neoadjuvant therapy (NAT); either in the form of chemotherapy (NACT) or chemoradiotherapy (NACRT) (5). Both have shown a survival advantage over surgery alone although debate continues over which regime provides better oncological outcomes (5–9). The Neo-Aegis Trial continued to support clinical equipoise in the absence of a clear survival advantage from either modality (despite a noticeably higher incidence of pathological tumour regression within the CRT arm) (10). However recent data from ESOPEC (a German trial again comparing neoadjuvant chemotherapy and neoadjuvant chemotherapy in locally advanced OC) seems to have finally claimed a putative victor, reporting superior survival and even a higher pathological regression within the chemotherapy arm (11). Despite this new chapter in the neoadjuvant therapy saga, challenges remain. Survival benefit from neoadjuvant therapy is not conferred universally, with Noble et al., previously demonstrating that meaningful response to NACT was only seen in those with Tumour Regression Grade (TRG) 1-2 (14.8% of the cohort) deemed "responders". Overall survival in this group was 7.68 years versus 2.22 years in those with TRG 3-5 ("non-responders", 85.2%)(12). Identifying "responders" prior to starting neoadjuvant therapy is a major challenge as neoadjuvant chemotherapy can decondition patients prior to surgery, even to the point of inoperability (13–15). Predictive modelling has offered modest success at best in this regard when limited to pre-treatment data (16,17).

Oesophageal cancer patients are hugely reliant on high-stakes decision-making in typically complex clinical contexts, carrying serious implications for their outcomes and quality of life (18). Since 1995, cancer treatment decisions for patients managed in the UK have been made collectively by multidisciplinary teams (MDT) leading to improved patient outcomes (19–21). Current oesophageal cancer MDTs are required to comprise a core quorum. This includes: two or more surgeons, a

gastroenterologist with specialist endoscopic skills, a clinical oncologist, a medical oncologist (where the responsibility of chemotherapy delivery is not assigned with the clinical oncologist), a histopathologist, imaging specialists (including an interventional radiologist), a nurse specialist, core members of the palliative care team, dieticians and an MDT co-ordinator (22). This group is mandated to meet regularly (typically weekly) to discuss new cancer cases, ongoing cancer patient care and recurrent cases or complex cases. However, these services face increasing caseloads, increasing treatment options and clinical complexity, leading to inconsistent and sometimes suboptimal decisions (23). Individual experience, perception and bias can also lead to discordance, creating "noise" within the process (24). In 2017 Cancer Research UK recognised the need for improving the operational effectiveness of the UK MDT framework (25). Their report confirmed a clear trend in the period of data analysis (2011-2014): a linear increase in caseload being managed with little to no increase in resources for the MDT to operate with. An aging population and widening treatment options increased the clinical complexity of cancer cases across cancers without a commensurate evolution in MDT structure to adapt to the rising workload. Modern MDTs have on average 2-3 minutes to discuss and agree a patient's treatment plan, reducing their ability to audit decisions, learn or reflect. These issues alone would be sufficient for an urgent call to reform the current MDT structure; however, the cost of MDTs pose an additional highly relevant consideration. At the time of the CRUK report, the mean cost per MDT discussion (i.e. per patient discussed, per meeting, across specialities) was approximately £100. The highest cost was associated with colorectal surgery (£132.95) and the lowest was breast (£91.84). The national cost of MDTs rose from £88 million in 2011/12, to £150 million in 2014/2015 and as of 2024 sits at £316 million (26–28). A process to prioritise, accelerate and rationalise weekly MDT discussion lists is therefore essential within the current economic and clinical climate where budgetary restrictions across the NHS remain austere.

Clinical uncertainty and equipoise remain a core facet of MDTs within a modern era of expanding treatment options in surgical oncology. Controversies in the management of oesophageal cancer include optimal neoadjuvant therapy regimes, the optimal surgical approach for surgical resection (open versus laparoscopic versus robot-assisted), the appropriate classification of junctional tumours (gastric versus oesophageal) and perhaps most recently the prognostic importance and implications of resecting locoregional lymph nodes on the host's immune response to cancer in the tumour microenvironment (10,29–31). The consequence of these uncertainties remains a consistent trend of variability in practice and decision-making (2,3,32–34).

The use of data-driven clinical decision-support systems (CDSS) represents a tangible solution which may allow more standardised decisions moving forward with the benefit of being able to update recommendations as new clinical research is released. CDSSs are becoming increasingly commonplace within medicine (QRISK®, IC-RISC™ and QCancer among numerous others) (35–37). They may range from simpler tools summarising national guidelines to more sophisticated recommendation systems leveraging unstructured data (38–44). Despite their growing popularity, routine adoption of these systems remains in its infancy (45). It has been demonstrated however that with appropriate stakeholder buy-in and regulatory support, CDSSs can significantly impact clinical practice. The National Emergency Laparotomy Audit (NELA) for example, has managed to achieve widespread use nationally in the UK as a tool for objective operative-risk stratification, predicting the need for higher levels of care following emergency laparotomy (46,47). However, while NELA at least in its current form represents a traditional, statistical framework for decision-support, the domain of Machine Learning (ML), a sub-division of Artificial Intelligence (AI)) offers a significantly more powerful vehicle through which to model complex decision-making paradigms (48). Machine Learning can handle huge datasets, capture complex relationships, non-linear interactions and produce highly accurate predictions, making it an appealing candidate for healthcare-based task-automation and more specifically; team-based frameworks such as the MDT (49). Thus far, ML-based decision-support models have been trialled within cardiac disease (43), breast cancer (44), lung cancer (42), pancreatic cancer (50) and dermatological cancers (41). Prior to this research however, no such approach has been ever made to the oesophageal cancer MDT representing a substantial opportunity within a high-stakes clinical arena. As demonstrated by work outlined in Appendix F, there is clear evidence of an appetite from MDT personnel within OC for data-driven decision support. Additionally, this work highlighted that there remain areas of discordance between what human agents making MDT decisions feel is important when compared to ML-driven insights into these patient cohort. While a significant portion of respondents from the National Survey conducted in Appendix F showed a positive mentality towards ML, it is not universal at present, and scepticism remains for many MDT attendees of its long-term benefits.

This may be in part due to the ongoing confusion among clinicians who do not routinely employ ML-based technologies as to how it improves upon or differs from more traditional statistical modelling, especially as ML is built upon statistical foundations. The first, and perhaps most significant, difference is that while traditional models are designed to provide insight into the relationships between predictor variables following which these relationships may also be used to make further predictions, ML modelling is designed with the sole aim of maximising

predictive performance even at the expense of explainability or transparency of the inter-variable relationships (51). In other words, traditional statistics cares more about the nature of the variables in the process, ML cares more about the outcome. Additionally, while ML may be comfortable with black-box models (this is gradually shifting), this is antithetical to a statistical paradigm where understanding the nature of the model variables and their interactions are paramount. While statistics is interested in causal inference, this is not necessarily a priority for ML. Finally, while ML algorithms aim to produce models which can make accurate decisions or predictions in the absence of explicit coding, statistical models focus on making inferences about the wider population based on the original sample (52).

The Artificial Intelligence boom within healthcare has brought with it a pressing need for trustworthy, safe, ethical and responsibly developed AI across disciplines, with medical AI (MAI) a particularly high priority (53–55). The appreciation for the dangers surrounding AI innovation have recently been reflected within the landmark EU AI Act 2024, the first of its kind for AI regulation and the de facto benchmark for global AI innovation going forward (56,57). While much of the literature has understandably focused on proving medical AI tools at a technical level, there is a grave paucity in considering the implications of innovations on stakeholders, patients and patient advocates from design-to-deployment (55). These include (but are not limited to) bias, quality control, data-drift detection and AI explainability (58,59). In response to this, the field of Responsible Research and Innovation (RRI) has grown in recent years with the intent of guiding research towards maximising societal benefit and minimizing harm (60). Frameworks such as “AREA” for instance (Anticipation, Reflection, Engagement and Action) and more recently “AREA-Plus” are well-known within the UK, providing a practical structure for integrating RRI within the life cycle of research programs (60–62). It follows then that integrating responsible innovation is going to be essential to driving adoption of AI technologies in healthcare settings.

The body of work I present within this thesis is intended to address this unmet need for trustable, responsibly co-designed, data-driven decision-support within Oesophageal Cancer. I have developed an externally validated AI clinical decision support system utilizing off-the-shelf machine learning algorithms which can be integrated into user-friendly interfaces as needed. It has been responsibly developed using an in-parallel Responsible Research and Innovation program of regular workshops and interviews working with domain experts in the form of patients, patient-advocates, clinicians, and explainable AI (XAI) computer scientists. The CDSS is trained to recommend oesophageal cancer MDT treatment plans early within a patient’s

clinical journey with additional prognostication for palliative cases. My work demonstrates that it is possible to harness AI-based technologies to replicate and simulate the oesophageal cancer MDT decisions as a practical and translatable means to streamlining, standardising and supporting the MDT operational framework while aligning with RRI principles.

The following sections aim to contextualize the MDT's role within OC in more depth, discussing the experimental applications of ML within OC to date (such as the prediction of treatment response and the emerging potential for radiomics for prognostication, nodal disease evaluation, and resectability) and finally outline the hypothesis, aims and objectives of my thesis which intends to contribute to this field.

1.1 Review of the literature

1.1.1 Acknowledgements

A version of the review outlined within the following sections of this chapter has previously been published within the Journal of Gastrointestinal Surgery (63). While the work presented here is my own, I wish to acknowledge my co-authors for their contributions: Dr. Ganesh Vigneswaran (supervision and manuscript review), Dr. Indu Bodala (supervision and manuscript review), Professor Tim Underwood (supervision and manuscript review).

1.1.2 The Multi-Disciplinary Team (MDT)

Clinical management for all UK cancer patients has been centralised through MDTs following the Calman-Hine report in 1995 (64). The aim of this reformation was to consolidate expertise from all clinical disciplines relevant to a patient's oncological treatment in a single place and time as opposed to a serial chain of clinical interactions. This requires a broad scope of healthcare professionals – surgeons, physicians, oncologists, radiologists, specialist nurses, physiotherapists, occupational therapists, palliative care teams and administrative staff. The benefit is rapid, nuanced, complex decision-making early and (theoretically) consistently during the assessment, treatment and follow-up stages of cancer care. The MDT evaluates and agrees on: cancer origin, anatomical location, disease stage, curability, resectability, fitness for surgery or aggressive oncological therapies and the patient's wishes. All of this is then be conveyed to the patient to discuss the options available and reach a mutually agreeable plan.

1.1.3 Strengths of the MDT framework

Numerous studies have proven the benefit of managing oesophageal cancer via an MDT over the historical practice of surgeons managing these cases independently (19–21,65). One Welsh study comparing 77 patients managed by surgeons independently against 67 cases managed by an MDT reported the incidence of open-and-close surgeries (laparotomies and thoracotomies) having dropped from 21% and 5% to 13% and 0% respectively when put through MDT ($p = 0.02$). Operative mortality also dropped (26% vs 5.7%, $p = 0.004$) and 5-year survival improved significantly (52% vs 10%, $p = 0.0001$). Additionally, multi-variate analysis found that in combination with lymph node metastases and American Society of Anaesthesiologist (ASA) grade (an anaesthetic marker of fitness for surgery, dependent on the patients' current comorbidities), MDT management, was

independently associated with improved survival (20). Rates of completed staging reportedly jumped from 67% to 97% ($p < 0.0001$), along with adherence to national guidelines for management (83% to 98%, $p<0.0001$) simply by introducing a formal thoracic MDT to the management oesophageal cancer patients (19). MDTs can influence and course-correct management strategies with one study reporting a change in over one third of management plans originally designed by individual clinicians for potentially curative OC cases after MDT discussion (21).

Palliative cases have also benefitted from MDTs. One Dutch registry-based study of palliative oesophageal cancer patients compared 389 cases discussed at MDT (MDT group) versus 547 cases that were not (65). Within the MDT group, the study reported a significantly shorter time to commencement of palliative therapy (20 days vs 30 days, $p < 0.001$), a higher incidence of palliative external beam radiotherapy (EBRT) (38% vs 21%, OR 2.7), higher incidence of systemic therapy (30% vs 23%, OR 1.6), fewer patients treated with palliative stents (4% vs 12%, OR 0.3) and longer overall survival (169 days vs 107 days, HR 1.3). While the authors acknowledged that prognostic factors not recorded within the registry may have also contributed to the survival difference, they attributed at least part of this survival advantage to the increased usage of tumour specific palliative therapies within the MDT group.

1.1.4 Vulnerabilities of the MDT

MDT frameworks are not invulnerable to clinical, inter-personal and logistical challenges. Rising caseloads, pressured preparation time, missing staging data, patient complexity, and intra-group disagreement can all lead to inconsistent and varied decision-making.

1.1.4.1 Workload

Deficiencies within the MDT workflow have been identified as far back as 2011. A systematic review in the same year explored the clinical, social, and technological factors influencing MDT decision-making across multiple specialities (23). The authors reported that definitive plans were reached at first discussion in only 47.6-73% of cases because of time pressure or missing information. Where plans were agreed, they were not implemented in anywhere from 1-16% of cases either because of differing patient wishes or perhaps more worryingly, wholly inappropriate management plans when patient co-morbidities were factored in. Excessive workload and time pressure were flagged as contributory to poor decision-making, lower team morale, and unmet need for protected preparation causing wasted time and/or increased workload. MDTs covering general surgery, soft tissue cancers and urology were found to have clinician-made decisions based almost entirely on

clinical information, infrequently factoring in patient wishes unless specialist nurses present felt empowered to bring those views up in the meeting (66). In a similar theme, the study noted that while physicians drove decision-making within the meetings, they often ignoring nurse-led input usually at the detriment of the MDT's overall efficacy. Interestingly, even at the time of the review in 2011, telemedicine-based tools were found to be cost-effective (where used in at least 20-30 meetings per year) and able to increase attendance without adversely affecting care. The authors did however recognise the negative impact it had on the feasible caseload per meeting. Despite this, virtual MDTs have successfully achieved widespread adoption since the 2020 SARs Cov2 pandemic.

1.1.4.2 Interpersonal dynamics

While the diverse composition of MDT attendees can positively influence performance, a lack of clarity or conflict over leadership can be a negative predictor for effective communication, clarity of objectives, team effectiveness, resource efficiency, or effective patient communication as demonstrated in breast cancer by Haward et al (67). Compellingly, the authors also noted that a single strong leader could be a negative predictor for support in innovation, indicating a delicate balance between a single strong "voice" and one supportive of team input. Perception of team-effectiveness varies significantly by discipline within MDTs, with Haward et al., reporting that breast surgeons and breast care nurses consistently rated their MDT's performance higher than their radiology and histopathology counterparts.

Communication of MDT decisions to patients is susceptible to compromise by inter-personal MDT disagreement. Hamilton et al., investigated 35 MDTs and 37 MDT clinics to evaluate the level of patient-inclusive decision-making used in head and neck cancer management across 3 centres (68). While the study sample was modest (20 patients and 9 MDT members), they utilised a combination of direct observation, informal interviews, and formal semi-structured interviews to identify significant barriers to shared decision-making between the MDT and their patients. Individual members often held a clear personal view of what they deemed the best course of action, yet this did not always align with fellow team members, posing a challenge for the MDT to convey this uncertainty to the patient. The authors reported that MDTs often felt such disagreements should be kept internally, and even when an individual was tasked with conveying the final MDT outcome to the patient, the conversation risked being "framed" in a manner filtered by that clinician's own biases. Ultimately this internal dissent can force the bulk of the decision-making to remain internal to the MDT, at the risk of excluding patient values and wishes.

Cancer MDTs are also subject to disagreements between MDTs even within a specialty. An observational multi-centre Danish study investigated this inter-observer variability between MDT decisions between 4 main upper gastrointestinal (UGI) cancer centres in Denmark following centralization of their UGI services in 2002 (69). The study sample was again small (20 oesophageal squamous cell carcinoma cases), each of which were repeatedly assessed as new referrals at each of the centres. Each MDT was asked to determine: resectability, curability (determined by operability status) and treatment strategy. The authors calculated a kappa-like coefficient for inter-observer variability, as well as the frequency by which disagreement between MDTs resulted in a different treatment recommendation and whether this had any clinical impact on the patient. The study reported "moderate" concordance between the 4 MDTs on classifying T-stage, M-stage, resectability, and curability while only "fair" concordance was reached for N-stage and operability. The biggest impact of their findings was that MDT disagreement led to a clinical impact in 60% of cases. The study was limited by the very small sample size (again owing to the busy caseloads experienced by the MDTs external to the study), and missing positron emission tomography (PET) computed tomography (CT) images for 5 of the 20 (25%) study cases. Operability was crucial to determining an appropriate treatment strategy and yet found to be most vulnerable to inter-MDT discordance due to often-incomplete data available to the MDT.

1.1.5 Current UK management guidelines used by Oesophageal Cancer MDTs

Table 1.1 outlines the 2018 National Institute for Heath and Care Excellence (NICE) guidelines for the management of oesophageal cancer (70). Notably while some authors categorize T2N0 disease as early and amenable to endomucosal resection (EMR), NICE supports the use of neoadjuvant therapy in this cohort, to minimise local recurrence risk from micro-metastases (71,72). Histology, Tumour-Node-Metastasis (TNM) staging, and an assessment of patient fitness (commonly quantified by the World Health Organisation (WHO) Performance Status classification) account for the bulk of decision-critical parameters. While the concept of comorbidity is acknowledged, especially when determining suitability for palliative chemotherapy, such guidelines remain simplistic, rarely factoring in dimensions such as high-risk comorbidities, social variables or even ease of patient access to chemoradiotherapy centres.

Table 1.1 - 2018 NICE guidelines for the management of OC

Disease stage	**OAC**	**OSCC**
T1aN0	Offer EMR	Offer EMR
T1bN0	Offer Surgery	Offer either - Definitive CRT - Surgical resection
T2-4 N0-3 M0	Offer either: - NACT ± ACT - NACRT Assess response Then surgery	Offer either - Radical CRT Or: - NACRT Assess response Then Surgery
Non-metastatic disease unsuitable for surgery	Consider - CRT if feasible within RT field Or: - Chemotherapy - Stenting - Palliative RT - Best supportive care	Consider - CRT if feasible within RT field Or: - Chemotherapy - Stenting - Palliative RT - Best supportive care
Metastatic disease	If HER2 +ve: - Trastuzumab (Herceptin) 1st line palliative chemotherapy (If performance status 0-2, no significant comorbidities): 2nd line palliative chemotherapy	If HER2 +ve: - Trastuzumab (Herceptin) 1st line palliative chemotherapy (If performance status 0-2, no significant comorbidities): 2nd line palliative chemotherapy

Abbreviations: OAC - oesophageal adenocarcinoma, OSCC – oesophageal squamous cell carcinoma, EMR – endomucosal resection, CRT – chemoradiotherapy, NACRT - neoadjuvant chemoradiotherapy, NACT - neoadjuvant chemotherapy, ACT - adjuvant chemotherapy, HER2 - human epidermal growth factor 2

1.1.6 A role for Machine learning?

Machine learning (ML) has gained popularity within healthcare for its ability to analyse large, complex datasets and provide, advanced predictive modelling. ML-driven decision-support

models can also actively improve as new data is obtained. An example of this includes the strong performance shown in the accurate predictive modelling of outcomes after oesophagectomy (73). However, while post-operative models show good discrimination and calibration, pre-operative models are typically more challenging as they are by nature trained on significantly fewer features (74).

The MDT discussion prior to first treatment is nevertheless a key mile-marker in the patient's care pathway and optimising the decision-making at this check point is critical to providing the best outcomes possible for patients. As MDTs typically assimilate information across clinical, pathological and radiological sources, each of these domains separately offers a potential focus for the application of ML. Aggregation of these data streams within machine learning models could then allow "mirroring" of the current human-led decision-making paradigms seen within MDTs.

Machine learning is traditionally divided into supervised and unsupervised learning with the former requiring 'labelled' data (the ground truth is provided to the machine during model training). The machine then compares the input and outcome data to determine the model which best fits the underlying structure of the data. Supervised learning is consequently well suited to smaller datasets, where the ground truth is known – a prime example being historic MDTs where treatment decisions for patients are already known. By comparison, unsupervised learning algorithms identify patterns within datasets to extract features that may cluster data points into separate groups. Such techniques are useful when the ground truth is unknown but requires large volumes of data - a challenge frequently encountered in cancer datasets. Ideal models balance "under-" and "over-fitting", learning from training data to make accurate predictions when fed new unseen data. Under-fitted models are typically too simplistic or inflexible to capture underlying relationships leading to high error rates in both training and testing (bias). Over-fitting occurs when the model is too complex resulting in high variance. These models perform well within training but struggle on test/validation sets (75). This may be mitigated by increasing the size of the training data and the diversity of the observations themselves, making it more representative of the theoretical population distribution. In real-world settings however, this is often difficult to achieve with health data especially for rarer clinical scenarios. Table 1.2 summarises some common ML based techniques along with their respective advantages and disadvantages.

Table 1.2 - Common Machine Learning techniques

Algorithm	**Summary**	**Benefits**	**Drawback**
Decision trees	Flow-chart based modelling whereby variables are trialled at each "node" of a tree (decision split point) to determine the best combination of root-, branch- and leaf nodes for the overall model	Provides an interpretable model, and easy to visualise No assumptions made about data distribution Can manage regression and classification tasks	Less well suited to continuous variable outcomes Produces a single tree but may be computationally expensive to grow tree as must trial every split of variables at each node Prone to overfitting, especially if large number of variables and small datasets
Random forest	A tree-based modelling technique which aggregates hundreds of individual decision trees, each composed of a random selection of predictor variables	Copes with large feature pools Randomly selecting a subset of variables for each tree rather than the full pool minimises overfitting and increases generalisability Can be used to assist feature selection based on relative importance of each variable	Sacrifices interpretability for overall model performance Vulnerable to outliers within dataset

Algorithm	**Summary**	**Benefits**	**Drawback**
Ridge regularization	Also known as L2 regularisation, is a form of regularisation method which acts to minimise a loss function (a penalty associated with misclassification).	Ridge regularization produces a more generalisable regression model by shrinking variable coefficients to reduce model overfitting	Ridge regression never shrinks coefficients to "0" thus maintaining all variables within the model. This in turn reduces interpretability
Least absolute shrinkage and selection operator (LASSO)	Also known as L1 regularisation. Similar approach to Ridge regression, however the penalty function is derived from the absolute sum of the coefficient as opposed to their square as is used in Ridge.	Allows automatic feature selection LASSO allows coefficients to be shrunk to "0" and effectively drops them from the model which allows for feature elimination. Used to minimise model overfitting	In situations where predictors outnumber the observations, LASSO will reduce variable pools even if non-significant variables are nevertheless relevant to the model as whole. Similarly, where variables may be correlated, LASSO may randomly select one and eliminate the other
Logistic regression	Form of regression analysis for outcomes which are categorical (and often binary). Learns a linear relationship in form $y = c + \beta_1 x + \beta_2 x + \beta_3 x + \ldots \beta_n x$ to predict probability of a given class.	Provides an interpretable model The variable co-efficient enumerate the relative weights of each variable to the overall model, and direction Easy to train and computationally inexpensive	Requires linearity between the predictors and outcomes Observations need to be independent of each other Limited to categorical outcome prediction

Algorithm	**Summary**	**Benefits**	**Drawback**
Support Vector Machine	Segregates data by creating a decision boundary of "hyperplane" to allow class separation	Useful in binary outcome predictions Capable of handling high-order data relationships Commonly used in radiomic tasks	For more complex higher-order data, requires elevation of data into higher dimensions to achieve hyperplane
Convolutional neural network	Uses multiple "hidden layers" of processing to analyse input data and provide a task outcome. Deep learning models are formed around the concept of recreating neural networks – comes under ML discipline of Deep Learning.	Powerful ML approaches Particularly suited to complex tasks such as audio and image analysis	Computationally intense Requires large volume datasets Sacrifices interpretability for overall model performance

1.1.7 Current ML applications in oesophageal cancer

1.1.7.1 Histopathological analysis

Machine learning applications within histological analysis have risen in popularity within research settings (76–78). Integrating machine learning with computer-vision (a branch of computer science dedicated to the recognition and extraction of meaningful information from images) offers cheap, automated, scalable analysis and decision-making within cancer care. This contrasts with techniques such as ribonucleic acid (RNA)- and Whole Genome Sequencing (WGS) where detailed and individualised data is available but ultimately cost-prohibitive and logistically challenging when attempting to acquire additional tissue retrospectively (79).

Despite the practical appeals, few studies have applied computer vision-based ML techniques to oesophageal cancer (Table 1.3) (79,80). Where this has been attempted, typically this has been done with the aid of convolutional neural networks (CNNs), a form of neural networks known to excel in tasks involving computer-vision-based classification. They include a convolutional layer within the neural network (layers of inter-connected nodes with associated activation weights) which include a kernel or filter designed to sweep across input images and produce a feature-map which can then be compared to pre-trained object features for feature detection. Pilot work by Rahman et al., attempted an innovative approach to predicting response to neoadjuvant therapy in oesophageal adenocarcinoma at the tissue level by combining automated visual capture and CNN processing of unlabelled digital histology slides (79). The CNN analysed high-resolution microscope slide images fragmented into distinct "patches", as a means of achieving Whole Slide Image (WSI) analysis without losing granular data at the pixel level when downsizing such massive images. The CNN was pre-trained using ImageNet (non-specific images from a vast online database of everyday images) to circumvent the need for high volume histology-specific training images that would otherwise be needed to train a sufficiently accurate model. Despite a small sample the authors reported an internally validated C-index of 0.836 (0.825-0.847) in training the CNN to distinguish between responders (Mandard Tumour Regression Grade or TRG) 1-2 and non-responders (TRG 3-5). While these results were promising, the study had some limitations. Their results have yet to be achieved in larger datasets, the use of both neoadjuvant chemoradiotherapy and chemotherapy within the patient cohort may have confounded their results, and finally, such deep learning approaches inherently create a "black box" problem as the underlying logic is obscured within the hidden CNN layers. This limits transparency, "explainability" and ultimately trust within the ML

solution, a scenario seen again with Tomita and colleagues who applied a ResNet based deep learning model to the detection of Barrett's and adenocarcinoma in oesophageal tissue biopsies, again using ImageNet trained platforms. Here they sought to show that strong classification performance could be achieved without the need for pre-annotating regions of interest, removing the human "bottleneck" in the process. While their results were promising in principle, their training sets were orders of magnitude smaller than what many accept is needed for deep-learning models, nor did they have a strategy to explain the assignment of attention provided by the model on the extracted images (80). Pre-trained networks nevertheless have performed competitively against models trained from scratch and still offer a possible solution to the issue of applications limited by inherently restricted datasets (81).

Table 1.3 - Studies applying ML to histopathological data within OC

Study	Country	Study size (n)	Histology	Image modality	ML techniques	Outcome Predicted	Model performance metric	Results
Rahman et al., 2021 (79)	UK	46	Mixed	WSI with patch conversion	CNN (Xception) + Elastic Net Regression	Response to neoadjuvant therapy (NACT/NACRT) comparing histopathological analysis vs RNAseq	AUC	AUC for histopathology slide features 0.763 vs RNAseq (0.782) AUC for segment slides exceeded both (0.870)
Tomita et al., 2019 (80)	USA	180	AC, BE, Dysplasia	WSI with patch conversion	CNN (ResNet-18) + Attention-based Neural Network	Classification of Barrett's ± dysplasia and oesophageal adenocarcinoma comparing tissue-level annotations vs traditional ROI segmentation	Accuracy, recall, precision, F1 Score	Mean Accuracy of 0.73 for differentiating BE, BE + dysplasia and AC F1 scores for differentiating BE, BE + dysplasia and AC were 0.72, 0.30, and 0.67 respectively.

AC = adenocarcinoma, BE = Barrett's oesophagus, WSI = whole slide image, AUC = area under receiver operator characteristic curve, NACRT = neoadjuvant chemoradiotherapy, NACT = neoadjuvant chemotherapy ANN = artificial neural network, CNN = convolutional neural network, RNAseq = sequenced ribonucleic acid

1.1.7.2 Radiomics

Over the last two decades, a substantial evidence base across numerous solid organ cancer types has developed in radiomics (82). Radiomics refers to the extraction of quantifiable, clinically significant, high-dimensional imaging-based biomarkers from standard-of-care medical imaging which may correlate with tumour phenotype and its molecular fingerprint (83). These markers are then treated as potential predictor variables when modelling a range of clinical outcomes (75). Human radiological assessments within MDTs are mainly qualitative, with some quantification of tumour size, number and position of suspected lymphadenopathy and the presence of distant metastases within the TNM staging classification (84). The underlying tenet of radiomics lies in the assumption that human assessment is biologically limited in its discrimination at a pixel/voxel level and inherently involves a degree of both selection bias as well as inter- and intra-observer variability (85). Radiomics aims to mine the image for more precise evaluation of disease burden. Coupling this to the infrastructure of the MDT could, in theory, achieve high-precision assessment of their disease, resectability and potentially reveal features indicative of likely treatment response to neoadjuvant therapy prior to even starting therapy.

1.1.7.2.1 Radiomic workflow

The radiomic process (Figure 1.1) can be summarised as: image acquisition, image pre-processing, segmentation, feature extraction, data preparation, feature reduction, and model development (75,86). Image acquisition involves the curating of imaging series containing regions of interest (ROI). Image pre-processing includes segmentation of regions of interest which may be manual (considered gold standard but resource intensive), automatic, or hybridized. Once ROIs are segmented, radiomic “features” can be extracted and converted into quantifiable vectors from within these regions, forming the functional core of radiomics (75,85). Vectors often differ in scale, thus the data preparation stage in this process frequently includes feature scaling, data continuization, discretization, and under- or over-sampling for class imbalances (87). The resultant features may however be hundreds in number, and consequently counter-productive to a well performing model (88). Dimensionality reduction and feature selection technique are typically used in this situation to minimise redundant, non-relevant features which may slow the model (89–91). Once this is done, the final feature pool can be used to train the final radiomics model towards its intended use. Validation of the

generated model must then be done internally and externally as this will inform the eventual generalizability of the final model in clinical use (92).

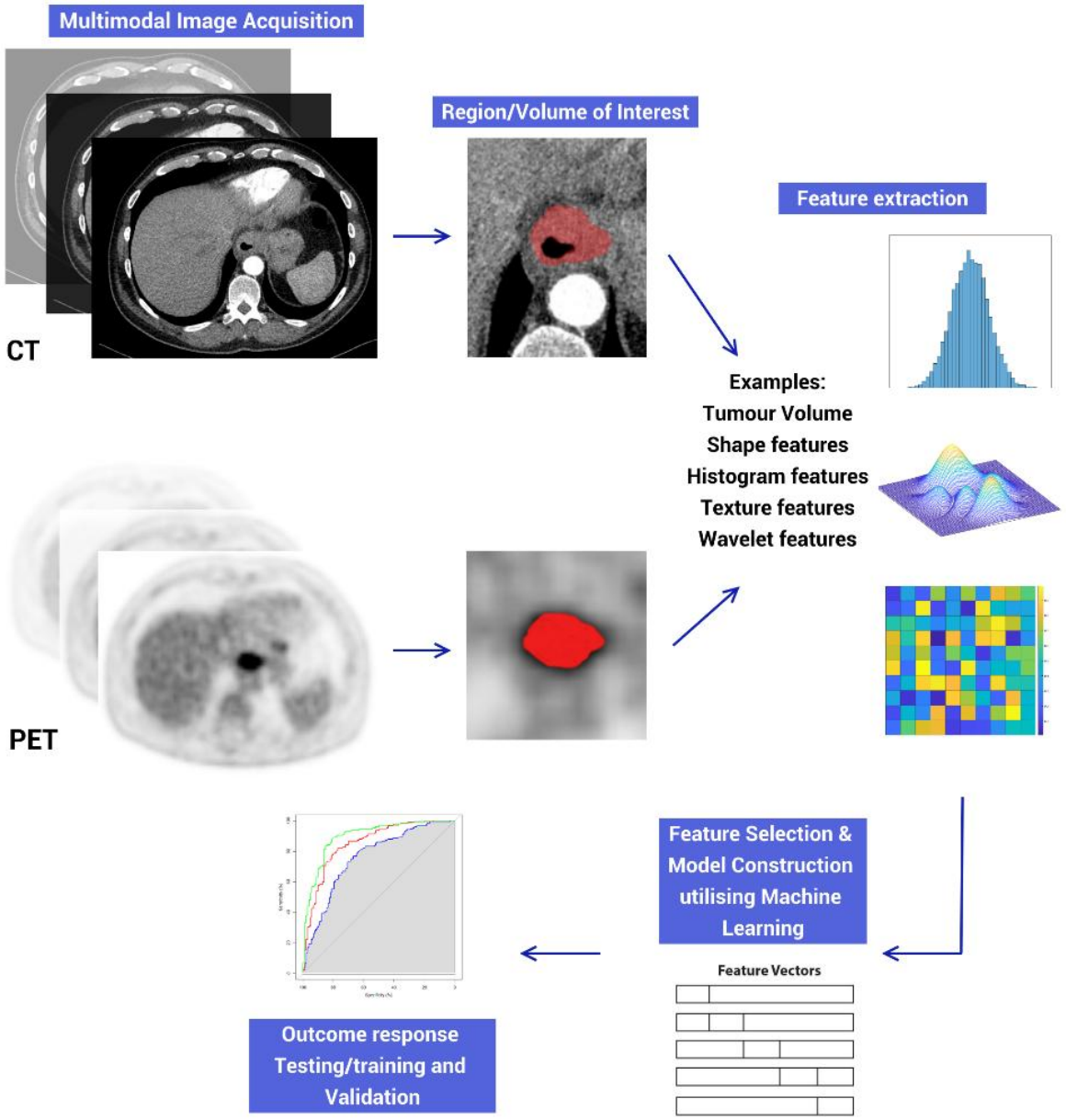


Figure 1.1 - A standard Radiomics workflow depicting the stages of image acquisition, processing, annotation, feature extraction, model training and validation (Thavanesan et al. 2023, doi: 10.1007/s11605-022-05575-8).

1.1.7.2.2 Radiomic studies within oesophageal cancer

An evolving body of evidence is now emerging within oesophageal cancer for predicting treatment response, prognosis, nodal status and even resectability (93). Using radiomics to improve the speed and accuracy of each of these facets of oncological decision-making can drive forward a significant portion of the MDT's weekly workflow. Table 1.4 summarises studies which have applied radiomics within the oesophageal cancer domain.

Table 1.4 - Studies applying Radiomics to OC

Study	Country	Study size (n)	Histology	Imaging modality	ML techniques	Outcome Predicted	Model performance metric	Results
Ou et al., 2019	China	591	SCC	CT	LASSO, Logistic regression, RF, SVM, XGBoost, Decision trees	Resectability of SCC	AUROC, Accuracy, F1 Score	Logistic regression radiomics model performed best (validation set AUC 0.87 ±0.02, accuracy 0.86, F1 score 0.86).
Hou et al., 2017	China	49	Mixed	CT	SVM, ANN	Therapeutic response to NACRT	AUROC	Radiomics based SVM AUC 0.891, ANN AUC 0.972 for responders vs non responders Skewness and Kurtosis features capable of differentiating partial response and stable disease, Kurtosis also discriminatory for partial versus complete response
Larue et al., 2018	Netherlands	239	Mixed	CT	RF	3 -year survival post NACRT	AUROC (95% CI)	Radiomics RF model validation set AUC 0.61 (0.47 – 0.75) vs Clinical parameter RF validation set AUC 0.62 (0.49 – 0.76)
Tan et al., 2019	China	230	SCC	CT	LASSO Logistic regression	Predicting LN metastases in resectable SCC	Discrimination, Calibration, and Reclassification	AUC of model combining radiomic signature with CT LN status was 0.773. Discrimination of signature significantly better vs LN size criteria alone (p = 0.005)
Beukinga et al., 2017	Netherlands	97	Mixed	PET/CT	LASSO Logistic regression	pCR following NACRT	AUROC	Model combining clinical parameters with PET/CT derived textural features outperformed SUVmax models (AUC 0.74 vs 0.54 on internal validation).

Study	Country	Study size (n)	Histology	Imaging modality	ML techniques	Outcome Predicted	Model performance metric	Results
Simoni et al., 2020	Italy	54	Mixed	PET/CT	Logistic regression	Pathological response to NACRT	ROC	MTV (AUC 0.74) and TLG (AUC 0.69) correlated with tumour regression at baseline PET. SUVmean (AUC 0.67) and TLG (AUC 0.64) related to higher chance of significant pathological response at second PET after induction chemotherapy
Cao et al., 2020	China	159	SCC	PET	LASSO Logistic regression	Treatment response following CCRT	AUROC	Validation set AUC for radiomics signature- based model was 0.835
Zhang et al., 2014	USA & China	20	Mixed	PET/CT	SVM & Logistic regression	Pathological tumour response to NACRT	AUROC (95% CI)	SVM combining classic PET/CT measures + clinical parameters + spatiotemporal PET features reached AUC of 1.0 vs 0.56 (0.07), 0.6 (0.06) and 0.94 (0.02) individually. SVM additionally outperformed LR (combined model AUC 0.9 (0.06)).
Qiu et al., 2020	China	206	SCC	CT	LASSO & Cox Proportional Hazards	Recurrence free survival following pCR post NACRT	Validation set C-Index (95% CI)	0.724 (0.696 – 0.752) with Radiomics + Clinical risk factors model vs Radiomics (0.671, 0.624 – 0.718) or clinical risk factors (0.629, 0.597 – 0.661)

Study	Country	Study size (n)	Histology	Imaging modality	ML techniques	Outcome Predicted	Model performance metric	Results
Yang et al., 2019	Taiwan	548	SCC	PET	18/34-layer CNN	1 year survival post-diagnosis	AUROC	AUC of 0.738. Patients predicted to expire at 1 year who survived had a lower 5-year survival than those predicted to survive 1 year (32.6% vs 50.5%, p <0.001) - Authors inferred that the CNN model also reflected aggressive tumour biology

SCC = Squamous Cell Carcinoma, MTV = metabolic tumour volume, TLG = total lesion glycolysis, SUV = standardized uptake value, PET = positron emission tomography, CT = computerised tomography, AUROC = area under receiver operator characteristic curve, LASSO = least absolute shrinkage and selection operator, SVM = support vector machine, pCR = pathological complete response, NACRT = neoadjuvant chemoradiotherapy, NACT = neoadjuvant chemotherapy, CCRT = concurrent chemoradiotherapy, RF = random forests, XGBoost = extreme gradient boosting, ANN = artificial neural network, CNN = convolutional neural network

1.1.1.1.1. Treatment response evaluation

It has long been appreciated that intra-tumour heterogeneity on cross sectional imaging is associated with aggressive tumour biology, and impaired treatment response in oesophageal cancer leading to many machine learning techniques being applied to this issue (94). As imaging is often one of the earliest potential sources of information on tumour biology for oesophageal cancer patients, and accurate characterisation here can tailor the oncological plan even before histology has been returned, this is a logical approach. Historically, most studies attempting to predict treatment response have focussed on neoadjuvant chemoradiotherapy rather than chemotherapy, often using oesophageal squamous cell carcinoma or mixed histology datasets (95–98). This is explained by the fact that many of these studies originate from China, where 90% of oesophageal cancers are the squamous subtype. This has unfortunately limited the relevance and utility of their findings in western populations.

Flurodeoxyglucose (^{18}F)-Positron Emission Tomography (FDG-PET) is used frequently in oesophageal cancer to assess for metastatic disease using uptake of FDG in metabolically active cells. Several studies have reported the use of Metabolic Tumour Volume (MTV) and Standardized Uptake Value (SUV) on FDG-PET as predictive (to variable degrees) of response to neoadjuvant chemoradiotherapy across serial imaging time-points as well holding prognostic significance for survival (93,99,100). One PET study taking inspiration from Deoxyribonucleic acid (DNA) microarray analysis combined a radiomics signature with a LASSO-logistic regression model to report an Area Under Curve (AUC) of 0.835 in predicting treatment response (101). While the authors contended with a class imbalance favouring responders and a radiomics signature derived from only 20 patients, their approach was nevertheless intriguing. However numerous drawbacks to using FDG-PET in this manner remain, including its expense, time-consumption, poor resolution and lack of the complete molecular characterization that is typically desirable when mining spatial heterogeneity within tissue architecture and metabolic activity (94). Contrast-enhanced Computed Tomography (CT) is comparatively ubiquitous in assessing treatment response, quick and readily accessible. In smaller case series it has even shown some success in predicted response to chemoradiotherapy using as few as five shape- and histogram-based metrics (AUC 0.686 – 0.727) (96). A weakness however is that while avidity on PET visually highlight regions of interest for suspicious tissue, grey-scale images on CT require human labelling.

Another recurrent theme within radiomic modelling studies continues to be the superiority shown when using multimodal datasets over single data streams. Zhang et al. predicted

pathological tumour response to chemoradiotherapy using both logistic regression (LR) and Support Vector Machine (SVM) models. They reported that a combination of conventional PET/CT response measures, clinical data (TNM staging, histology, patient demographics), and spatio-temporal PET/CT features offered superior predictive performance over individual feature sets (AUC of 1.0 for SVM vs 0.9 for LR) (102). While these results might seem impressive at first, the study did not account for nodal disease and lacked any statistical power (N= 20), making model over-fitting highly probable in the absence of any external validation data. Another study assessing treatment response to chemoradiotherapy in 97 patients also combines clinical information, geometry, PET textural features and CT textural features used a LASSO regularised regression model. They reported an AUC of 0.78 versus 0.58 when modelling SUVmax alone and while their sample size was limited, the study's internal validation procedures remained robust (94).

1.1.1.1.2. Prognostication

Accurate prognostication in oesophageal cancer is of obviously beneficial in optimising cancer care decisions as precise prognostication allows clinicians to quantify the cost-benefit analysis of treatment options to patients. Several studies have attempted to predict prognosis using machine learning models. Qiu et al. reported disease recurrence in one third of patients who experienced a pathological complete response following chemoradiotherapy and surgery for squamous cell carcinoma (103). Their CT-based nomogram combined clinical risk factors and a radiomic signature comprising eight features. This proved superior (Concordance (C) -Index of 0.746) versus either radiomic (0.685) or clinical (0.614) features alone ($p < 0.001$ in all cases) and could effectively stratify patients into high and low risk categories to help tailor adjuvant therapy post-resection.

One Dutch study predicting 3-year survival from pre-treatment CT features used a random-forest model to compare clinical and radiomic feature sets on (97). The authors reported an AUC of 0.61 on external validation for their radiomic model versus 0.62 for their clinical dataset. Despite a clear survival difference between Tumour Regression grade (TRG) 1-2 and TRG 3-5 patients within the study cohort the study did not show a statistically significant difference in survival within validation sets. This again echoes the disconnect seen between pathological response and survival reported in the NeoAegis trial (104).

Deep learning models using Convolutional Neural Networks (CNN) have also proved capable of predicting 1-year survival in squamous cell carcinomas using PET images. A Taiwanese group pre-trained a ResNet 3D CNN using a mixed dataset of 1,107 oesophageal squamous cell cancers and lung cancer (105). Their best model attained an AUC of 0.738, outperforming clinical data alone. The authors found that the CNN model's predictions were in and of themselves positive predictors for survival on multivariable analysis suggesting that prognostically significant hidden data could be extricated from the scans. The authors postulated that their model was able to identify indirectly more aggressive tumour biology based on their 1-year risk however lacked any cross-linked "-omics" data to test this hypothesis further.

1.1.7.2.3 Nodal status

The presence of lymph node (LN) disease carries significant implications for prognosis and potential treatment options. However, few studies have turned their attention to this aspect of oesophageal cancer. Tan and colleagues modelled a predictor of lymph node metastases in resectable squamous cell cancer patients with a test-set AUC of 0.773 using LASSO-Logistic regression, outperforming size criteria alone on CT imaging (98). Another CT-based study in 197 patients reported near-identical performance in testing using an elastic-net approach across what was implied to be a mixed histological cohort. While the study implemented multiple validation measures during and after model training, it did not indicate if it mitigated the noticeable class imbalance present in the patient cohort with a significant incidence of node-negative disease (106).

1.1.1.1.3. Other clinical outcomes targets

Less conventional radiomic-based problems have also been explored. For example, resectability was predicted in one study of 591 OSCC patients using a LASSO-enhanced dimensionality reduction technique which showed multivariable logistic regression (MLR) models to offer strong predictive performance (AUC 0.87, accuracy 0.86) (95). Another study in radio-genomics used CT imaging from 92 patients to help predict microRNA (miR)-1246 expression, a biomarker linked with prognostic significance in squamous cell carcinoma (107). In the study correlation analysis extracted image features correlating to miR-1246 levels after which linear regression separated patients into low- and high-expression groups. Unfortunately, while miR-1246 levels were significantly raised in Stage 2 disease, no difference

was seen between healthy controls and stage 1 disease, thereby limiting miR-1246's potential for screening.

This summarization of the state of the art to date demonstrates that ML has been trialled across several data modalities relevant in the MDT assessment of OC. The MDT itself however remains an unexplored domain which offers significant potential for ML application. The following section therefore outlines the aims and objectives of my thesis which sought to trial ML within the operational framework of the OC MDT.

1.2 Stakeholder engagement

The literature as described above demonstrates a technical benefit for applying Machine Learning to an MDT workflow however it would still rely on stakeholder engagement to drive long-term adoption and success. To this end, a qualitative assessment of the sentiment within the field was made in conjunction with this research to determine both how MDT personnel assimilate clinical data within oesophageal cancer MDTs as well as how they perceive the discipline of artificial intelligence support within that framework. The results of this work are included in Appendix F and represent an effort to include the human agents who would potentially interact with machine-learning driven digital solutions in the future. The results of the study highlighted many useful insights into how non-technical experts faced with potential interactions with AI might react and the type of safeguards they would wisht to see implemented in deployment. These insights do not relate solely to oesophageal cancer, and in fact provide essential, universal needs that apply to any form of medical AI. Common themes such as technical prowess of such AI tools, the rift between human intuition and skill versus cold machine logic, and how such tools would need to evolve to rise to changes in practice are all essential considerations throughout this work. A significant positive in the results of this preliminary work was the largely positive sentiment most respondents within the study felt towards the idea of AI support. It should be borne in mind however that there remains potential for reporting bias in view of the participant attrition during the study which could leave more ML-friendly individuals accounting for the greater proportion of responses. With this in mind, it would be incumbent on this research to demonstrate at all times that patient safety, reliability and equity are full-time considerations during model development and testing.

1.3 Research framework

1.3.1 Research Question/Hypothesis:

The overarching hypothesis of this research is that oesophageal cancer MDT treatment-decisions can be replicated through machine learning classification models with sufficient accuracy and explainability as to be useable for the semi-automation of MDT discussions within their current configuration. This in turn should eventually standardise decision-making (and by extension provide equality of healthcare) as well as improved efficiency and reduction in cognitive loading on clinicians should they choose to accept the machine's recommendation.

1.3.2 Aims & Objectives:

The aims and objectives for this research are split into Primary and Secondary modules as follows:

1.3.2.1 Primary aims and objectives.

1. **Pilot the concept of an MDT treatment classifier model by applying established machine learning algorithms to historic curative oesophageal cancer cases.**

   To train novel ML models capable of replicating the MDT's decision-making in this subset of patients, I will use cases extracted from of a prospectively managed oesophagectomy database. This dataset will comprise University Hospital Southampton oesophageal cancer MDT patients who have all undergone curative surgery with or without some form of neoadjuvant therapy – chemotherapy or chemoradiotherapy. The initial outcome classes would thus comprise:
   a. Surgery
   b. Neoadjuvant chemotherapy prior to surgery
   c. Neoadjuvant chemoradiotherapy prior to surgery

2. **Develop separate treatment classifier models for palliative oesophageal cancer patients who are defined by a separate set of management needs.**

This would be dictated by the level of success achieved with the curative treatment classifier model as a template pipeline for extension to the palliative domain. Currently no local palliative patient database is maintained or exists at University Hospital Southampton, this will require the creation of a brand-new de novo database of non-curative patients. As the nature of palliative treatment pathways typically focusses on oncological treatment without debulking surgery, this potentially lends itself to additional survival modelling at an early point in the patient's care pathway. The palliative treatment classifier model would comprise the following outcome classes:

a. Best Supportive Care
b. Palliative Chemotherapy
c. Palliative Radiotherapy (to the primary tumour specifically)
d. Palliative Stent
e. Palliative Stent + Chemotherapy/Radiotherapy

3. **Incorporate eXplainable AI (XAI) methodologies during this research to build and develop long-term clinician-trust within the ML models derived.**

The use of XAI methods is essential to building trust and willingness to engage/use AI-based models in the MDT setting. To demonstrate that these models do accurately represent the MDT I will also use XAI methods to produce insights into the logic behind predictions made by the treatment classifier models on both global and local levels. This would include:

a. Variable importance measures (Global)
b. Partial Dependence Analysis (Global)
c. Local Interpretable Model-Agnostic Explanations (LIME) (Local).

4. **Externally validate all derived ML models.**

While models trained on local data may perform sufficiently well within a local population, they remain vulnerable to potential overfitting and falsely high-performance metrics. To assess their generalisability, I will utilise internal validation methods during model training. The gold-standard test for generalisability however remains the

application of these models to data sourced from external units. I will therefore use data from another tertiary referral unit to assess the robustness of my models when provided data independent of the original training set.

1.3.3 Supplementary (non-core) aims and objectives.

The following aims while not anticipated to form the core narrative of this thesis will add value to the work by providing a clinically translatable vehicle through which to utilise the generated models and embedding the ethos of responsible research and innovation within the AI space.

1. **Develop a user-friendly interface with which to interact with the models to generate new predictions.**

   Once models are trained and validated, they will need a Human-AI interface or "tool" through which clinicians may interact with and make use of the trained models for new predictions. This will be attempted using a Shiny dashboard (Shiny is a package within the statistical computing program R which provides a web-application framework for data-visualisation and interaction with R-based computer models).

2. **Embed responsible research and innovation (RRI) within the development of MDT classifier models and any resulting MDT tool.**

   To ensure that AI-based technology is used, researched and implemented fairly, safely and with the key stakeholders central to the process, I will collaborate with researchers trained in qualitative methodologies to undertake national surveys, patient focus groups and clinician interviews.

   a. Expert consensus opinion from the Association of Upper GI Surgery (AUGIS), UK Acute Oncology Society (UKAOS), British Society of Gastroenterologists (BSG) and UK and Ireland Oesophagogastric Cancer Group (UKIOG) through a National Qualtrics Survey would provide expert input from a national base of MDT attendees and domain experts to help define the appetite for AI-based decision support as well as understand the human perspective in reaching these decisions with which to compare the machine output.
   b. Qualitative exploration of how key stakeholders would view our model and what key trust issues influence their acceptance through clinician interviews and

patient/patient relative focus groups. We will investigate what factors appeal or prevent clinicians adopting AI-based decision-support tools and how patients' feel about the use of AI in the streamlining of their management. This process will also help implement any resulting tool within current regulatory and legal frameworks.

1.4 The structural narrative of the thesis

In the beginning of this thesis, I have sought to demonstrate that there is already a body of evidence to support the use of Machine Learning techniques within clinical oncology and that some early work has also established a role for these techniques within oesophageal cancer (OC). Yet, as described previously within this chapter the MDT itself has largely been untouched and provided a natural and promising target.

The logical starting point was a pilot study in a limited high-quality dataset on curable patients to demonstrate proof of early principle. This is presented in Chapter 2, and has been published in the European Journal of Surgical Oncology (EJSO, DOI: 10.1016/j.ejso.2023.106986) (108). This paper introduced original work into the AI-MDT arena and allowed me to plant my flag in that research space. I demonstrated viable treatment plan models in a dataset focussed purely on oesophageal cancer curative treatment plans and provided an early introduction into explainability techniques by proving that age was significant to numerous OC treatment decision scenarios.

Once the technical concept of classifying curative OC treatments was borne out, I had the means to extend treatment planning into palliative cohorts. This required establishing a de novo palliative dataset which could then be used for model training. Chapter 3 outlines this study in which I developed sequential models (firstly to classify treatment) and then to estimate survival (factoring in the treatment type involved). The ML models were capable of successfully predicting palliative treatment pathways for OC patients and, meaningfully, the generated survival estimates were by nature tailored to the treating hospital which is important when counselling patients on their likely progress during and after treatment. The palliative models in this chapter exploit a clear gap in the literature a) for treatment decision support in palliative settings and b) OC specifically. The bulk of models published are generally aimed at prognostication with almost nothing on treatment recommendations. The benefit for this work is that within this gap in the market we can now offer palliative patients a “double-act” of models that firstly predict the likely initial treatment and then prognosticate based on that treatment. The palliative survival model was validated internally to approximately 12 months

post-diagnosis. The other key benefit of the random survival forests model in this study was the ability to generate and superimpose survival curves for alternative treatments with which to compare to the recommended option. While this will not represent a completely "individualised treatment effect" (i.e. personalised to the patient) it is tailored to the specific treating unit and thus significantly more personalised than national statistics which are typically the standard benchmark for counselling.

In Chapter 4 I take the opportunity to extend the explainability work I introduced towards the end of Chapter 2, the results of which study are published in Computers for Biology and Medicine (CBM DOI: 10.1016/j.compbiomed.2024.108978) (109). This study takes the concepts of variable importance (introduced in Chapter 2) as well as a technique called Partial Dependence Analysis (with the assistance and guidance of our collaborator in Texas, Assistant Professor Arya Farahi). Both techniques come under the umbrella of explainable AI (XAI) and represent global techniques (global techniques allow understanding of how our model structures its use of training variables, local techniques apply to understanding how the model reached an individual prediction). I wished to bridge the gap between clinicians and trustworthiness of Machine Learning models. If XAI techniques can open the "black box" and what we see is reflective of true clinical practice – then it would be much more likely for clinicians to trust these models enough to use them in practice. At that point we could then trial local explainability techniques to then explain how a given treatment recommendation is reached to further reassure them. Additionally, when trust is developed, we are also able to use the same techniques to highlight areas of aberrant, questionable or simply inconsistent decision-making. In this paper I applied it to one of our algorithms (random forests) and again found the role age continues to play in treatment decisions to be significant, with treatment probabilities looking very different between patients younger than 75 years and those older than 75 years. The study further highlighted heterogenous decision-making for cT2N0 oesophageal cancer patients reflecting ongoing controversy within this cohort. This paper's strength is in employing a known ML technique in a novel use case within OC to demonstrate that ML could be used both going forward to predict on new patients but also in reverse to audit and interrogate team-based decisions (a process the CRUK report discussed in Chapter 1 found there simply is no time for anymore within modern MDTs).

Thus far my work has confirmed technical viability of modelling MDT decisions and deriving logic insights from the models themselves. To move towards a long-term, viable translatable

tool, we required direct input from clinicians themselves regarding how they feel they make decisions in OC MDTs and how they view AI tools in the medical setting. This is necessary to introduce innovations which both met user needs and addressed user concerns. In Appendix F we present a qualitative analysis of a national survey of health care physicians within the UK who routinely attend and contribute to OC MDTs. This piece of work was done in collaboration with my co-first author Dr. Catherine Webb. The survey explored factors which guided clinicians with their current decision-making at MDT as well as discuss the respondents' sentiments towards AI and ML tools in general to establish key barriers to adoption and uptake. The respondents' weighting of clinical factors in their decision making was then compared to a random forest model. This allowed direct comparison between "human" perception and the "AI" on MDT treatment decisions. The results of this study offered key insights into how health care physicians perceive their own decision-making and importantly, the barriers they report which may prevent adopting AI-based Decision-support tools in the future. This work is currently under re-review with Computers in Biology and Medicine, with a decision in principle indicating willingness to accept the paper pending minor revisions.

The insights derived from the national survey proved invaluable in guiding the development of a trustable, useable decision-support tool, however, to fulfil this aim, I needed to be able to validate my ML models externally, demonstrate that they could generalise to new and unknown patients who were not part of the training process, and consider the impacts such AI innovations may pose to clinicians, patients and the wider society. This body of work is represented in Chapter 5 and has been published by The Lancet's eClinicalMedicine Journal. It represents the culmination of my doctoral thesis presenting the full width of my research activity through this project. The paper details the external validation of three separate models which, in summation, act as a complete clinical decision support system (CDSS) and summarises the process I have collaborated in to ensure that my CDSS aligns with principes of Responsible Research and Innovation. With regards to model validation, the first is a primary classification model trained using the same principles outlined in Chapter 2, triaging new patients to either a specific curative treatment plan (neoadjuvant chemotherapy + surgery, neoadjuvant chemoradiotherapy + surgery, surgery alone, endoscopic resection), or "palliative therapy". This "primary" model leverages excellent training data (approx. 1000 patients locally and validated on nearly 1000 patients externally) and is currently the only known ML-based MDT treatment recommendation system for OC. The second model is a bespoke palliative treatment classifier model (trained on approximately 440 patients locally and tested on 475 externally), which will predict a treatment pathway of: best supportive care, palliative chemotherapy,

palliative radiotherapy, palliative stent-only or stent + oncological adjunct (be it chemotherapy or radiotherapy). The third and final model is a palliative random survival forests model. The model is trained on the same variables as the palliative treatment classifier but also includes the planned treatment as an additional variable. This is particularly beneficial in clinical practice as it means our MDT prototype CDSS models can be "chained" sequentially drawing off almost all the same original user-inputs without needing to be re-inputted when interacting with each new model in the chain. I tested my local Southampton-trained models using data from Oxford (N=975). This allowed me to show that all three main models (primary model, palliative treatment classifier and palliative Survival) all generalise externally, representing the first-ever externally validated ML MDT model for oesophageal cancer MDT decisions. The RRI work incorporated into the CDSS development included introduction of the MDT prototype tool to the scientific community at-large as well as discussing targeted elements of the Responsible Research and Innovation process I have followed to co-design the tool (this included regular RRI workshops with interdisciplinary attendees, Patient and Public Involvement representatives as well as semi-structured interviews with expert clinicians).

This thesis therefore aims to establish a clear and coherent narrative from conception of modelling the MDT through each treatment cohort, the introduction of explainability into the tool, and finally externally validating the models, all in the context of a RRI program which considered performance, trustworthiness, transparency and risks of bias.

The following chapter thus details the first stage of this process where the initial attempt at generating simpler models from a baseline cohort of curative OC patients is described and establishes the rationale for the core ML algorithms used throughout this thesis.

Chapter 2 Machine learning to predict curative multidisciplinary team treatment decisions in oesophageal cancer

Journal: European Journal of Surgical Oncology, 2023, Impact Factor 3.5, CiteScore 6.4

Eur J Surg Oncol. 2023 Nov;49(11):106986. doi: 10.1016/j.ejso.2023.106986

Navamayooran Thavanesan[1], Indu Bodala[2], Zoë Walters[1], Sarvapali Ramchurn[2], *Timothy J Underwood[1], *Ganesh Vigneswaran[1]

[1] School of Cancer Sciences, Faculty of Medicine, University of Southampton

[2] School of Electronics and Computer Science, University of Southampton

*These authors are Joint Last Author for this manuscript

Corresponding Author: Navamayooran Thavanesan

Address: School of Cancer Sciences, Faculty of Medicine, University of Southampton, South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, UK, SO16 6YD

Email: N.Thavanesan@soton.ac.uk

ORCID ID
NT – 0000-0002-7127-9606

IB – 0000-0002-7547-2526

ZW – 0000-0002-1835-5868

SR – 0000-0001-9686-4302

TJU 0000-0001-9455-2188

GV 0000-0002-4115-428X

Twitter (TJU): @TimTheSurgeon

Twitter (GV): @ganesh_vignes

Funding Support Acknowledgement: NT receives a joint studentship from the Institute For Life Sciences (University of Southampton) and University Hospital Southampton

Conflicts of Interests to declare: None

Manuscript category: Original Article.

2.1 Acknowledgements

The study outlined within this chapter has previously been published in the European Journal of Surgical Oncology (108). While the scientific work presented here is my own, I wish to acknowledge my co-authors for their contributions.

Contributions:

1) **Navamayooran Thavanesan was involved in the conception of this work, primary data collection, primary data analysis, its drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy**
    - **NT performed the data collection, collation, cleaning and coding. He performed the coding for the ML models and model evaluation methods in R. He drafted the initial manuscript based on the results he generated and made amendments to the subsequent drafts based on feedback by his supervisory team who are among the co-authors on this paper. He undertook the submission process, received and acted on reviewer comments and performed the final submissions too.**
2) Indu Bodala was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
3) Zoe Walters was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
4) Sarvapali Ramchurn was involved in final approval
5) Timothy J Underwood was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
6) Ganesh Vigneswaran was involved providing a limited section of R code to assist the modelling process, and in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy

The CRediT taxonomy is as follows:

NT - conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing– review & editing.

IB - methodology, supervision, and writing– review & editing.

ZSW - funding acquisition, investigation, methodology, project administration, supervision, and writing– review & editing.

TJU - conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, and writing– review & editing.

GV - conceptualization, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

2.2 Abstract

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.

Retrospective complete-case analysis of oesophagectomy patients ± neoadjuvant chemotherapy (NACT) or chemoradiotherapy (NACRT) between 2010-2020. 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.

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 ($p < 0.05$).

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.

2.3 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 (110,111). Treatment decisions for OC cancer patients in the UK are managed by multidisciplinary teams (MDT) integrating healthcare expertise for shared decision-making (112). 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 (113). OC treatment decisions thus carry implications for patient quality of life (114). 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 (64,112,115,116).

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 (117,118). One Danish study reported clinical impact in as many as 60% of test cases on subsequent management because of MDT disagreement (69).

Predictive modelling to assist decision-making for OC patients has demonstrated excellent results when predicting survival post-surgery in OC patients (119,120). These studies have generally accessed both pre- 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 (63,121). Accurate predictive models would provide for consistent clinical assistive decision tools (CADT) capable of standardising 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 to provide model explainability within OC decision-making.

2.4 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.

2.4.1 Study cohort

All patients who underwent an oesophagectomy for oesophageal adenocarcinoma or oesophageal squamous cell carcinoma from 2010 - 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 (122). 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. 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.

2.4.2 Model development

2.4.2.1 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), Neoadjuvant chemotherapy + surgery (NACT+S), or Neoadjuvant chemoradiotherapy + surgery (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 2.1).

2.4.2.2 Machine learning algorithms

Four established ML algorithms were selected and implemented via the "caret" package; Multinomial Logistic Regression (MLR)(123), Random Forests (RF)(124) , Extreme Gradient Boost (XGB)(125) and Decision Tree (DT) analysis(126). 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).

2.4.2.3 Validation and model performance

All models were developed using nested cross-validation (CV) and optimised for accuracy. A 5x10 configuration was chosen (10-fold cross-validation within the inner loop with 5-fold outer loop). The Receiver Operator Characteristic (ROC) values for the best model from each outer fold (N = 5) were then averaged to generate a mean Area Under the ROC 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).

2.4.2.4 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 multinomial logistic regression, random forests and decision tree, and the xgb.importance() function (xgboost package) for the XGBoost model. Absolute values were scaled (0-100) to allow comparison between algorithms.

2.4.2.5 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.

2.5 Results

2.5.1 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 (Table 2.1).

Table 2.1 - 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	**"Chemo" (N = 172) (%)**	**"CRT" (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%)	

Pre-treatment variables	**"Chemo" (N = 172) (%)**	**"CRT" (N = 127) (%)**	**"Surgery" (N = 100) (%)**	**Total (N = 399) (%)**	**P Value**
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%)	

Pre-treatment variables	"Chemo" (N = 172) (%)	"CRT" (N = 127) (%)	"Surgery" (N = 100) (%)	Total (N = 399) (%)	P Value
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

2.5.2 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 (multinomial logistic regression (MLR) 0.865, random forests (RF) 0.859, XGBoost (XGB) 0.805, decision trees (DT) 0.802). All algorithms perform less confidently in predicting neoadjuvant chemoradiotherapy + surgery (NACRT+S) (MLR 0.772, RF 0.699, XGB 0.696, DT 0.651) and neoadjuvant chemotherapy + surgery (NACT+S) (MLR 0.704, RF 0.651, XGB 0.644, DT 0.704). Individual ROC curves for each algorithm are

illustrated in Figure 2.1 (additional ROC curves for models trained solely on adenocarcinoma are in Supplemental Figure 1).

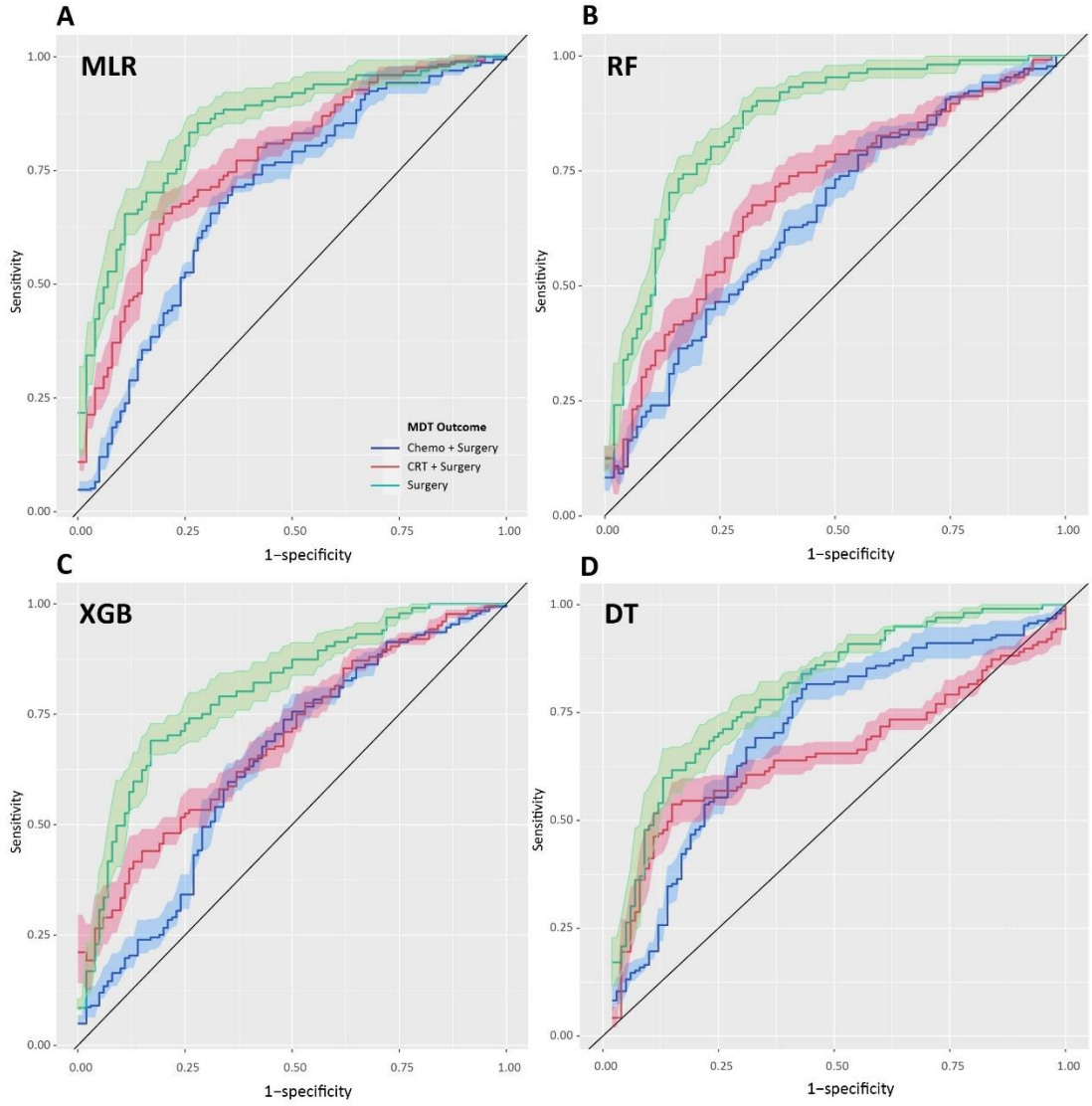


Figure 2.1 - ROC curve for averaged nested, cross-validated model performance given with +/- 1x standard error of the mean (SEM), A = Multinomial Logistic Regression, B = Random Forests, C = Extreme Gradient Boost and D = Decision Tree. AUROC = Area under Receiver Operator Characteristic.

2.5.3 Comparison of algorithms

Repeated, nested cross-validation was used to assess for statistical differences in area under ROC between algorithms (Supplemental Table 1). MLR outperformed RF and XGB on Kruskal-Wallis analysis when predicting neoadjuvant chemotherapy + surgery (P = <0.001) and neoadjuvant chemoradiotherapy + surgery (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 2.2). 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).

Table 2.2 - Mean performance metrics by algorithm (best performance metric in bold). Abbreviations – sd = Standard Deviation, AUC = Area Under Curve, PRAUC = Precision Recall AUC.

Model	Mean Balanced Accuracy (± sd)	Mean AUC (± sd)	Mean Kappa (± sd)	Mean LogLoss (± sd)	Mean F1 (± sd)	Mean PR AUC (± 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

2.5.4 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)

2.5.5 Variable importance

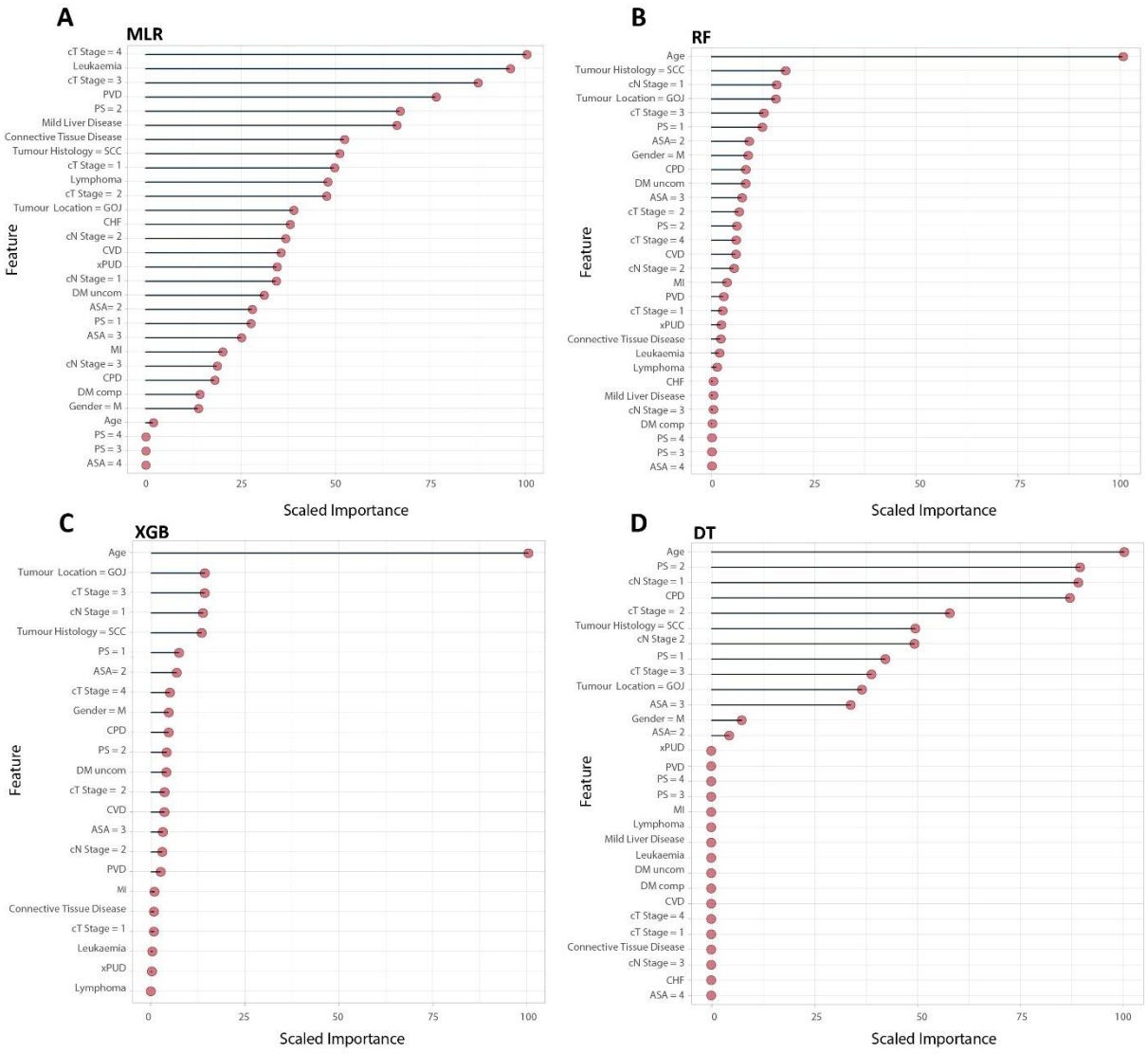


Figure 2.2 - Variable importance analysis for each trained algorithm. A - Multinomial Logistic Regression (MLR), B - Random Forests (RF), C - Extreme Gradient Boost (XGB), D - Decision Tree (DT).

Variable importance analysis highlighted factors critical to model formation (Figure 2.2). 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 Figure 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).

2.5.6 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 (Figure 2.3).

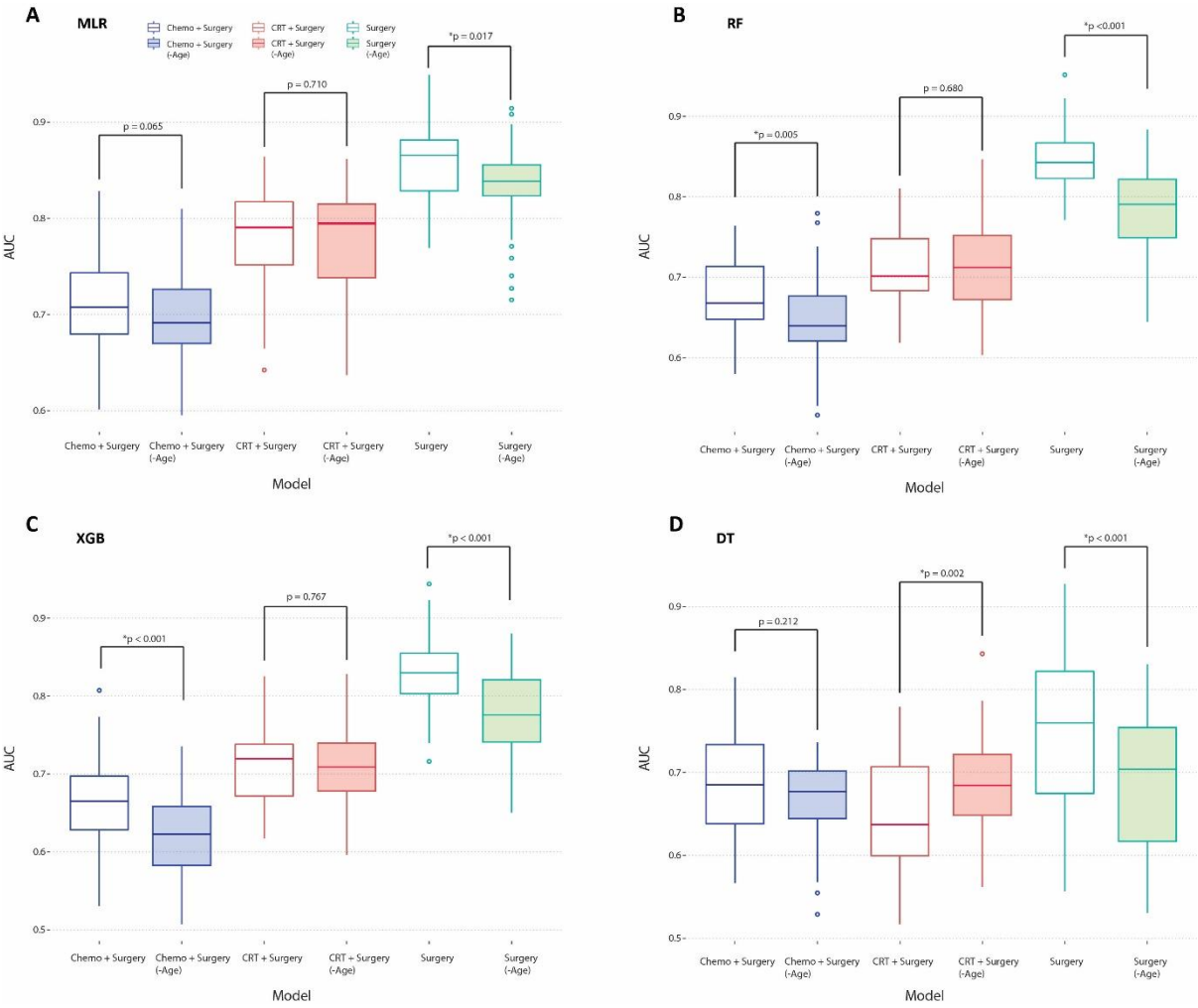


Figure 2.3 - Boxplot comparison of mean model AUCs for models with and without Age. MLR (A), RF (B), XGB (C) and DT (D). Significant P values denoted with and asterisk.

Difference in AUROC for all algorithms ± age were then compared statistically (Kruskal-Wallis test, P values provided in Figure 2.3). Across all algorithms, the removal of age as a predictor 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 = 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). The

removal of age as a predictor did not reduce predictive performance for 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 Figure 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).

2.6 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 electronic healthcare infrastructure. While multinomial logistic regression performed best, all models demonstrated good AUROCs and were confident discriminating between patients recommended surgery versus those offered neoadjuvant therapy (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 (9,127–129). 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 (104). 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 (130). 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 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 (131,132). 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 (14). 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 (133). This is problematic for deep-learning platforms which are inherently “black-box” solutions (134). 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 (135). Simple visual analysis of the scaled variable importance plots in Figure 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 Esophagectomy (TIME) and MIRO trials showed equivalence in survival benefits compared to open resection (136–140). Early Indications from the Randomised Oesophagectomy: Minimally Invasive or Open (ROMIO) study (31) 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 (141,142). 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.

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) (117,118). 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.

2.7 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 (143–146). 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.

2.8 Research in Context

This chapter outlined the efforts to test early proof of principle: that tabular MDT data could be leveraged towards ML models for treatment prediction. It was kept simple: three outcome classes and only curative patients. However naturally, the field of oesophageal cancer management is inevitably more complex and the majority of patients at presentation are advanced if not outright incurable. This cohort is poorly represented within the field when testing AI-based CDSSs and is a critically underserved population. The following chapter therefore addresses this gap in the literature by shifting the techniques described so far into a non-curative cohort. It establishes a de novo dataset for University Hospitals Southampton, derived, augmented and quality-checked from the National Oesophago-gastric Audit (NOGCA) combined with locally sourced data and significantly upscales the data pool this thesis uses to model with.

Chapter 3 Chained decision-support modelling combines treatment recommendation with treatment-related prognostication for palliative oesophageal cancer patients.

Navamayooran Thavanesan[1], Charlotte Parfitt[3], Saqib Rahman[1], Sam Luke Hill[1], Zoë Walters[1], Sarvapali Ramchurn[2], *Timothy J Underwood[1], *Ganesh Vigneswaran[1]

[1] School of Cancer Sciences, Faculty of Medicine, University of Southampton

[2] School of Electronics and Computer Science, University of Southampton

[3] University Hospitals Southampton, Department of General Surgery

*These authors are Joint Last Author for this manuscript

Corresponding Author: Navamayooran Thavanesan

Address: School of Cancer Sciences, Faculty of Medicine, University of Southampton, South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, UK, SO16 6YD

Email: N.Thavanesan@soton.ac.uk

ORCID ID

NT – 0000-0002-7127-9606

CP – 0000-0002-5486-6331

SAR – 0000-0001-9489-0021

SH – 0000-0002-9590-2116

ZW – 0000-0002-1835-5868

SR – 0000-0001-9686-4302

TJU 0000-0001-9455-2188

GV 0000-0002-4115-428X

Twitter (TJU): @TimTheSurgeon

Twitter (GV): @ganesh_vignes

Funding Support Acknowledgement: NT receives a joint studentship from the Institute For Life Sciences (University of Southampton) and University Hospital Southampton, with additional project funding from the UKRI Trustworthy Autonomous Systems Hub (TASHub) Pump-Priming grant

Conflicts of Interests to declare: None

Manuscript category: Original Article.

3.1 Acknowledgements

While the scientific work presented here is primarily my own, I wish to acknowledge my co-authors for their contributions.

Contributions:

1) **Navamayooran Thavanesan was involved in the conception of this work, primary data collection, primary data analysis, its drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy**
    - **NT performed data collection, collation, cleaning and coding. He performed the coding for the ML models and model evaluation methods in R. He drafted the initial manuscript based on the results he generated and made amendments to the subsequent drafts based on feedback by his supervisory team who are among the co-authors on this paper. He undertook the submission process, received and acted on reviewer comments and performed the final submissions too.**
2) Charlotte Parfitt assisted in data collection, and reviewing of the final manuscript.
3) Saqib Rahman provided coding assistance in R to assist NT with survival model training and internal validation. He was also involved in reviewing and recommending revisions for critical and important intellectual content, and agreement of accountability for accuracy
4) Sam Luke Hill was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
5) Zoe Walters was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
6) Sarvapali Ramchurn was involved in final approval
7) Timothy J Underwood was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
8) Ganesh Vigneswaran was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy

The CRediT taxonomy is as follows:

NT - conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing– review & editing.

CP - data curation, and writing– review & editing.

SAR - formal analysis, methodology, software, validation, visualisation, and writing– review & editing.

SLH - methodology, writing– review & editing.

ZSW - funding acquisition, project administration, supervision, and writing– review & editing.

TJU - funding acquisition, project administration, resources, supervision, and writing– review & editing.

GV - funding acquisition, project administration, supervision, visualization, and writing– review & editing.

3.2 Abstract

Palliative treatment plans dominate oesophageal cancer MDT caseloads. Leveraging ML offers a transformative approach to semi-automate and streamline this workflow. We present ML models which are trained to predict treatment decisions and provide prognostic insights for palliative Oesophageal cancer (OC) patients working towards personalised cancer care.

Using clinicopathological data from 437 palliative OC cases treated at a single tertiary centre over 12 years, we trained several ML algorithms (Multinomial Logistic Regression [MLR], Random Forests [RF], Extreme Gradient Boost [XGB], Decision Tree [DT], and Random Survival Forests [RSF]) to predict treatment pathways (best supportive care, chemotherapy, radiotherapy, palliative stent, or stent with an oncological adjunct) and survival prognoses. Model performance was evaluated using Area Under the Curve (AUC) for classifier models and calibration plots along with error rate (1-concordance) for the survival model.

Mean (±SD) AUCs for the classifier models were: MLR 0.801±0.090, RF 0.806±0.078, XGB 0.817±0.079 and DT 0.762±0.108. Mean error rate for the RSF survival model was 0.330±0.018 while calibration curves showed good calibration within the first 6-12months (median survival for the cohort was 6.31 months (range 0.1-105.8 months)).

This study represents the first use of ML to predict palliative treatment plans linked with treatment-related prognostication for OC patients. It offers significant potential to streamline MDT caseload and provide data-driven decision support for clinicians counselling their patients within the clinic room setting.

3.3 Introduction

Oesophageal cancer (OC) management has significantly evolved over the last five decades, yet overall prognosis remains poor. While 5-year survival for curative cases approaches 50% (147), this number is closer to 20% at one year for those with Stage 4 disease (148). Patients with incurable OC at diagnosis nonetheless account for 60% of presentations, requiring considered decision-making by MDTs to offer the best balance between prolonging survival and quality of life (32,34).

The perpetual increase in caseloads has been recognised as a major factor hindering the efficiency of Multidisciplinary teams (MDTs) in various specialities. The escalating workload has led to challenges such as, insufficient time to discuss patients, limited ability to focus on more complex cases, sub-optimal attendance, and heterogeneity within the clinical impact on treatment plans (26,149).

Clinical decision support tools are becoming increasingly commonplace within healthcare settings with notable examples including QRISK®, IC-RISC™ and QCancer among many others (35–37). However, tools to assist multidisciplinary teams (MDTs) with treatment-planning for palliative cancers early in the care pathway are historically overlooked, with the minority which have been trialled, focussed primarily on survival prediction (39). I have previously demonstrated that Machine Learning (ML) can successfully provide decision-support within oesophageal cancer when applied to the curative setting (108). However, to date few ML-based models exist for palliative oesophageal cancer patients with those that do exist again focussing on prognostication alone (150,151).

The purpose of this study was to develop and validate ML models capable of offering clinicians' informative decision-support for palliative OC patients going through MDT discussion. Our approach involved two distinct models, a classifier model to predict the likely palliative therapeutic pathway recommended by the MDT, and a second, survival model which is trained on a feature set which includes treatment pathway for prognostication purposes.

3.4 Methods

This study was a retrospective complete-case analysis of non-curable oesophageal cancer patients at a single tertiary referral centre (University Hospital Southampton) under the ethical approval of IRAS 233065 and 319540.

3.4.1 Study Cohort

The palliative cohort within this study were oesophageal cancer patients who were discussed and deemed unsuitable for curative management at MDT between 2010 – 2022. We established this cohort as a de novo dataset identified from unit submission records to the National Oesophagogastric Audit (NOGCA). Treatment decisions at our institution were made in line with National Institute for Health and Care Excellence (NICE) guidelines (122). Patients not suitable for curative management are typically offered one of five possible therapeutics outcomes: best supportive care (BSC), palliative chemotherapy ("Chemo"), palliative radiotherapy ("RTX", typically to either the primary tumour and/or symptomatic secondary sites amenable to radiotherapy, however for the purposes of this study, RTX was defined as therapy to the primary tumour), a palliative oesophageal stent alone, or in conjunction with an oncological adjunct (chemotherapy or radiotherapy).

Predictor variables for model training were again derived from clinicopathological data recorded routinely for patients discussed at the MDT. 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 (7^{th} edition until 2017 and 8^{th} edition thereafter). Novel molecular markers and immunotherapies which have been approved for metastatic disease within the UK since 2021 were not built into this first generation of palliative models as these are emerging treatments and consequently my unit has not yet accrued sufficient training data at present.

3.4.2 Data preparation and Analysis

I conducted data analysis, model training and model validation in R (version 4.2.2) with relevant packages described where first used. Excepting age and overall survival which were treated as continuous variables, the remaining covariates were treated as categorical. Two separate decision-assistance models were developed: classification models were trained to predict the palliative therapy assigned by MDT for a palliative patient, while a random survival forest model was trained for prognostication.

3.4.3 Survival Analysis

We used a Kaplan-Meier survival estimator ("survival" package) for preliminary survival analysis of the study cohort. Median survival time was stratified by treatment modality with a log-rank test-of-significance between curves. Overall survival was defined as survival from date of diagnosis to date of death or last recorded follow-up.

3.4.4 Machine Learning Algorithms

The methodology of training the MDT treatment-decision classifier has been described previously in Chapter 2 and was adopted again for this cohort (108). Multinomial Logistic Regression, Decision Tree, Random Forests and Extreme Gradient Boost models were trained using "caret", "nnet", "RandomForest", "xgboost" and "rpart" packages respectively (123–125,152). Survival modelling was undertaken using Random survival forests (RSF). We used an RSF model for this study as these have been previously proven to outperform traditional Cox Proportional Hazard models for prognostication in OC patients post-oesophagectomy (randomForestsSRC package) (73,153). The randomForestsSRC package by Ishwaran et al., is also computationally rapid, able to implement parallel processing and importantly generates predicted survival probabilities for every patient at every unique death time point within the training cohort providing significant granularity of prognostication which is invaluable when counselling patients on their anticipated clinical trajectory. Consequently, once a classification model produces a treatment recommendation, this prediction can then be re-inputted into the RSF model as a variable to provide survival probabilities factoring in that treatment, or an alternative pathway altogether.

3.4.5 Validation and Model Performance

Internal validation for the treatment classifier models was again using cross-validation with 5 outer folds and 10 inner folds (108). Mean-model performance assessed on area under the curve (AUC) for each outcome class. Additional metrics such as balanced accuracy, LogLoss, Kappa, Precision-Recall AUC (PR AUC) were also calculated for each classifier model using the resamples() function in the caret package.

Survival forests were internally validated using bootstrapped resampling to train 1000 forests (ntree = 1000 per forest) as per Rahman et al., (73) with hyperparameter tuning via the Tune() function to determine optimal nodesize and number of variables trialled per split ("mtry").

Mean-model performance was assessed across 3 methods: Error rate, Continuous Rank Probability Score (CRPS) and Calibration.

Error rate was defined as 1- Concordance (153). Here, concordance is the percentage of observation-pairs where the probability of a true event is greater than a true non-event (and thus a perfect model would have an error rate of 0%) (154). Error rate was extracted for each bootstrapped model and averaged across all time points.

The Continuous Rank Probability Score (CRPS), another measure of prediction calibration was averaged across all bootstrapped models. It is defined as Integrated Brier Score divided by time where the Brier score is the mean squared difference between the predicted probability and observed probability of an event, where the lower the score the better a model is calibrated (perfect accuracy = 0, perfect inaccuracy = 1) (153,155). As for the Brier score, a perfect model scores 0 and a perfectly inaccurate model scores 1.

Mean-model calibration curves were plotted both by survival probability quintile (cases were stratified using predicted 1-year survival), and by event probability at sequential time points (3,6,12,18 and 24 months, "pec" package). Survival curves were derived from mean test-set predictions (predicted probability) averaged at each time point across each bootstrapped model and plotted against the corresponding Kaplan Meier (observed) survival probability. Quintile-based plots allow for direct comparison of model predictions for patients ranked by survival probability at a specific time-point, while calibration curves at sequential time-points allow for comparison of model predictions across the whole cohort at a range of milestones. The combination thus offers a clearer insight to the optimal operating window for the model longitudinally and by patient-risk.

3.4.6 Variable Importance Analysis

Variable importance ranking for each predictor variable was derived from the classification models using the varImp() function ("caret" package). Variable importance for the final random survival forest model was by comparison extracted via in-built VIMP() function ("randomForestsSRC" package). The varImp() function expands variable-weighting to the sub-class level while the Random survival Forest (RSF) VIMP() function presents weights at the variable-level only and this is presented accordingly within their respective importance plots. Derived absolute values were scaled (0-100) to allow comparison between algorithms.

3.5 Results

3.5.1 Cohort Demographics

A cohort of 437 patients were used in this complete case analysis (Table 3.1). Median age was 75.2 years (range 29.8 – 96.7) with a male to female ratio of 2.7:1. Table 1 outlines the cohort demographics by treatment outcome.

Table 3.1 - Palliative cohort demographics

Pre-treatment variables	"BSC" (N =56) (%)	Chemotherapy (N = 148) (%)	Radiotherapy (N = 78) (%)	Stent (N = 113) (%)	Stent + Onc (N= 42) (%)	Total (N = 409) (%)
Gender						
Male	33 (58.9%)	119 (80.4%)	56 (71.8%)	80 (70.8%)	30 (71.4%)	318 (72.8%)
Female	23 (41.1%)	29 (19.6%)	22 (28.2%)	33 (27.4%)	12 (28.6%)	119 (27.2%)
Median Age in years (Range)	79.07 (40.27-94.27)	66.27 (29.78-87.15)	80.16 (51.21-96.71)	82.34 (55.47-95.59)	73.97 (49.81-94.16)	75.17 (29.78 – 96.71)
Performance status						
0	2 (3.6%)	49 (33.1%)	7 (9.0%)	10 (8.8%)	6 (14.3%)	74 (16.9%)
1	11 (19.6%)	64 (43.2%)	22 (28.2%)	17 (15.0%)	17 (40.5%)	131 (30.0%)
2	17 (30.4%)	31 (20.9%)	32 (41.0%)	44 (38.9%)	16 (38.1%)	140 (32.0%)
3	23 (41.1%)	4 (2.7%)	16 (20.5%)	41 (36.3%)	3 (7.1%)	87 (19.9%)
4	3 (5.4%)	0 (0%)	1 (1.3%)	1 (0.9%)	0 (0%)	5 (1.1%)
cT stage						
IS	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
0	0 (0%)	1 (0.7%)	0 (0%)	0 (0%)	0 (0%)	1 (0.2%)
1	1 (1.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.2%)
1a	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
1b	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2	3 (5.4%)	11 (7.4%)	11 (14.1%)	13 (11.5%)	2 (4.8%)	40 (9.2%)
3	19 (33.9%)	80 (54.1%)	46 (59.0%)	59 (52.2%)	23 (54.8%)	227 (51.9%)
4	12 (21.4%)	35 (23.6%)	16 (20.5%)	28 (24.8%)	11 (26.2%)	102 (23.3%)
4a	6 (10.7%)	14 (9.5%)	2 (2.6%	4 (3.5%)	2 (4.8%)	28 (6.4%)
4b	5 (8.9%)	3 (2.0%)	1 (1.3%)	2 (1.8%)	3 (7.1%)	14 (3.2%)
X	10 (17.6%)	4 (2.7%)	2 (2.6%)	7 (6.2%)	1 (2.4%)	24 (5.5%)

Pre-treatment variables	"BSC" (N =56) (%)	Chemotherapy (N = 148) (%)	Radiotherapy (N = 78) (%)	Stent (N = 113) (%)	Stent + Onc (N= 42) (%)	Total (N = 409) (%)
cN stage						
0	11 (19.6%)	16 (10.8%)	27 (34.6%)	31 (27.4%)	5 (11.9%)	90 (20.6%)
1	16 (28.6%)	46 (31.1%)	30 (38.5%)	33 (29.2%)	18 (42.96%)	143 (32.7%)
2	14 (25%)	52 (35.1%)	16 (20.5%)	32 (28.3%)	13 (31.0%)	127 (29.1%)
3	6 (10.7%)	31 (20.9%)	3 (3.8%)	13 (11.5%)	6 (14.3%)	59 (13.5%)
X	9 (16.1%)	3 (2.0%)	2 (2.6%)	4 (3.5%)	0 (0%)	18 (4.1%)
cM Stage						
0	22 (39.3%)	27 (18.2%)	54 (69.2%)	58 (51.3%)	15 (35.7%)	176 (40.3%)
1	30 (53.6%)	121 (81.8%)	24 (30.8%)	53 (46.9%)	27 (64.3%)	255 (58.4%)
X	4 (7.1%)	0 (0%)	0 (0%)	2 (1.8%)	0 (0%)	6 (1.4%)
Tumour location						
Oesophagus						
Proximal	4 (7.1%)	3 (2.0%)	9 (11.5%)	0 (0%)	2 (4.8%)	18 (4.1%)
Middle	8 (14.3%)	21 (14.2%)	13 (16.7%)	16 (14.1%)	5 (11.9%)	63 (14.4%)
Distal	30 (53.6%)	73 (49.3%)	48 (61.5%)	72 (63.7%)	29 (69.0%)	252 (57.7%)
GOJ						
GOJ Siewert 1	1 (1.8%)	7 (4.7%)	1 (1.3%)	10 (8.8%)	1 (2.4%)	20 (4.6%)
GOJ Siewert 2	7 (12.5%)	30 (20.2%)	5 (6.4%)	12 (10.6%)	2 (4.8%)	56 (12.8%)
GOJ Siewert 3	6 (10.7%)	14 (9.5%)	2 (2.6%)	3 (2.7%)	3 (7.1%)	28 (6.4%)
Tumour Histology						
Adenocarcinoma	44 (78.6%)	121 (81.8%)	45 (57.7%)	83 (73.5%)	29 (69.0%)	322 (73.7%)
Squamous Cell	12 (21.4%)	27 (18.2%)	33 (42.3%)	30 (26.5%)	13 (31.0%)	115 (26.3%)
Difficulty passing gastroscope and/or severe dysphagia						
Yes	17 (30.3%)	31 (20.9%)	23 (29.5%)	111 (98.2%)	40 (95.2%)	222 (50.8%)

Pre-treatment variables	"BSC" (N =56) (%)	Chemotherapy (N = 148) (%)	Radiotherapy (N = 78) (%)	Stent (N = 113) (%)	Stent + Onc (N= 42) (%)	Total (N = 409) (%)
Co-morbidities						
History of MI (MI)	5 (8.9%)	12 (8.1%)	10 (12.8%)	12 (10.6%)	3 (7.1%)	42 (9.6%)
Chronic heart failure (CHF)	3 (5.4%)	3 (2.0%)	5 (6.4%)	6 (5.3%)	2 (4.8%)	19 (4.3%)
Chronic pulmonary disease (CPD)	10 (17.9%)	10 (6.8%)	14 (17.9%)	16 (14.2%)	7 (16.7%)	57 (13.0%)
Connective tissue disease	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Peripheral vascular disease (PVD)	3 (5.4%)	6 (4.1%)	4 (5.1%)	8 (7.1%)	4 (9.5%)	25 (5.7%)
Cerebrovascular disease (CVD)	15 (26.8%)	18 (12.1%)	18 (23.1%)	29 (25.7%)	4 (9.5%)	84 (19.2%)
Dementia	5 (8.9%)	2 (1.4%)	0 (0%)	7 (6.2%)	0 (0%)	14 (3.2%)
History of Peptic Ulcer Disease (XPUD)	2 (3.6%)	4 (2.7%)	3 (3.8%)	5 (4.4%)	2 (4.8%)	16 (3.6%)
Uncomplicated diabetes (DM uncomp)	12 (21.4%)	17 (11.5%)	16 (20.5%)	21	5 (11.9%)	71 (16.2%)
Complicated diabetes (DM comp)	0 (0%)	1 (0.7%)	2 (2.6%)	0 (0%)	0 (0%)	3 (0.7%)
Leukaemia	0 (0%)	0 (0%)	1 (1.3%)	0 (0%)	0 (0%)	1 (0.2%)
Lymphoma	2 (3.6%)	0 (0%)	1 (1.3%)	2 (1.7%)	0 (0%)	5 (1.1%)
Mild liver disease	1 (1.8%)	1 (0.7%)	1 (1.3%)	1 (0.9%)	0 (0%)	4 (0.9%)
Hemiplegia	0 (0%)	1 (0.7%)	1 (1.3%)	1 (0.9%)	0 (0%)	3 (0.7%)
Renal failure	5 (8.9%)	1 (0.7%)	11 (14.1%)	15 (13.3%)	2 (4.8%)	34 (7.8%)
Severe dysphagia or difficulty passing gastroscope	17 (30.4%)	31 (20.9%)	23 (29.5%)	111 (98.2%)	40 (95.2%)	222(50.8%)

3.5.2 Palliative cohort survival

Kaplan-Meier (KM) survival analysis for the cohort is summarised in Table 3.2. Median Survival for the cohort was 6.3 months (range 0.1 – 105.8). Sub-group KM analysis showed the greatest overall survival was seen in patients who received palliative chemotherapy with a median survival of 11.1 months (95% CI 9.66-12.16). This was followed by the palliative radiotherapy

group (median 8.4 months (7.06-12.88)), the palliative stent + oncological therapy group (6.0 months (4.27-8.51)), palliative stent alone (3.9 months (3.12-4.21)) while patients receiving best supportive care alone survived a median 2.2 months (1.28-3.55). A significant difference in overall survival among these treatment groups was confirmed by the log-rank test (P <0.001). Kaplan-Meier survival curves by treatment group are illustrated in Figure 3.1.

Table 3.2 - Kaplan-Meier Median Survival analysis by treatment type

Treatment	N	Events	Median Survival (Months)	95% CI (months)
BSC	56	56	2.15	1.28-3.55
Chemotherapy	148	134	11.07	9.66-12.16
Radiotherapy	78	72	8.41	7.06-12.88
Stent	113	113	3.88	3.12-4.21
Stent + Oncological therapy	42	41	6.01	4.27-8.51

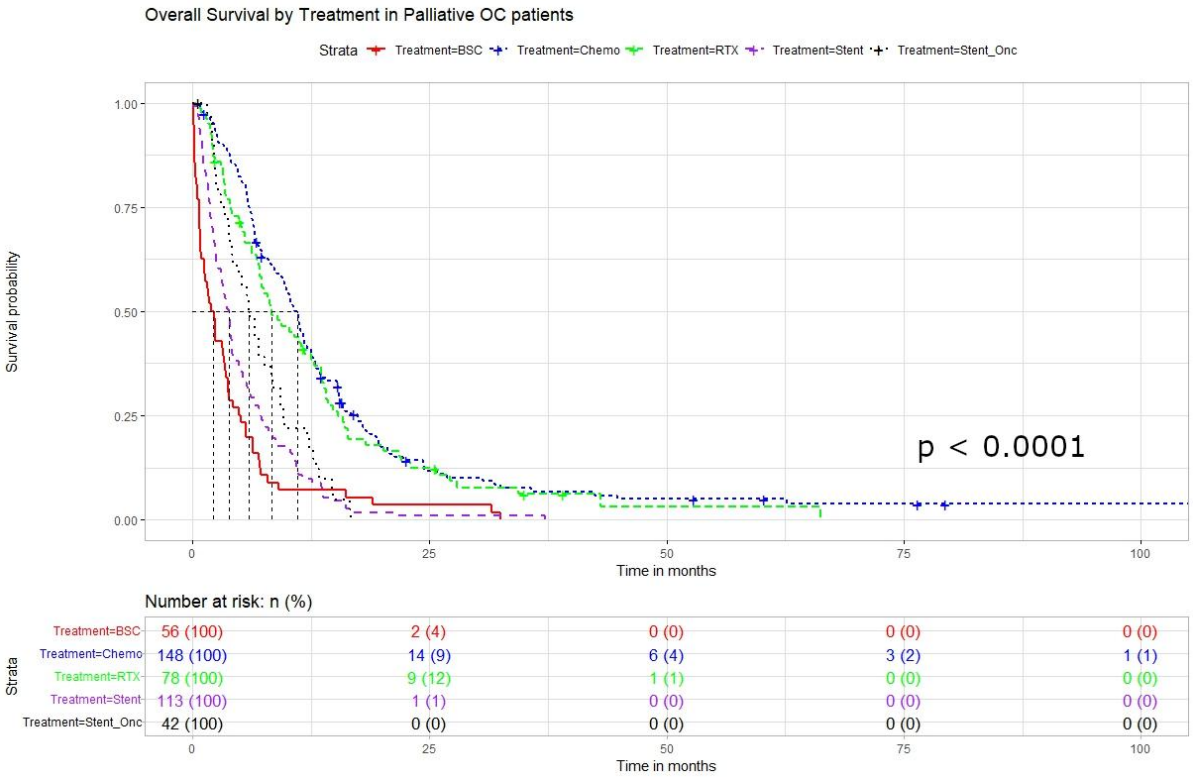


Figure 3.1 - Kaplan Meier survival curve for palliative cohort by treatment. Labels: BSC = Best Supportive care, Chemo = Palliative Chemotherapy, RTX = Palliative Radiotherapy, Stent_Onc = Stent + oncological adjunct.

3.5.3 Algorithm performance

Predictive performance for all four algorithms produced models with good mean AUCs (Multinomial Logistic Regression (MLR) 0.791, Random Forests (RF) 0.799, eXtreme Gradient Boost (XGB) 0.817 and Decision Tree (DT) 0.752). Across all performance metrics, the XGB model performed best for classifying MDT palliative treatment decisions (Table 3.3). Across algorithms, patients predicted for palliative chemotherapy were most reliably predicted, followed by the palliative stent-only, palliative stent with oncological therapy and palliative radiotherapy. Individual class Receiver Operator Characteristic (ROC) curves for all 4 classifiers can be found in Figure 3.2.

Table 3.3 - Palliative treatment classifier mean-model performance by ML Algorithm (MLR = Multinomial Logistic Regression, RF = Random Forests, XGB = Extreme Gradient Boost, DT = Decision Tree).

Palliative model resample metrics (N = 437)	Mean Balanced Accuracy (± sd)	Mean AUC (± sd)	Mean Kappa (± sd)	Mean LogLoss (± sd)	Mean PR AUC (± sd)
MLR	0.685 ± 0.012	0.791 ± 0.021	0.435 ± 0.028	1.227 ± 0.054	0.428 ± 0.023
RF	0.695 ± 0.016	0.799 ± 0.015	0.452 ± 0.031	1.318 ± 0.071	0.447 ± 0.021
XGB	**0.711 ± 0.001**	**0.817 ± 0.016**	**0.499 ± 0.016**	**1.064 ± 0.072**	**0.464 ± 0.017**
DT	0.678 ± 0.032	0.752 ± 0.021	0.434 ± 0.030	1.705 ± 0.413	0.126 ± 0.063

The survival model was trained using a random survival forest. The final model, trained on the full cohort after internal validation, demonstrated a mean prediction error of 0.323 and a CRPS of 0.068. On internal validation over 1000 bootstrapped models, mean error rate was 0.330±0.018 while mean CRPS was 0.112±0.019 (Table 3.4). Calibration curves (observed vs predicted survival probability) were stratified by 1-year survival quintiles (Figure 3.3) as well as by whole-cohort survival at defined time points (Figure 3.4). Calibration curve analysis suggested best model calibration within the first 12 months after which the model's prognostic accuracy diminished. Quintile-based analysis further indicated that best calibration was seen for the three highest-risk quintiles (Q1-3). Model predictions were ‘pessimistic’ for Q4 patients but ‘over-optimistic’ by Q5.

Table 3.4 - Survival model metrics with interpretation guidance

	Mean	SD	Reference	Interpretation
Mean Prediction Error (1-Concordance)	0.330	0.018	0 = perfect concordance 1 = perfect non-concordance	Good
CRPS (Integrated Brier Score/time)	0.112	0.019	0 = perfectly accurate model 1 = perfectly inaccurate model	Very Good

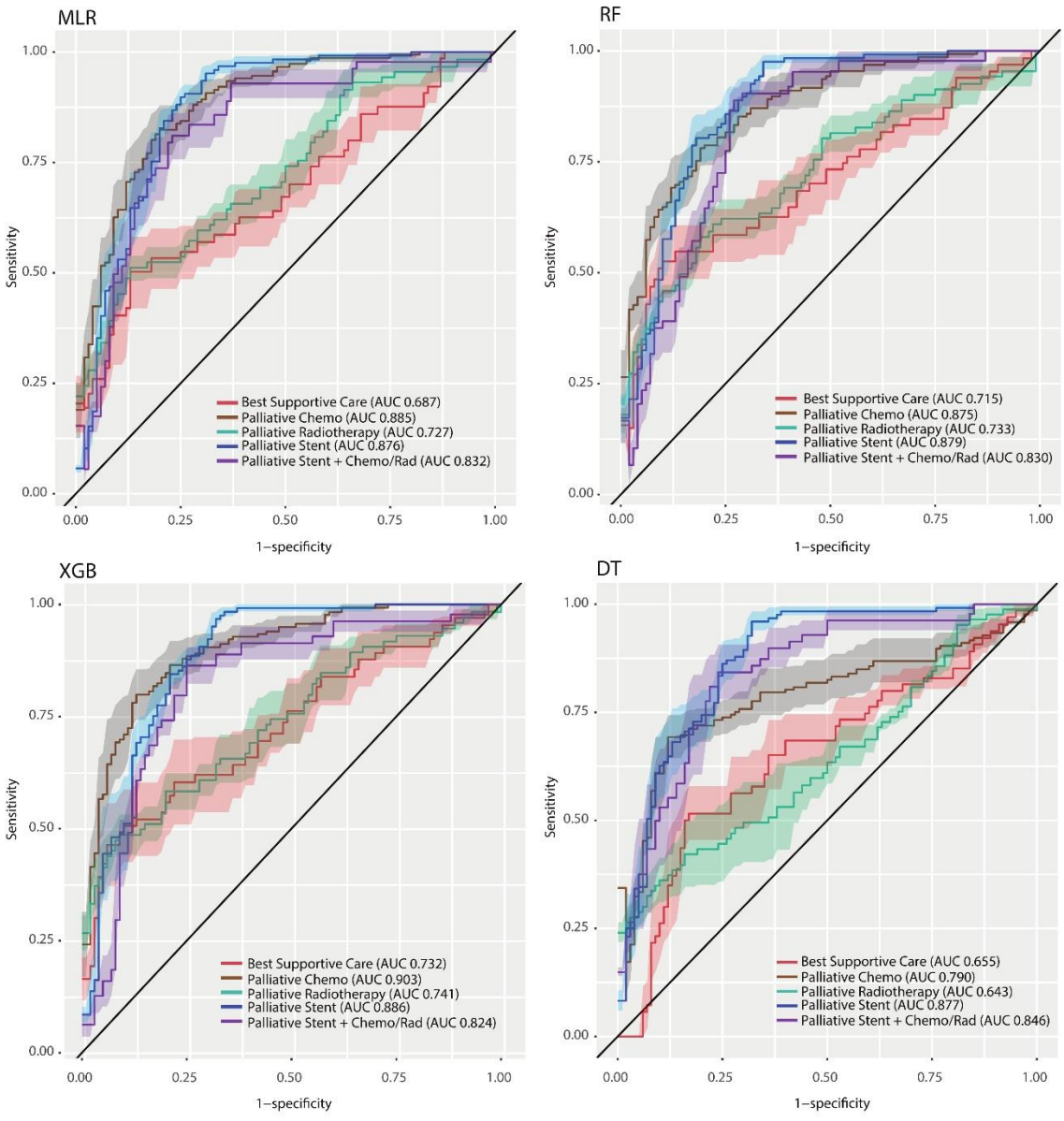


Figure 3.2 - Area Under Curve for treatment modality classification by algorithm (MLR = Multinomial Logistic Regression, RF = Random Forests, XGB = Extreme Gradient Boost, DT = Decision Tree)

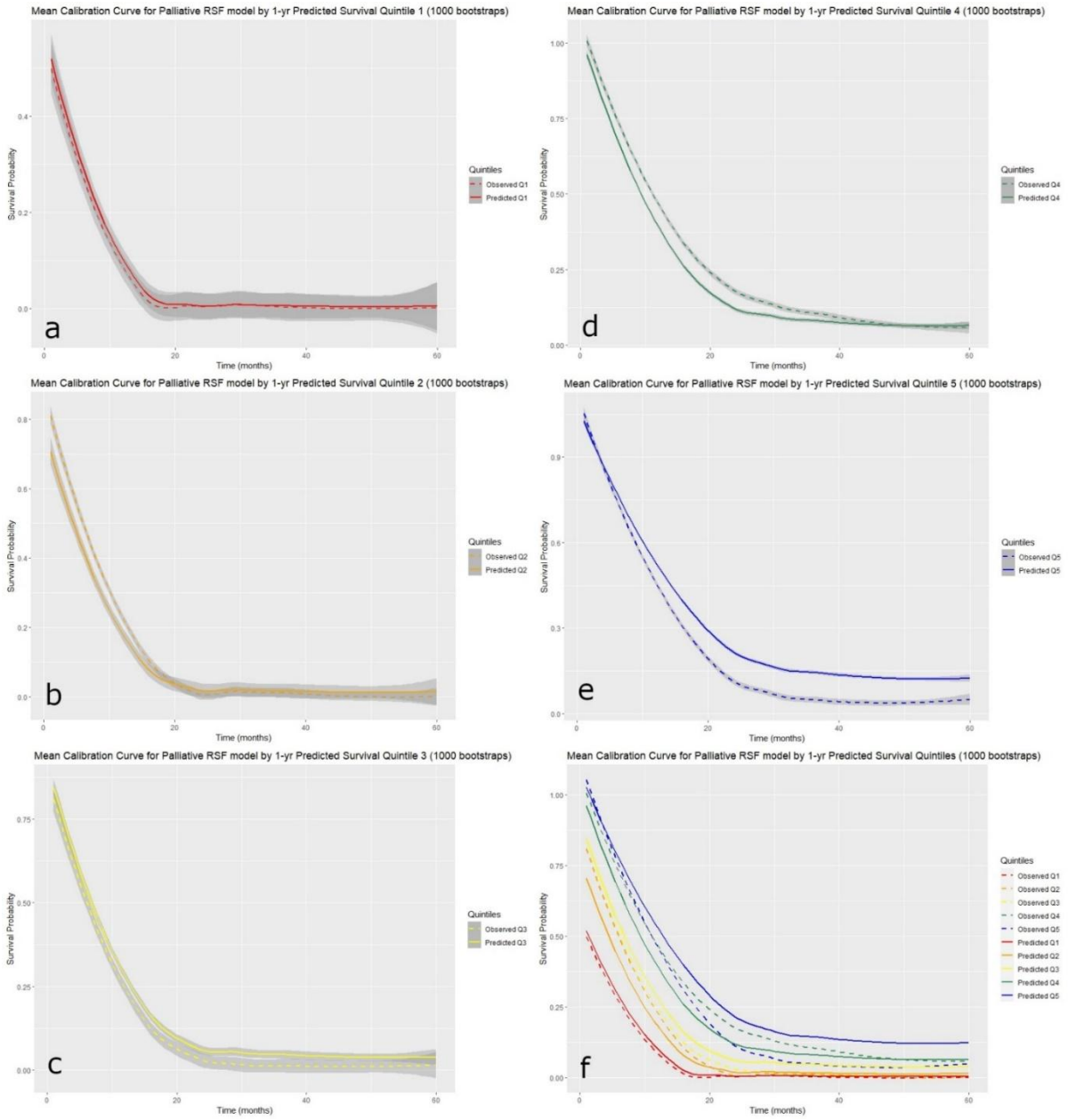


Figure 3.3 - Quintile Calibration curves plotted over 60 months with cases stratified by predicted 1-year survival probability (Quintile 1 = 0-20% (a), Quintile 2 = 20-40% (b), Quintile 3 = 40-60%(c), Quintile 4 = 60-80% (d), Quintile 5 = 80-80-100% (e) and all Quintiles (f). Solid lines are predicted survival probability by the RSF model, dashed lines are observed probability curves by the Kaplan-Meier estimator. Shaded errors represent standard error.

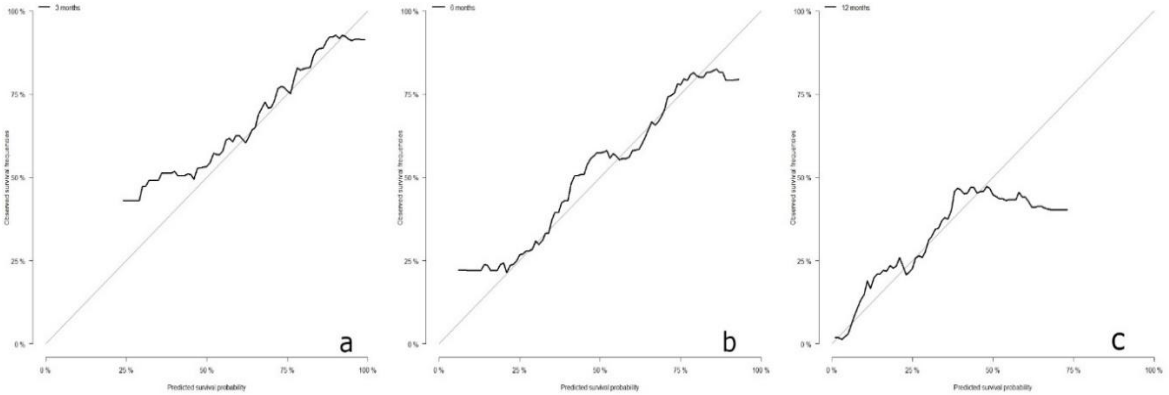


Figure 3.4 - Calibration curves for RSF model at 3months (a), 6 months (b), 12 months (c).

3.5.4 Variable importance

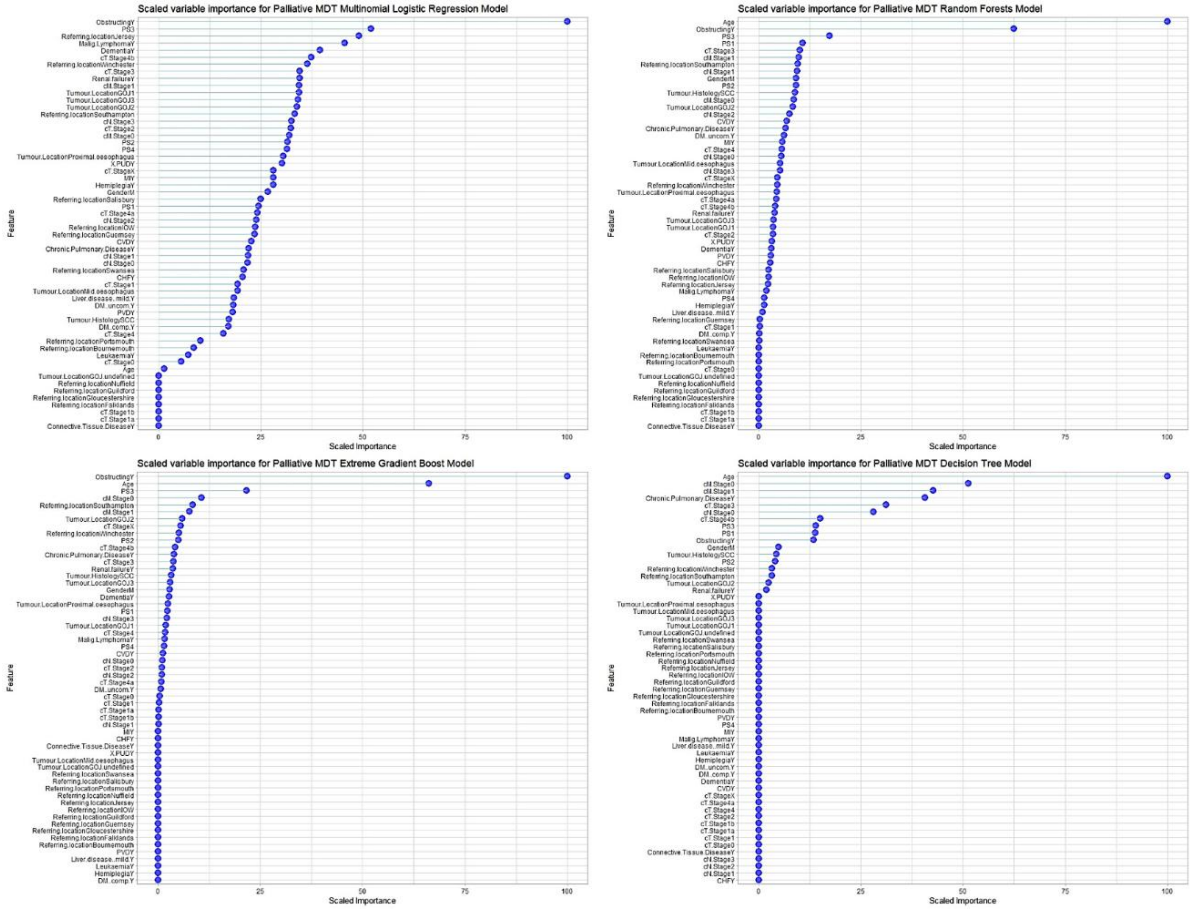


Figure 3.5 - Scaled variable importance by algorithm for treatment classifier models (MLR = Multinomial Logistic Regression, RF = Random Forests, XGB = Extreme Gradient Boost, DT = Decision Tree). Larger, high-resolution versions of this panel are available in supplemental figures 4-7

Variable importance analysis for each treatment classifier model is shown in Figure 3.5 and for the final Random Survival Forests (RSF) survival model in Figure 3.6 (for larger versions of Figure 3.5, please see Supplemental Figure 4, Supplemental Figure 5, Supplemental Figure 6, and Supplemental Figure 7). In the XGB and MLR models which were best performing, clinical or endoscopic signs of obstruction was most important to the models while Age ranked highly in all tree-based models. Treatment choice was most significant to the survival model followed by cT stage and performance status.

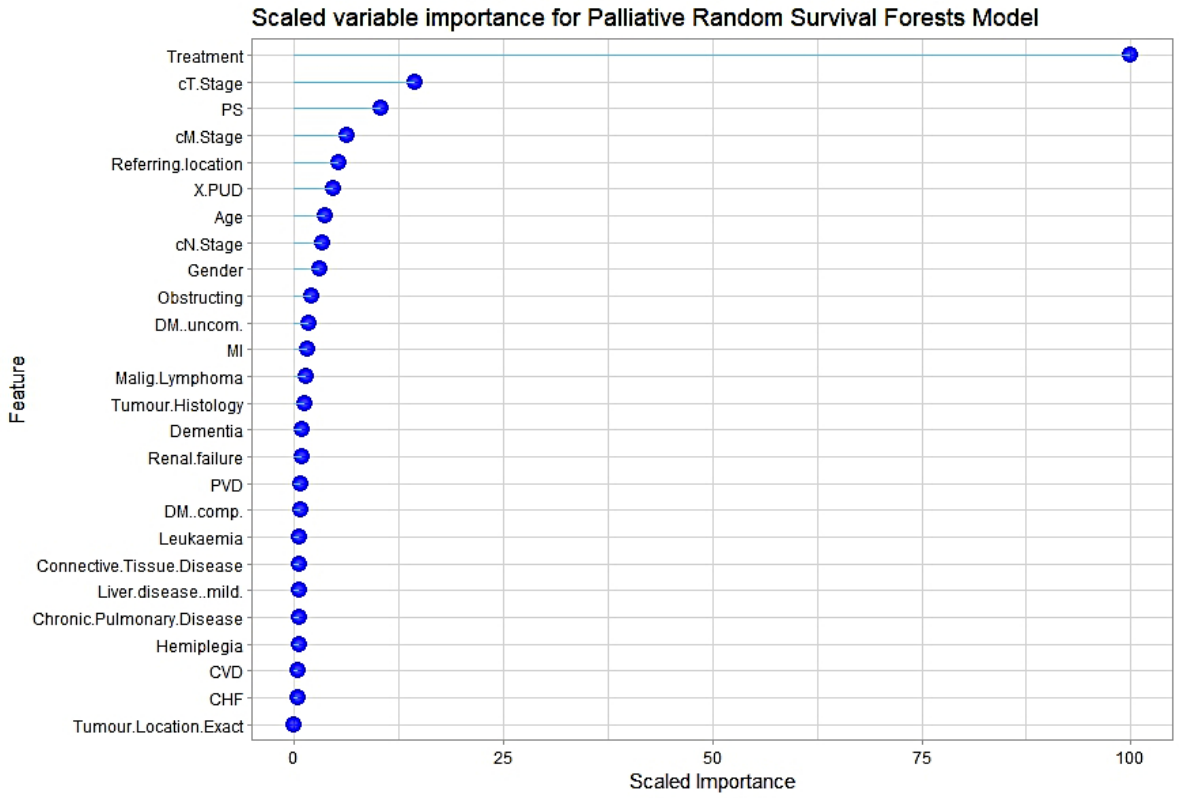


Figure 3.6 - Scaled variable importance for final RSF survival model.

3.6 Discussion

This study demonstrated that the methods I adopted in Chapter 5 using curative training cohorts could also be applied successfully to the palliative setting. All palliative treatment classifiers performed well irrespective of algorithm, with the XGB model performing best of all. The study also presents a prognostic RSF model predicting survival probabilities across a range of user-definable time-points. Calibration analysis confirmed that forecasting was most reliable over the first 12 months post-diagnosis, which is of material benefit for clinicians counselling these patients on their likely clinical trajectory where historically this has been based primarily on national staging-based statistics (148). As the survival model accounts for treatment modality, it is possible to compare prognosis for both the machine-recommended treatment but also for alternative treatment pathways, personalised to the unit-level based – offering data-driven counselling support for clinicians and their patients.

Across algorithms strong class separation (Area Under Curve (AUC) > 0.790) was achievable for palliative chemotherapy, palliative stent-only and stent + oncological adjunct prediction. Best Supportive Care (BSC) and palliative radiotherapy prediction were less confident independent of the algorithm, reflecting a combination of overall class-size limitation and less clearly defined criteria for patient selection within practice (122). Despite this, ensemble tree-based models performed (RF, XGB) strongly in the study as they typically tolerate class-imbalances well (156).

Variable importance analysis again demonstrated common themes: all tree-based classifier algorithms (DT, RF, XGB) highlighted the importance of Age and cM-stage, with XGB and RF, also recognising the importance of 'obstructive' clinical signs and Performance Status (PS) ≥ 3 to that final decision. As demonstrating non-inferiority versus the current paradigm is essential for translating this technology, these findings are reassuring in their consensus with National Institute for health and Care Excellence (NICE) guidance (NG83) which recommends stenting for luminal obstruction and relief of dysphagia, as well as combination chemotherapy in those with advanced oesophagogastric cancer, minimal comorbidity, and a PS score of 0-2 (122). Within the survival model, treatment modality proved most important. As the survival benefit here is likely a function of both patient selection and treatment effect, this further illustrates the benefit of non-linear ML-based methodologies which can account for these interactions (157).

The Random Survival Forests algorithm offered good predictive performance on bootstrapped internal validation within an appropriate time-period post diagnosis for palliative OC patients. This is realistic within the context of a cohort whose overall median survival was 6.3 months, and where only 6% of patients remain alive at 2 years. Kaplan-Meier analysis of our cohort highlighted a hierarchy in survival, with the greatest survival associated with palliative chemotherapy (11.1 months), commensurate with historically reported survival rates (158,159), followed by radiotherapy, stenting with palliative oncological therapy, stenting alone and finally best supportive care, again supporting current guideline recommendations and previous palliative therapy outcomes (157). While immunotherapy could not be factored into the present models owing to a paucity of training data, they can be readily included within future iterations once sufficient training data is accrued (160).

Considering the proportion of MDT decisions which relate to incurable patients, semi-automating palliative treatment decisions offers potential for cost-saving, standardized care and workflow efficiency (26,34). While a few studies have applied ML models towards gastric cancers to aid clinical management, these have mainly sought to prognosticate, comparing model performance against the arguably inadequate TNM staging benchmark (although Jiang et al did also look to predict likely benefit of adjuvant chemotherapy within their high and low risk cohorts) (161,162). ML models for oesophageal cancer remained survival-focussed usually in the context of a single pre-specified treatment modality. Liao et al., for instance analysed Surveillance, Epidemiology, and End Results (SEER) database data to determine a survival advantage in metastatic oesophageal adenocarcinoma who underwent palliative surgical interventions on Kaplan-Meier analysis. Their decision tree-based binary classifier predicted candidates likely to benefit from such palliative surgery with an AUC of 0.710 (150). Unfortunately, despite a large training cohort, they were unable to validate their data either internally or externally, curtailing the generalisability of their findings. Furthermore, as palliative resection within the UK is not current practice, the utility of such models is limited presently. A recent US study sought to predict patient-response to palliative chemotherapy in end-stage gastric and oesophageal cancers, testing several algorithms both at the beginning of therapy and again after 2 cycles on a binary classification task predicting response (151). They reported an average accuracy of 80% for 6-month survival prediction, rising to 85% after 2 cycles of chemotherapy irrespective of the algorithm used. These studies currently remain a minority within decision-support models being tested for palliative OC. The models I have presented here however seek to integrate ML both at the treatment-planning level and the patient-

counselling level and in doing so, seeks to optimise the decision-making process for clinicians and patients in tandem.

This study has natural limitations. Within a palliative cohort where gold standard therapy may at best offer a median survival of 11 months, prognostication beyond this time point is unlikely to be reliable. This in turn speaks to a second key challenge in the limited training cohort. Multi-centre datasets offer larger training sets but also introduce more variability within decision-making paradigms. Modelling over country-wide data may provide a generalisable model but ultimately lacks nuance, trading accuracy locally for generalisability nationally. Within the era of personalised medicine, I believe that a unit-based modelling approach offers clinicians the ability to counsel patients with prognostic data personalised to their own hospital rather than a generic inference derived from national cancer statistics. This is enhanced by the RSF model's ability to provide probable prognosis for not only the recommended treatment but also for alternative treatment pathways as a means of more thoroughly counselling patients.

It is important to note that the treatment classifiers we have trained in both this chapter and indeed Chapter 2 have mapped the current MDT rather than attempt to model the "best decision". No agreed-upon metric currently exists for such a concept within OC which can adequately encapsulate the myriad outcomes salient to these patients. For many patients, survival does not represent the most important outcome measure, yet it remains by far the most prolifically used to quantify treatment "success" for oncological strategies. This formed the rationale for using it in this study, and we intend it to be a starting point from which future models could springboard to other metrics such as quality of life and re-admission rate. For this technology to translate to clinical use, it must first prove capable of mapping what "is" while the field attempts to agree upon what "should be".

In order to maximise training data, we restricted predictors to those accurately and consistently available across the cohort during the study period. Immunotherapies and other novel systemic modalities did not feature within this generation of models as insufficient data was available to train with. In future we envisage these models as being able to support an expanding array of systemic treatments such as Chemotherapy +/- anti HER2, anti-PD-1/PD-L1, Claudin 18.2, and MMR-d/MSI-H (163–166). Nevertheless, we have established proof-of-principle for personalised unit-level ML-based decision-support in a subset of patients historically overlooked in decision-support resources despite accounting for the majority of MDT caseload (39).

3.7 Research in context

The research outlined within the preceding chapters focussed primarily on prediction generation. The ability to leverage tabular data to predict the treatment for the next patient seen in the MDT is core to this doctoral thesis, however as we see from the CRUK report, there is also a need to afford clinicians breathing space and bandwidth for auditing and re-considering the appropriateness of historic decisions too. The acceptability and explainability of MDT models will be a major consideration when integrating AI-driven decisions-support tools into healthcare where regulatory approval will almost certainly hinge upon explainable and interpretable solutions (167). This is problematic for advanced deep-learning platforms which are inherently "black-box" solutions (168). Linear models by comparison, such as MLR which has consistently proven effective in this use case represents one of the most interpretable options available. Decision-trees are also members of explainable AI (XAI), however, once the model training involves many hundreds of trees (RF and XGB-models) explainability becomes challenging, requiring post-hoc explainability methods (169). While the literature remains rife with the benefits of ML and AI for forward-prediction, there is a paucity on utilising ML techniques to provide the human agents insights into their decision-making paradigm. The field of explainable AI (XAI) offers great benefit in this regard, not just for regulatory transparency but for deriving actionable intelligence into past trends and team-based dynamics by leveraging the models themselves as MDT microcosms from which to derive idiosyncratic insights. The following chapter focusses on the utilization of ML techniques to extract these insights, demonstrating how potential treatment allocation biases may be identified, and identifying areas of variability in decision-making for OC patients and demonstrating that AI in this domain need not function solely as prediction engines.

Chapter 4 Insights from explainable AI in oesophageal cancer team decisions.

Journal: Computers in Biology and Medicine July 2024, Impact Factor 7.0, CiteScore 11.7

Comput Biol Med. 2024 Sep;180:108978. doi: 10.1016/j.compbiomed.2024.108978

Navamayooran Thavanesan[1], Arya Farahi[2], Charlotte Parfitt[3], Zehor Belkhatir[4], Tayyaba Azim[4], Elvira Perez Vallejos[5], Zoë Walters[1], Sarvapali Ramchurn[4], Timothy J Underwood[1], Ganesh Vigneswaran[1]

[1] School of Cancer Sciences, Faculty of Medicine, University of Southampton

[2] Department of Statistics and Data Science, University of Texas at Austin

[3] University Hospitals Southampton NHS Foundation Trust.

[4] School of Electronics and Computer Science, University of Southampton

[5] School of Computer Science, Horizon Digital Economy Research, University of Nottingham

Corresponding Author: Navamayooran Thavanesan

Address: School of Cancer Sciences, Faculty of Medicine, University of Southampton, South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, UK, SO16 6YD

Email: N.Thavanesan@soton.ac.uk

ORCID ID

NT – 0000-0002-7127-9606

EPV – 0000-0002-7547-9440

AF – 0000-0003-0777-4618

ZW – 0000-0002-1835-5868

CP – 0000-0002-5486-6331

SR – 0000-0001-9686-4302

ZB – 0000-0001-7277-3895

TJU – 0000-0001-9455-2188

TA – 0000-0002-6940-7955

GV – 0000-0002-4115-428X

Twitter (TJU): @TimTheSurgeon

Twitter (GV): @ganesh_vignes

Funding Support Acknowledgement: NT receives a joint studentship from the Institute for Life Sciences (University of Southampton) and University Hospital Southampton. The project receives funding from the UKRI Trustworthy Autonomous Systems Hub (TAS Hub) Pump Priming Fund.

Conflicts of Interests to declare: None

Manuscript category: Original Article.

4.1 Acknowledgements

The study outlined within this chapter has previously been published in the journal: Computers in Biology and Medicine (109). I wish to acknowledge my co-authors for their contributions.

Contributions:

1) **Navamayooran Thavanesan was involved in the conception of this work, primary data collection, primary data analysis, its drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy**
    - **NT performed the data collection, collation, cleaning and coding. He performed coding for the initial ML models and model evaluation methods in R as well as the variable importance plots which were then repeated in python by Ganesh Vigneswaran for interval verification. He drafted the initial manuscript based on the results he generated and made amendments to the subsequent drafts based on feedback by his co-authors on this paper. He undertook the submission process, received and acted on reviewer comments and performed the final submissions too.**
2) Arya Farahi was involved in methodology guidance for the PDP analysis process, data analysis advice, reviewing and recommending revisions for critical and important intellectual content and final approval
3) Charlotte Parfitt was involved in primary data collection, and final approval
4) Zehor Belkhatir was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval
5) Tayyaba Azim was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval
6) Elvira Perez Vallejos was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval
7) Zoe Walters was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
8) Sarvapali Ramchurn was involved in the final approval

9) Timothy J Underwood was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
10) Ganesh Vigneswaran provided python coding for ML models (to verify initial models in R), partial dependence plots (PDPs), the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy. I would like to specifically thank Dr Ganesh Vigneswaran for his assistance with the python-based Partial Dependence analyses and associated figures presented within this chapter.

The CRediT taxonomy is as follows:

NT - conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing– review & editing.

AF - methodology, supervision, and writing– review & editing.

CP - data curation, and writing– review & editing.

ZB - investigation, methodology, supervision, and writing– review & editing.

TA - investigation, methodology, supervision, and writing– review & editing.

EVP - supervision, and writing– review & editing.

ZSW - funding acquisition, project administration, supervision, and writing– review & editing.

SR - methodology, resources, supervision, and writing– review & editing.

TJU - funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft, and writing– review & editing.

GV - conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

4.2 Abstract

Clinician-led quality control into oncological decision-making is crucial for optimising patient care. Explainable artificial intelligence (XAI) techniques provide data-driven approaches to unravel how clinical variables influence this decision-making. We applied global XAI techniques to examine the impact of key clinical decision-drivers when mapped by a machine learning (ML) model, on the likelihood of receiving different oesophageal cancer (OC) treatment modalities by the multidisciplinary team (MDT).

Retrospective analysis of 893 OC patients managed between 2010-2022 at our tertiary unit, used a random forests (RF) classifier to predict four possible treatment pathways as determined by the MDT: neoadjuvant chemotherapy followed by surgery (NACT+S), neoadjuvant chemoradiotherapy followed by surgery (NACRT+S), surgery-alone, and palliative management. Variable importance and partial dependence (PD) analyses then examined the influence of targeted high-ranking clinical variables within the ML model on treatment decisions as a surrogate model of the MDT decision-making dynamic.

Amongst guideline-variables known to determine treatments, such as Tumour-Node-Metastasis (TNM) staging, age also proved highly important to the RF model (16.1% of total importance) on variable importance analysis. PD subsequently revealed that predicted probabilities for all treatment modalities change significantly after 75 years ($p < 0.001$). Likelihood of surgery-alone and palliative therapies increased for patients aged 75-85yrs but lowered for NACT/NACRT. Performance status divided patients into two clusters which influenced all predicted outcomes in conjunction with age.

XAI techniques delineate the relationship between clinical factors and OC treatment decisions. These techniques identify advanced age as heavily influencing decisions based on our model with a greater role in patients with specific tumour characteristics. This study methodology provides the means for exploring conscious/subconscious bias and interrogating inconsistencies in team-based decision-making within the era of AI-driven decision support.

4.3 Introduction

As with all cancers managed within the UK, Oesophageal cancer (OC) treatment plans are determined by a multidisciplinary team (MDT). Since their introduction in the mid-1990s, they have been shown to improve cancer outcomes, especially within OC, which remains the 6th leading cause of cancer-related death globally and is still characterised by dismal 5 & 10-year survival rates (148,170). With MDTs, the incidence of futile surgical procedures, operative mortality, and incomplete disease burden assessment (“cancer staging”) dropped significantly (19–21). However, this same framework which centralises a diverse group of domain experts in a single place and time is also vulnerable to challenges stemming from increased caseload pressure, reduced preparation time, interpersonal dynamics. Perhaps most importantly they continue to experience inadequate time for reflection or self-audit for the decisions they make, thereby limiting experiential growth, and in some cancer types leading to a growing pursuit to pre-select MDT cases by complexity as a means of improving workflow (23,26,66,171–175).

Objective, data-driven insight into oncological decision-making allows clinicians to interrogate, validate and ensure the appropriateness of their treatment choices over time which directly impacts patient outcomes and quality of life, something exemplified in OC (18). OC treatment decisions are highly complex; heavily influenced by primary tumour characteristics, metastatic spread and the physiological robustness of the patient (122,176). Peak incidence is between 85-89 years, with this cohort often experiencing polypharmacy, poor nutrition, frailty and disability, all of which impact clinical outcomes (177,178). Almost 80% of patients over the age of 85 years have two or more co-morbidities (increasing co-morbidity is known to be a negative prognostic marker of 90-day mortality post-surgical resection (179)) leading to age historically acting as a barometer of perceived risk for intensive therapeutic interventions (180–183). Judicious patient selection is critical; surgery alone is a monumental physiological stressor, further compounded by toxicity associated with neoadjuvant therapies (NAT), while even eligibility for palliative oncological therapies necessitates significant physiological reserve (14).

Additionally, while it is established that treatment decisions should be based on "physiological" over "chronological" age (184), it has recently been shown through ML that age plays a disproportionate role in treatment choice for curative OC patients at MDT. This bias is particularly evident when determining eligibility for multimodal versus unimodal therapy even when chronological age is not necessarily a guarantee of a negative outcome (185,186). It is not yet clear whether this is a conscious or unconscious bias nor if there is interplay between age and a patient's performance status (an oncological surrogate measure of baseline physical activity and thus a marker of resilience to otherwise deconditioning therapies). Implicit bias is a recognised aspect of healthcare, and while such bias has been reported for OC treatment allocation based on gender, race and socioeconomic status previously, how it manifests within more clinical parameters is currently unknown (187–190).

OC decision-making clearly carries high stakes, and yet while many of the clinical variables considered at MDT may be known or derived from guidelines, experience and current oncological doctrine (122,191,192), the relative weighting of these factors within the final decision is not currently known. This is salient when we consider the well-established literature surrounding vulnerabilities of cancer MDTs to inefficiency and sub-optimal decision-making in surgical oncology (23,66,171,172,175,193).

Machine Learning (ML), a branch of Artificial Intelligence (AI), is rapidly evolving within this aspect of healthcare, offering huge potential in multiple avenues relevant to OC. ML techniques can characterise complex patterns within current decision-making paradigms, inform future

decision-making within human-AI and Group-AI collaborative (HAIC) processes, theoretically transforming multi-disciplinary team (MDT) efficiency (63,194,195). Over the last decade, AI-based decision-support has also developed within MDT-type use cases with a view to changing the narrative from one of "human-versus-AI" to "human-and-AI" (196). The architectures being tested within oncology have ranged from traditional tree-based ML models and neural networks, through complex natural-language decision-support systems aiming to assimilate up-to-date clinical knowledge such as IBM's Watson, to more recently still, conversation-style, Large Language Model-based (LLM) architectures such as ChatGPT (39,197–199). This utility of AI however must be balanced with sufficient transparency and explainability to preserve clinician-AI trust within the recommendations and insights generated (55,59,200).

Within OC there clearly remains a research gap in how clinicians routinely utilise clinical variables in for oncological decision-making. The aim of this study was therefore to demonstrate a viable approach to leveraging eXplainable AI (XAI) in order to characterise in-detail, the influence these clinical variables exert (of which some may have subconscious impact) on OC treatment decisions. Combining explainable ML techniques, our goal is to offer clinicians a clearer perspective into decision-making variation for OC patients in a trustworthy and explainable fashion. This in turn sets the foundations for trust in future Human-AI collaborations within the inevitable clinical decision-support space and represents a novel application of XAI in OC surgical oncology to date.

4.4 Methods

This study was a retrospective complete-case analysis of oesophageal cancer patients at a single specialist cancer centre (University Hospitals Southampton) under the ethical approval of IRAS 233065 and 319540.

4.4.1 Patient Selection and Data Collection

OC patients who underwent MDT discussion from 2010 - 2022 were identified from a prospectively maintained oesophagectomy database combined with unit-submission records for the UK National Oesophagogastric Audit (NOGCA). Patients selected underwent either a curative pathway (surgery +/- NAT) or a non-curative (palliative) pathway (best supportive care, palliative stenting, palliative chemotherapy, palliative radiotherapy or a combination thereof).

Definitive chemoradiotherapy was excluded as this strategy occurred too infrequently for adequate model training. 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.

4.4.2 Statistical Analysis

Data analyses and model training were conducted using R (version 4.2.2) and Python (version 3.10.11). Sub-group comparison of continuous variables was made by Kruskal-Wallis analysis (adjusted with the Benjamini-Hochberg correction).

4.4.3 Data pre-processing and feature selection

Clinicopathological data within this study were analysed as structured tabular data. 'Label encoder' was employed within python to encode categorical variables for analysis. Features were selected through a combination of a priori domain expertise and established features form current UK clinical guidelines for oesophageal cancer management (122,179,186,191,192,201).

4.4.4 Treatment classifier model development and performance

MDT treatment-decisions were modelled using a Random forests (RF) classifier in Python ("Ranger" Library, sklearn v1.2.2) using variables consistently available to the MDT prior to a final treatment decision (Table 4.1 & Supplemental Table 3). Using k=5 cross validation, optimal max depth was determined as 6 which was used to train the final model on the whole dataset. The remaining hyper-parameters were set as default as RF models are not sensitive to small variations in these. The Random Forests algorithm is well-established and capable of handling higher-order interactions within classification tasks using both numerical and categorical features to produce strong predictive performance (108,124). It has been utilised in numerous healthcare settings (73,202,203) and has already been shown to perform well in classification tasks as related to MDT treatment plans (108). As this pilot study aimed to test whether XAI techniques could enhance complex decision-making processes, unlike linear models, random forest models can capture interactions, non-linear relationships, is recognisable and accessible for the analysis and future reproducibility.

Year of diagnosis was incorporated into model training within defined time-periods (termed "Epoch" for the purposes of this study) relative to the dissemination of key randomised clinical trials to account for, (and assess changes in) clinical practice over time. Treatment outcomes were classified into neoadjuvant chemotherapy prior to surgery (NACT+S), neoadjuvant chemoradiotherapy prior to surgery (NACRT+S), surgery-only (Surgery) or palliative therapy (Palliative). Model performance was assessed via multi-class area-under-the-curve (AUC)/Receiver Operator Characteristic (ROC), balanced accuracy and calibration. For evaluating initial model-generalizability we used a 5-fold cross validated approach, following which hyperparameters were fixed allowing for training on the whole dataset. This preserves statistical power during partial dependence analysis as we discover what the model has learned (204,205). For assessment of different algorithmic performances see Thavanesan et al., 2023 (108).

4.4.5 Variable Importance Analysis

Variable importance analysis of the final model was undertaken using all variables included in model-training (Table 4.1, Supplemental Table 3). The 'sk-learn' and 'caret' library functions were employed, where for RF, the significance of a feature is determined by averaging its value over all trees in the forest. Each characteristic gains greater significance as the impurity lowers. The total of these normalised importance values is 1.

4.4.6 Partial-Dependence Analysis

Partial-dependence (PD) analysis visualises how given predictor variables may influence predicted probabilities of a specified outcome across a range of values within the trained machine learning model including tree-based algorithms and allows for causal interpretations (205). PD has been utilised previously to evaluate and explain predictive models in a wide array of use-cases (200,206–208). PD selectively perturbs variables of interest incrementally while preserving the remaining variables to generate new predicted probabilities from the model after each perturbation. These may then be plotted either for individual patients (individualised conditional expectation plots), as an averaged curve, or as probability contours providing an intuitive, visual, model-agnostic approach to global interpretability of the ML model and so was chosen for this study especially as it offers causal interpretations.

Tools such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) offer insight into predictions at the instance-level (although SHAP values

can be aggregated over predictions to provide global insights too) (209,210). Such local explanation tools are however principally beneficial in explaining predictions for individual patients once a clinical decision-support tool has already been deployed. PD, (as with variable importance) by comparison offers clinicians value earlier in the development of such HAIC processes by conveying global model interpretability as a surrogate microcosm of their team's decision-making paradigm and increasing trust in the validity of the underlying model as a result. While PD allows for causal interpretation, LIME creates new hyper-localised models for a given instance and is thus inappropriate for this, while SHAP has been shown to be unreliable in causal interpretations (211).

4.5 Results

4.5.1 Cohort demographics

Of 938 initially identified cases, 13 were excluded as relating to patients who underwent failed endoscopic resection prior to salvage oesophagectomy. A further 32 cases with cT stages "cT0" (N = 4), "cTis" (N = 3) and "cTX" (N = 25) were excluded for low numbers and to allow examination of any ordinal relationships. The final cohort of 893 cases are summarised by predictor variable in Table 4.1 with additional referral unit data presented in Supplemental Table 3.

Table 4.1 - Patient demographics and model predictor variables by sub-group. Referral unit statistics are provided in Supplementary Table. Performance status is measured as per the Eastern Cooperative Oncology Group (ECOG) Performance status scale.

Pre-treatment variables	"NACT +S" (N = 209) (%)	"NACRT +S" (N = 196) (%)	"Surgery-only" (N = 102) (%)	"Palliative" (N=386) (%)	Total (N = 893) (%)
Gender					
Male	179 (85.6%)	137 (69.9%)	80 (78.4%)	280 (72.5%)	676 (75.7%)
Female	30 (14.4%)	59 (30.1%)	22 (21.6%)	106 (27.5%)	217 (24.3%)
Median Age, Years (Range)	65.7 (21 – 81.8)	66.6 (40.0 – 81.0)	73.4 (33.7 – 83.0)	74.8 (32.0-96.7)	69.1 (21.0-96.7)

Pre-treatment variables	"NACT +S" (N = 209) (%)	"NACRT +S" (N = 196) (%)	"Surgery-only" (N = 102) (%)	"Palliative" (N=386) (%)	Total (N = 893) (%)
ECOG Performance status					
0	120 (57.4%)	138 (70.4%)	34 (33.3%)	68 (17.6%)	360 (40.3%)
1	84 (40.2%)	54 (27.6%)	56 (54.9%)	122 (31.6%)	316 (35.4%)
2	5 (2.4%)	3 (1.5%)	12 (11.8%)	124 (32.1%)	144(16.1%)
3	0 (0%)	1 (0.5%)	0 (0%)	69 (17.9%)	70 (7.8%)
4	0 (0%)	0 (0%)	0 (0%)	3 (0.8%)	3 (0.3%)
cT stage					
1	0 (0%)	0 (0%)	8 (7.8%)	1 (0.3%)	9 (1.0%)
2	35 (16.7%)	44 (22.5%)	49 (48.0%)	40 (10.4%)	168 (18.8%)
3	149 (71.3%)	138 (70.4%)	43 (42.2%)	211 (54.7%)	541 (60.6%)
4	25 (12.0%)	14 (7.1%)	2 (2.0%)	134 (34.7%)	175 (19.6%)
cN stage					
0	40 (19.1%)	64 (32.7%)	53 (52.0%)	82 (21.2%)	239 (26.8%)
1	138 (66.0%)	112 (57.1%)	42 (41.2%)	131 (33.9%)	423 (47.4%)
2	31 (14.8%)	19 (9.7%)	6 (5.9%)	121 (31.3%)	177 (19.8%)
3	0 (0%)	1 (0.5%)	1 (1.0%)	52 (13.5%)	54 (6.0%)
cM stage					
0	209 (100%)	196 (100%)	102 (100%)	162 (42.0%)	669 (74.9%)
1	0 (0%)	0 (0%)	0 (0%)	224 (58%)	224 (25.1%)

Pre-treatment variables	"NACT +S" (N = 209) (%)	"NACRT +S" (N = 196) (%)	"Surgery-only" (N = 102) (%)	"Palliative" (N=386) (%)	Total (N = 893) (%)
Tumour location					
Oesophagus					
Proximal	0 (0%)	3 (1.5%)	0 (0%)	18 (4.7%)	21 (2.4%)
Middle	5 (2.4%)	22 (11.2%)	7 (6.8%)	59 (15.3%)	93 (10.4%)
Distal	103 (49.3%)	148 (75.5%)	64 (62.7%)	235 (60.9%)	550 (61.6%)
GOJ					
GOJ Siewert 1	24 (11.5%)	8 (4.1%)	4 (3.9%)	20 (5.2%)	56 (6.3%)
GOJ Siewert 2	39 (18.7%)	10 (5.1%)	19 (18.6%)	54 (14.0%)	122 (13.7%)
GOJ Siewert 3	23 (11.0%)	1 (0.5%)	5 (4.9%)	0 (0%)	29 (3.2%)
GOJ Siewert Undefined	15 (7.2%)	4 (2.0%)	3 (2.9%)	0 (0%)	22 (2.5%)
Tumour Histology					
Adenocarcinoma	197 (94.3%)	134 (68.4%)	93 (91.2%)	274 (71.0%)	698 (78.1%)
Squamous Cell (SCC)	12 (5.7%)	62 (31.6%)	9 (8.8%)	112 (29.0%)	195 (21.8%)

Pre-treatment variables	"NACT +S" (N = 209) (%)	"NACRT +S" (N = 196) (%)	"Surgery-only" (N = 102) (%)	"Palliative" (N=386) (%)	Total (N = 893) (%)
Co-morbidities					
History of MI (MI)	9 (4.3%)	11 (5.6%)	10 (9.8%)	34 (8.8%)	64 (7.2%)
Chronic heart failure (CHF)	1 (0.5%)	1 (0.5%)	2 (2.0%)	17 (4.4%)	21 (2.4%)
Chronic pulmonary disease (CPD)	26 (12.4%)	28 (14.3%)	19 (18.6%)	48 (12.4%)	121 (13.5%)
Connective tissue disease	2 (1.0%)	5 (2.6%)	1 (1%)	0 (0%)	8 (0.9%)
Peripheral vascular disease (PVD)	6 (2.9%)	7 (3.6%)	5 (4.9%)	21 (5.4%)	39 (4.4%)
Cerebrovascular disease (CVD)	8 (3.8%)	6 (3.1%)	7 (6.7%)	65 (16.8%)	86 (9.6%)
Dementia	0 (0%)	0 (0%)	0 (0%)	10 (2.6%)	10 (1.1%)
History of Peptic Ulcer Disease (XPUD)	8 (3.8%)	7 (3.6%)	5 (4.9%)	14 (3.6%)	34 (3.8%)
Uncomplicated diabetes (DM uncomp)	21 (10.0%)	20 (10.2%)	16 (15.7%)	60 (15.5%)	117 (13.1%)
Complicated diabetes (DM comp)	0 (0%)	1 (0.5%)	1 (1.0%)	3 (0.8%)	5 (0.6%)
Leukaemia	0 (0%)	0 (0%)	3 (2.9%)	1 (0.3%)	4 (0.5%)
Lymphoma	1 (0.5%)	2 (1.0%)	3 (2.9%)	4 (1.0%)	10 (1.1%)
Mild liver disease	2 (1.0%)	0 (0%)	0 (0%)	4 (1.0%)	6 (0.7%)
Hemiplegia	0 (0%)	0 (0%)	0 (0%)	2 (0.5%)	2 (0.2%)
Renal failure	0 (0%)	1 (0.5%)	3 (2.9%)	33 (8.5%)	37 (4.1%)
AIDS	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Treatment-allocation over time was plotted to visualise general trends within the context of the landmark CROSS (neoadjuvant chemoradiotherapy + surgery) and FLOT4 (neoadjuvant chemotherapy + surgery) trials as well as assessed on partial dependence by epoch for effect on treatment probabilities (Figure 4.1 & Supplemental Figure 8 respectively) (8,9).

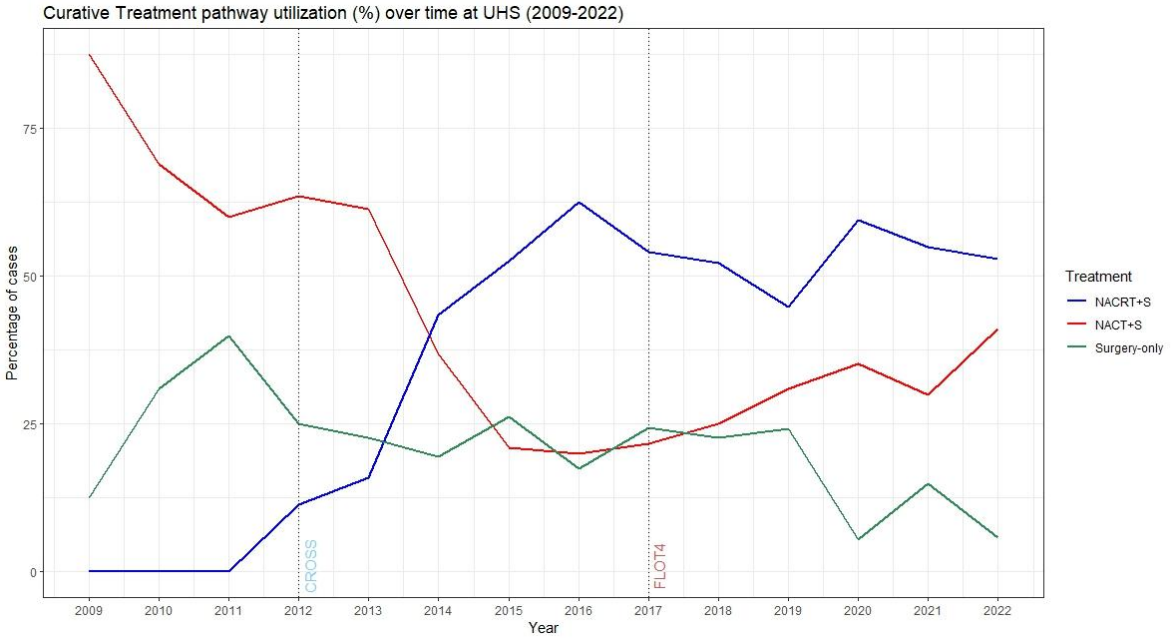


Figure 4.1 - Curative treatment allocation between 2009-2022, by year at UHS. Approximate time points for dissemination of the CROSS and FLOT4 trials which provided seminal evidence for NACRT and NACT respectively are also overlayed for reference. The pre-CROSS time-period, Cross-to-FLOT4 time period and post-FLOT4 period were incorporated into the classification model as epochs.

4.5.2 Model performance

Classification performance for the RF classifier model using multi-class receiver operator characteristic area under curve (ROC AUCs) is illustrated in Figure 4.2. All classes were separable with excellent AUCs, (neoadjuvant chemotherapy + surgery (NACT+S) 0.90, neoadjuvant chemoradiotherapy (NACRT+S) 0.88, Surgery-only 0.88, Palliative therapies 0.99) with reasonable calibration (Figure 4.3) and mean balanced accuracy (0.795±0.008). This again aligns with previous experience of the use of random forest models classifying curative OC treatment plans in a smaller dataset (108).

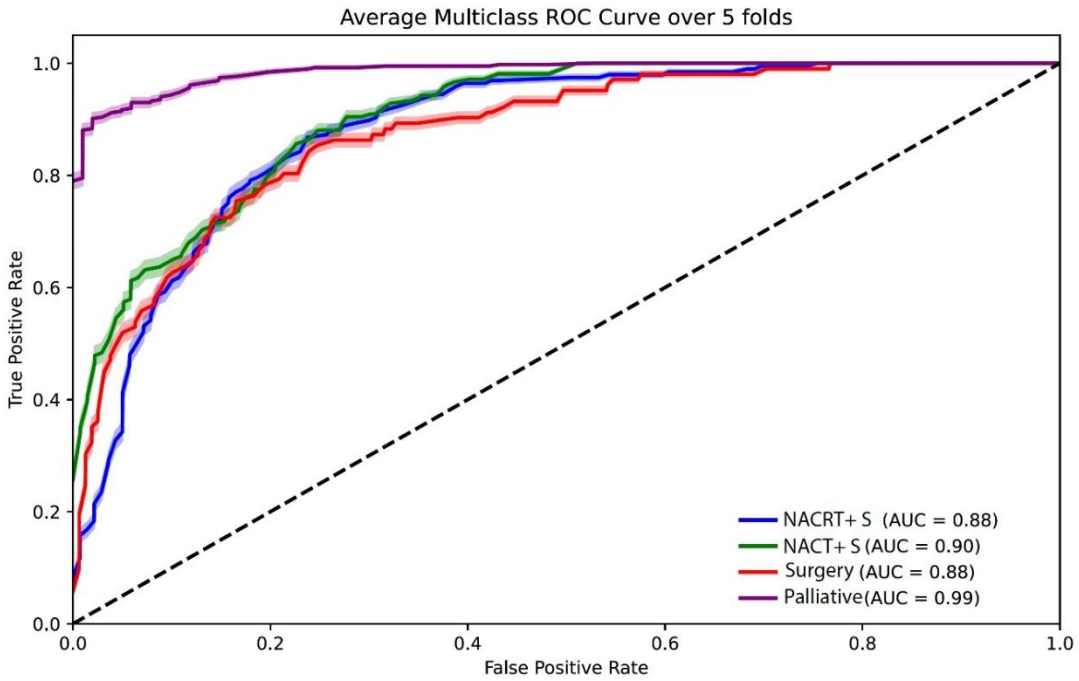


Figure 4.2 - Multiclass ROC curve for random forests treatment classifier representing a one vs others class-prediction performance. K=5 Cross-validation was conducted using an 80:20 split. Mean ROC is presented ±1x Standard Error of the Mean.

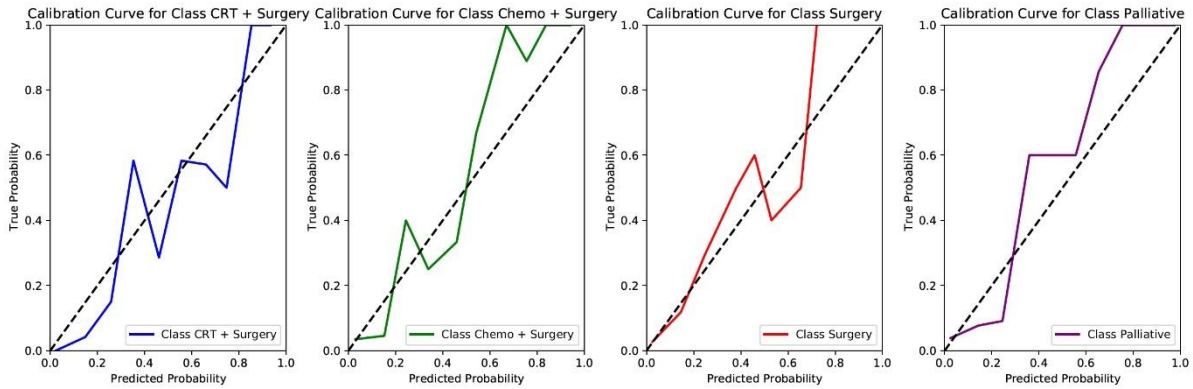


Figure 4.3 - Calibration plots for the RF model by outcome class prediction. Key: CRT + Surgery = NACRT+S, Chemo + Surgery = NACT + S.

4.5.3 Variable importance

Variables such as clinical TNM stage and tumour characteristics (location & histology) comprise standard criteria for treatment planning within national guidelines with cM stage and performance status key differentiators for curative versus palliative pathways (122). On relative

variable importance however, age notably ranked third when trained on the full cohort (Figure 4.4a) after cM stage and performance status, followed by, epoch, cN stage, cT stage, referring location, tumour site and histological subtype. In view of its consistently high ranking, we focussed on age in PD analysis both in isolation and in combination with these variables to examine their interrelations further. Within a second 'curative-only' model age ranked first, further validating its focus within this study (Figure 4.4b).

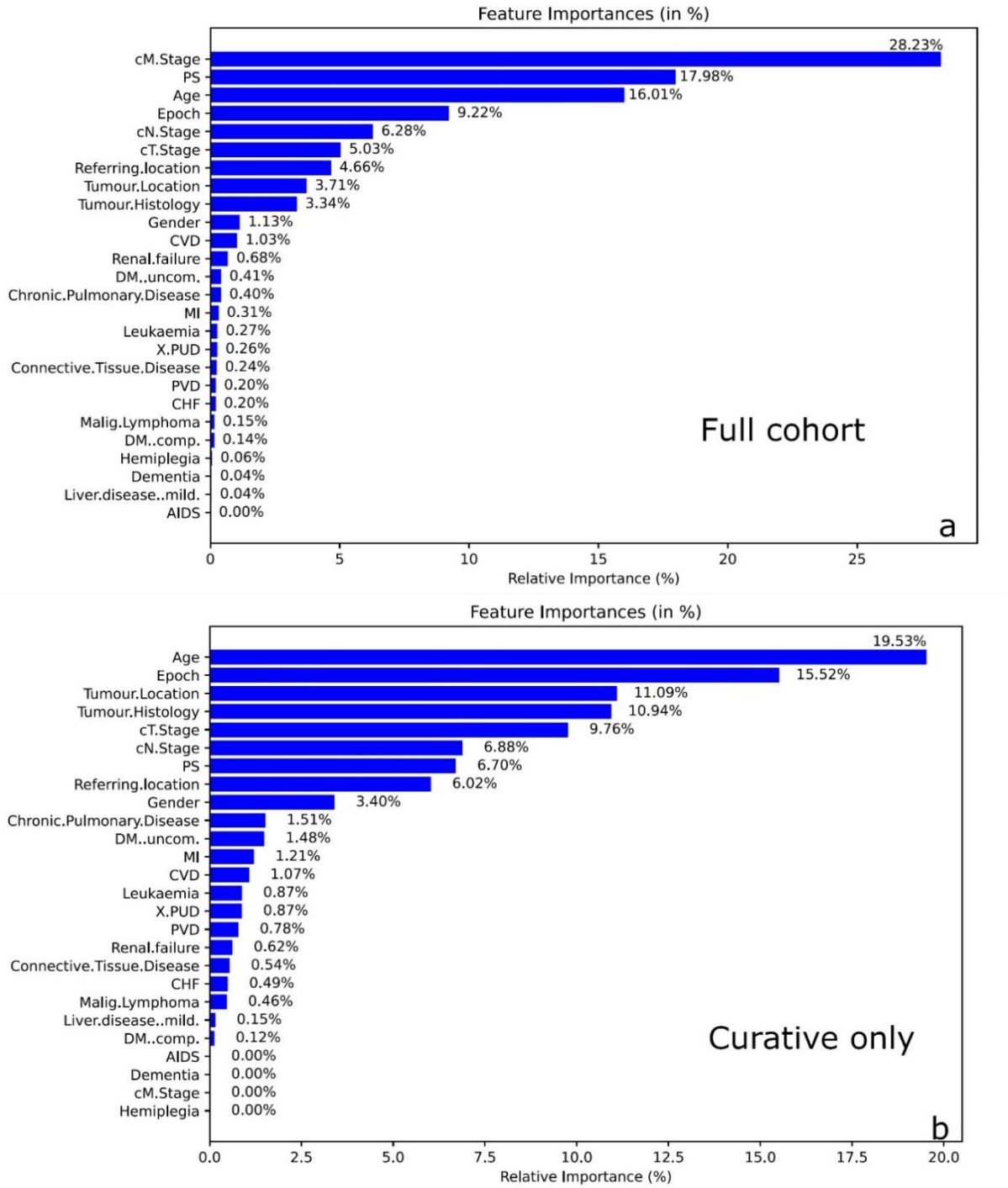


Figure 4.4 – Relative variable importance plot of the Random Forests classifier model. Importance values are plotted for all patients (a) and curative patients only (b) in rank order with most important at the

4.5.4 Influence of Age on Treatment Decisions

Variation in treatment probability due to age alone was investigated using individual conditional expectation plots (Figure 4.5). In all groups, a noticeable change in probabilities occurs after 75 years. Patients predicted for surgery-alone experience a probability rise between 75 – 85 years

after which they return to pre-75-year baselines. For neoadjuvant therapy (NAT), probabilities fall sharply after 75 years, however this decline starts as early as 70 years in the NACT+S group. Palliative pathway probabilities are largely consistent prior to 75 years however a clear upshift is seen beyond this time point.

The patient cohort was segregated into two subgroups (< 75 years vs 75+ years) to statistically test for age-related differences between treatment classes (Table 4.2). No significant difference was found between treatment groups within the younger subgroup or between NACT vs NACRT within the older cohort. A significant difference is seen between the palliative cohort against curative treatments as well as between Surgery and both NAT modalities within the older cohort.

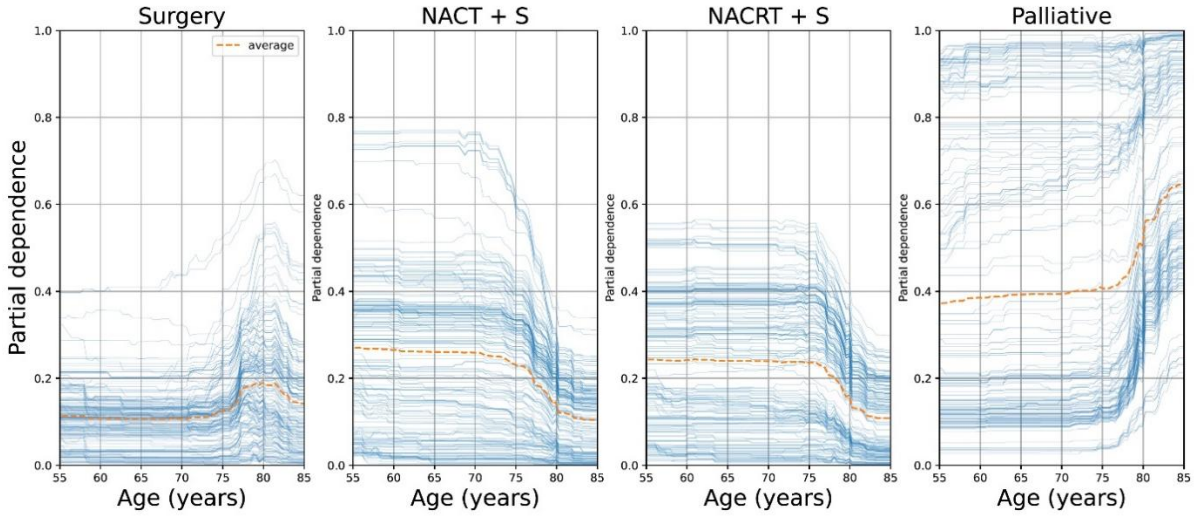


Figure 4.5 - Individual conditional expectation plots for predicted probability of treatment decision against age. Predicted probability (y axis) of each treatment pathway is plotted against the age range of the cohort (x axis) for each patient (blue lines). The averaged curve is also provided (orange dotted line).

Table 4.2 - Kruskal-Wallis test for median age difference by subgroup outcome class (significant differences in bold)

Δ Median age (75+) **(P value)**	**NACT+S**	**NACRT+S**	**Surgery**	**Palliative**
NACT+S	No data	0.3 years (P = 0.470)	**1.6 years** **(P = 0.015)**	**6 years** **(P < 0.001)**
NACRT+S	No data	No data	**1.9 years** **(P < 0.001)**	**6.3 years** **(P = <0.001)**
Surgery	No data	No data	No data	**4.4 years** **(P < 0.001)**
Palliative	No data	No data	No data	No data
Δ Median (<75) **(P = 0.09)**	NACT	NACRT	Surgery	Palliative
NACT	No data	0.7 years (P = 0.531)	0.5 years (P = 0.531)	1.4 years (P = 0.079)
NACRT	No data	No data	1.2 years (P = 0.925)	2.1 years (P = 0.391)
Surgery	No data	No data	No data	0.9 years (P =0.531)
Palliative	No data	No data	No data	No data

4.5.5 Age vs Tumour Staging

The relationship between age and tumour staging (cT/cN) stage was assessed (Figure 4.6, purple regions represent low probability, yellow regions represent high probability). For surgery-alone strategies, age proved minimally influential under 75, directed instead by disease-stage. From 75-85yrs however, probabilities increase independently of staging. The probability contours demonstrated most variation for the surgery-alone group at approximately cT2 N0

indicating this group may experience significant variability in treatment plans. For NACT+S, highest likelihood (yellow) was focussed on cT3-4 N1 for under 75s after which likelihood dropped in line with advancing age. A similar pattern was observed for NACRT+S however the high probability zone is comparatively larger, extending from cT1-3 and cN0-1. For palliative therapies, advancing age acted synergistically with stage. As cM stage only applies to non-curative patients it could not be meaningfully assessed across pathways, however it demonstrates a binary influence across all treatments (Supplemental Figure 9).

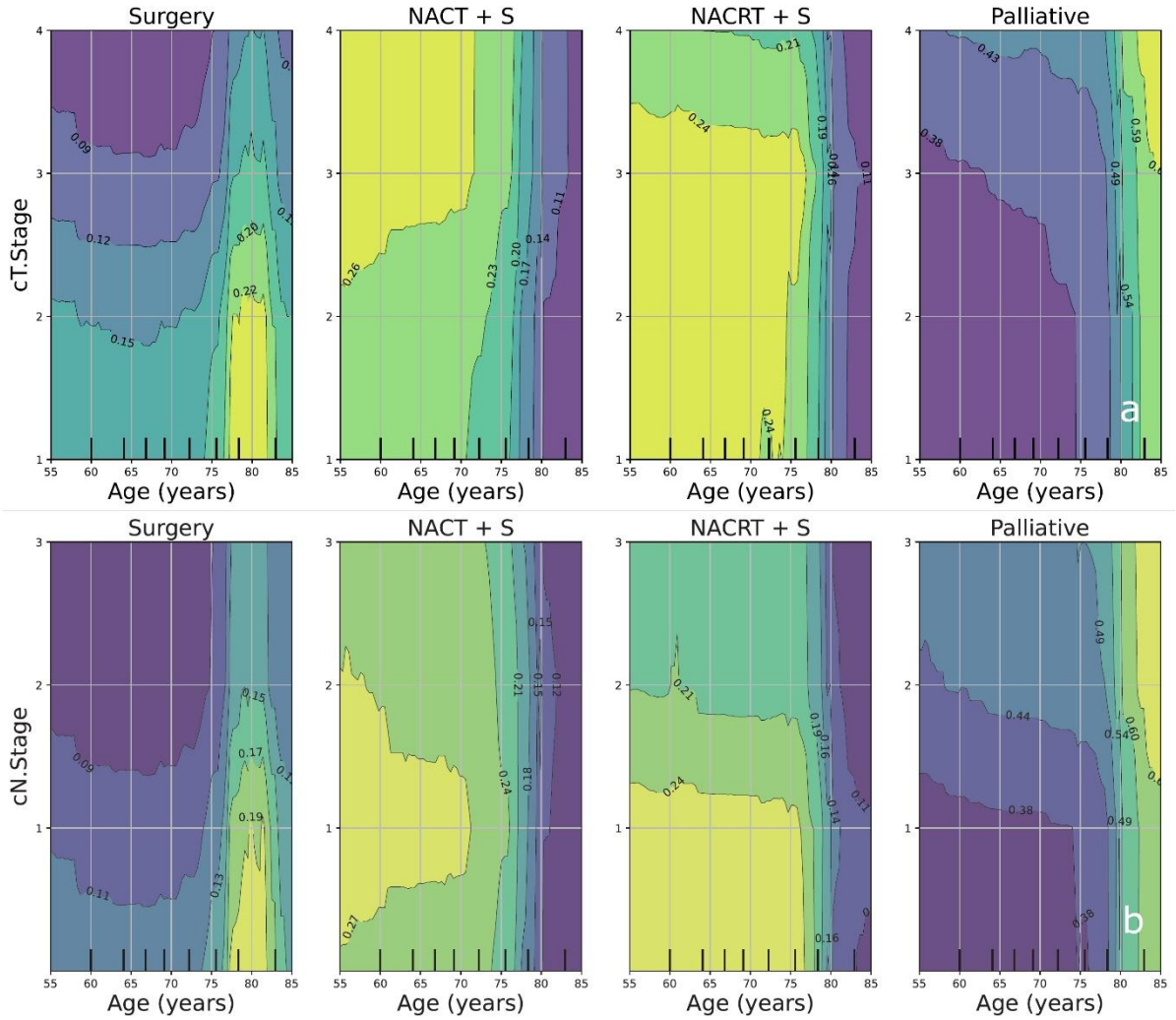


Figure 4.6 - 2-Dimensional Partial Dependence contour plot of Age vs cT Stage (a) and cN stage (b) on predicted probability of a treatment pathway. Interrelation between disease staging and patient age is mapped against four distinct OC treatment modalities: Surgery (S), NACT+S, NACRT+S and Palliative management. The x-axis delineates the age range of the patient cohort, while the y-axis captures the various cT/N staging levels on a continuous axis. Intensity of the colour gradients within the contour plot signifies the likelihood of selecting a particular treatment, with yellow shades indicating higher probability while purple regions indicate lowest probability and numbered contours equate to that probability (e.g., 0.24 = 24%).

4.5.6 Age vs Tumour characteristics

Tumour location demonstrated a hierarchical influence, conferring greater likelihood for surgery-alone strategies with progressively more distal tumours (Figure 4.7a). A similar, exaggerated effect is seen in NACT+S cases whereas this grouping is closer for NACRT+S. Mid-distal oesophageal tumours showed higher likelihood for NACRT+S while gastro-oesophageal junction (GOJ) type 1-2 and proximal oesophageal tumours exhibited a lower probability.

Proximal tumours were associated with highest likelihood for palliative pathways. Across modalities age continued to exert little influence under 75 years.

Histology separated base probabilities for all treatment choices independently of age (Figure 4.7b). Irrespective of age, adenocarcinomas were more likely to receive surgery-only and NACT+S over squamous cell carcinomas (SCCs) which were more likely to be assigned NACRT+S or palliative pathways. Palliative therapy likelihood rose in step with advancing ages regardless of histology.

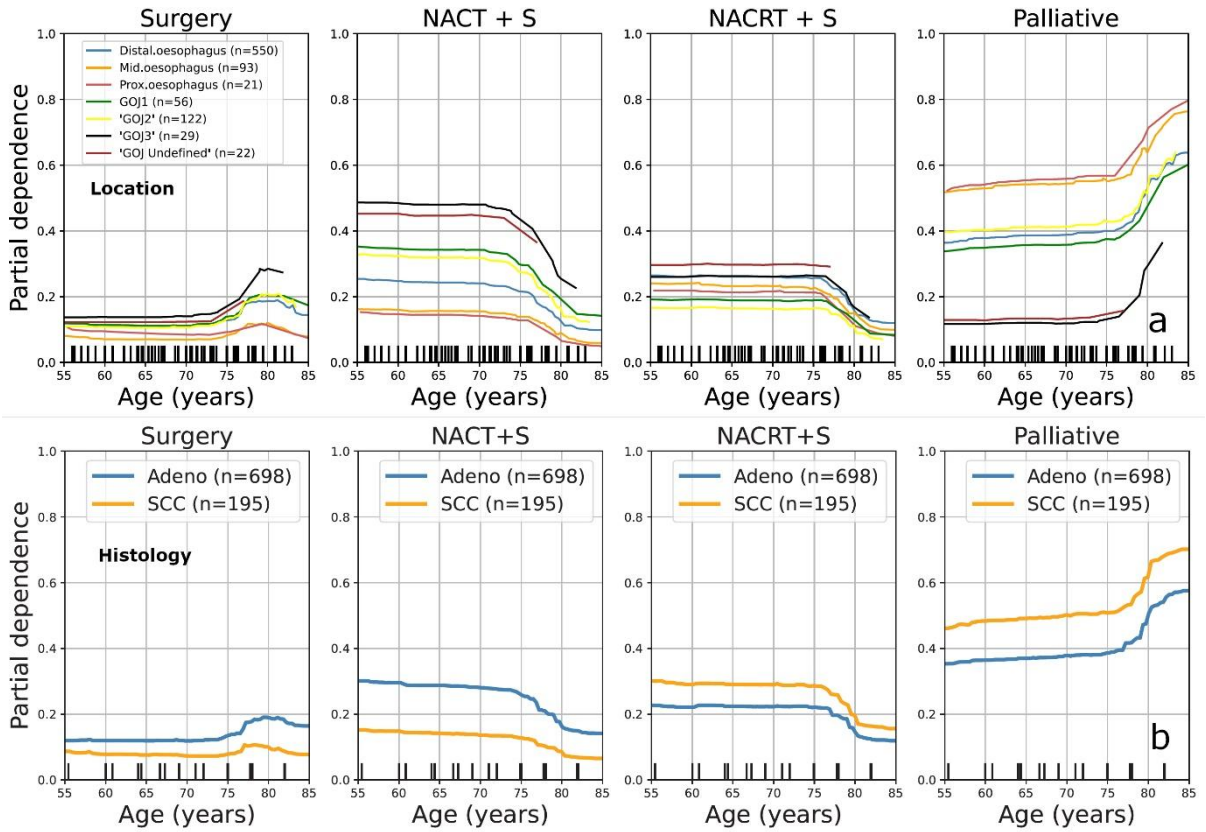


Figure 4.7 - Averaged Partial Dependence Plot of Age vs Tumour Location (a) and Tumour Histology (b) on treatment decision probability. to visualise interrelationships between the covariates.

4.5.7 Age vs Performance Status

The relationship between age and Performance Status (PS) demonstrated clear clustering into two patient cohorts across modalities: PS0-1 and PS2-4 (Figure 4.8). Under 75 years, age exerts minimal influence on surgery-alone probability. In older patients, PS0-1 cases experience a probability rise while PS2-4 patients follow a consistent low-probability trajectory, confirming

that advanced age forced selection of the fittest patients for surgery-alone strategies. PS0-1 patients were significantly more likely to get either NAT modality under 75 years after which probabilities re-converged with the PS2-4 cohort. PS2-4 patients were again much more likely to be assigned palliative pathway designation for across all age groups while PS1 patients only start to converge with the PS2-4 cohort after 75 years.

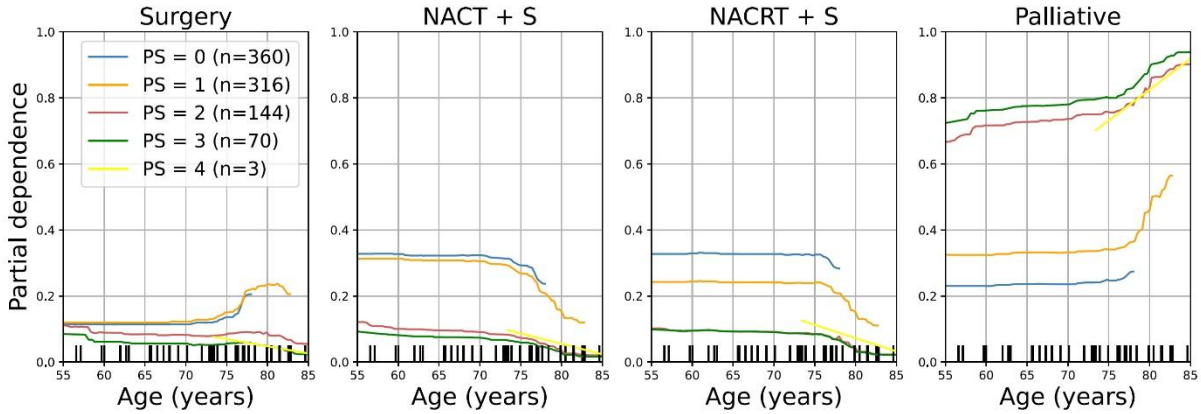


Figure 4.8 - Partial Dependence Plot of Age and Performance status (PS) on treatment decision predicted probability.

4.6 Discussion

4.6.1 Summary of findings

This study applied XAI techniques to quantify the influence specific clinical variables exert on the probability of a given treatment decision by the OC MDT. The study's findings of a model demonstrating strong AUC, balanced accuracy and calibration show that ML combined with XAI techniques can act as a vehicle to interrogate and analyse team-based decision-making dynamics with a granularity superior to classical statistical approaches. The ability to extract quantifiable objective insights, the majority of which align with observed clinical practice reinforces trust within the underlying model as a microcosm of the human MDT from which it draws inferences. As a proof-of-principle, the modelling in this study was not aimed towards clinical outcomes downstream of the decision (such as survival or quality of life), instead intentionally focussed on the route towards the treatment-decision itself in the first instance.

4.6.2 Age as a potential subconscious bias

Age, while not traditionally a criterion within management guidelines proved significant to OC treatment decisions, a finding consistent with our previous work which we are able to examine

in detail here (108). An important checkpoint within the seventh decade of life is highlighted which splits patients into two cohorts experiencing differing probabilities for treatment pathways. Patients over 75 years remain more likely to receive surgery-alone or palliative strategies and less likely to be offered NAT. As previous studies have historically highlighted a change in risk profile at 75 years this remains in keeping with our findings (212,213). Furthermore, this study indicates age may act as a surrogate marker of patient fitness even in the presence of functional metrics such as performance status. We shared these findings with our MDT and asked if they recognised chronological age as an influential to their decision making. Initially, members believed that age was not a routine consideration in their decision-making. However, after engaging in reflective feedback sessions, they recognised that age did play a role, albeit subconsciously though they had not initially been able to place a specific age cut-off. This led some members to consider other possible subconscious influences and whether these were biases or simply based on experience (190).

4.6.3 Variability in treatment decisions

We have explained the relationship within our model between disease-staging co-variates and age, with the former more important in the under-75 group and the latter driving choices thereafter. Of interest, we observed the greatest variability in decision-making (depicted by a broad range of partial dependencies) for surgery-only strategies in those with cT2N0-1 disease. This fits a long-established controversy within the UK regarding the optimal management of this cohort. By definition cT2N0 disease breaches the muscularis propria with further potential for submucosal lymphatic invasion, leading to unpredictable tumoral behaviour within this group (214). Compounded with historically high rates of under-staging, this cohort poses a therapeutic dilemma – utilise potentially toxic NAT (presuming undetected nodal disease) and risk deconditioning patients out of surgical fitness with potentially no additional survival advantage (215,216).

NAT decisions were mainly influenced by advancing age over staging with NACRT deployed over a wider age and staging range than NACT, but a drop in use of NAT altogether in older patients. This is attributable to a broadly held view that NACT regimes such as FLOT (Fluorouracil, Leucovorin, Oxaloplatin, Docetaxel) may be more toxic or less tolerated than NACRT (10,14,217–220). It is worth noting however that while successful completion of all cycles for NACT regimens (e.g., pre- and post-operative FLOT) are lower versus NACRT, a high proportion still manage all pre-operative cycles to reach surgery (10,221). Furthermore, concern over

adverse effects with NACRT on tissue friability and anastomotic leakage rates intra- and post-operatively has prompted some Chinese units to favour NACT, even in OSCC for those with perceived poor treatment tolerance or frailty (218,220,222). PD analysis suggested that NACT+S use within our unit dropped during epoch 2 (CROSS-FLOT4) but without rebounding post-FLOT4 as NACRT+S did after CROSS. This may be due to slower uptake by those keenly established in using NACRT+S especially while clinical equipoise persists regarding survival advantage. Modelling with trial epochs thus allows for changes in practice over time and requires periodic re-evaluation following future trials (10,223). Predictably, staging was synergistic with age on palliative pathway prediction reflecting the combination of disease burden and frailty associated with advanced age.

4.6.4 Is Age perceived as a surrogate marker of functional fitness?

The interrelation between age and performance status is particularly interesting within this study as historically the former has sometimes been treated as a surrogate marker of frailty, prejudicing older patients away from aggressive treatments (177,213,224). While it is important to identify and mitigate against inequitable biases such as treatment allocation driven primarily by age alone in favour of a more objective metric of functional capacity it is also important to recognise that some of this perceived bias may in fact represent earned and lived experience clinicians who have treated and operated on the older patient population and their associated risk (225,226). Furthermore, there is evidence to argue that not all “bias” necessarily equates to inequity, as some biases may in fact ensure that a patient unlikely to cope or tolerate intensive therapies is not needlessly overtreated without pause for what may in their best interests (227).

PD analysis in this study also grouped patients into two dominant performance status clusters independently of age: PS0-1 versus PS2-4. The PS2-4 cluster experienced a significantly lower likelihood for NAT and were much more likely to be offered palliative treatments fitting a well-established prognostic significance of pre-treatment patient physical activity. Metabolic Equivalents or METS (measured by oxygen consumption at rest and used in anaesthesia to quantify perioperative functional capacity) are predictive of poor outcomes at scores of 4 or less (228). Physical activity commensurate with such scores approximate to PS2 or worse, suggesting that this clustering reflects anticipation for treatment-related morbidity in this cohort. While national guidance on stratifying PS in curative OC cases is not currently offered, PD analysis allows for ML-driven benchmarking of observed clinical practice against current recommendations, a concept being explored in other surgical specialties (229). It seems

ultimately that clinicians may in fact be using both age and PS in combination to reach decisions for treatment allocation, with age perhaps holding more weight in those over the 77-year-old threshold and PS coming to the forefront in the younger population.

4.6.5 Tumour characteristics on neoadjuvant therapy choice

Tumour characteristics also outweighed age in the under-75s in driving treatment probabilities. GOJ tumours were more likely to receive surgery-alone versus oesophageal lesions and significantly more likely to receive NACT than NACRT. This is partly over concern for collateral radiation-induced damage to the planned gastric conduit at surgery, and in part to a historical body of trial data focussed primarily on oesophageal tumours (8,230–233). Across NAT, these decisions remain consistent until late into the 7th decade, at which point the deleterious effect of age is observed. High oesophageal tumours were additionally more likely to receive palliative outcomes versus distal lesions, in keeping with the significant challenges curative management for such lesions pose, and where resection in particular may be extensive (234). Histology and age followed a similar pattern with adenocarcinomas more likely to receive surgery or NACT+S independently of age while SCCs were favoured for NACRT+S and palliative outcomes. This fits with the radiosensitivity of SCC subtypes coupled with greater potential for tumour response however a survival benefit from NACRT for adenocarcinomas however remains debateable (235,236).

4.6.6 Implications of this study

In 2016, Cancer Research UK, demonstrated that MDTs within the UK were operating under significant strain and resource scarcity (26). Among many of their key findings was a significant challenge for MDTs finding the time to audit and reflect on their decision-making processes. Although numerous studies have, in recent years, demonstrated the capabilities of AI to support, replicate, or even beat the human clinician in clinical tasks, none to date have considered the benefit of AI in auditing or unpicking the human decision-making process (237). This study shows that AI may also provide significant benefit as vehicle for early-warnings of subtle shifts in practice, sub-conscious or even unconscious bias, and identifying areas where variability indicates a definitive knowledge gap which may in turn guide research questions downstream. XAI techniques offer the best way forward by championing accurate, capable high-functioning AI while balancing this with the need for transparent, auditable processes.

When working in symbiosis with human counterparts within the MDT, this can provide for the ideal of "AI-augmented clinicians" (196).

4.6.7 Study limitations and strengths

This was a single-centre retrospective analysis of 893 OC patients over a 13-year period during which a number of shifts in oncological practice have undoubtedly occurred in both NACT regimens and emerging immunotherapies. However, little clarity has been achieved even now in optimal NAT regimens or management of cT2N0 patients. The strength of this study is in its novel use of XAI on a large single-centre cohort of nearly 900 patients evaluating both curative and non-curative treatment pathways which broadens its generalisability. We have demonstrated how transparency can be introduced for team-based oncological treatment decisions, detailing clear shifts in human decision-making when faced with specific clinicopathological scenarios in oncological settings known to suffer chaotic leadership styles (238). This approach allows us to examine and re-examine the robustness of our decision-making to standardise practice for OC patients (especially given that there is evidence to indicate heterogeneity of decision-making even between OC MDTs (24)) and can be applied to other MDTs regionally, nationally or internationally in future for direct comparison of MDTs as well as being translatable to MDTs from other cancer types. Where MDTs have little time across cancer types for audit, self-reflection or learning (23,26,66,149,174), global XAI could be integrated into MDT workflows as part of annual departmental audits for quality control, sense-checking shifts in practice. Training data drift can be tracked to ensure models remain appropriate and true to the local population (239). As CDSS tools the evolve, local XAI techniques such as LIME and SHAP may be integrated within the user-interface to offer additional instance-level explanations in real-time tailored to the individual patient (209,210). While the present study is not designed to determine the clinical justification for decisions influenced by variables such as age, it highlights scenarios for MDTs to focus upon during clinical governance processes while introducing clinicians to the capabilities of AI-derived decision support.

However, while XAI techniques explain recommendations, this does not automatically guarantee clinician uptake of that recommendation. A recent study testing clinicians' fluid-prescriptions when offered additional advice from simple AI or XAI noted little difference on self-reporting, in outcome whether explanations were provided or not, questioning whether explanations were of material benefit above an AI recommendation (240). The study faced

some methodological challenges, namely the reliability of self-reporting, sample size and the generalisability of the clinical scenario. The question it raises however is valid, engagement often depends on the user's level of technical understanding, the effectiveness of communication methods for explanations, and whether clinicians perceive the explanations as beneficial beyond ML experts (195,241). Despite this, the prevailing wind within healthcare remains a need for trustable AI solutions which open the “black box”. Bridging the gap to non-technical clinicians must the occur through education programs to ensure they can critically appraise not only AI models but the explanations which may accompany their outputs(242).

4.6.8 Future work

Future work will include applying the technique to external centres to compare and contrast our findings both within OC but also in other cancer-types. Testing other well-known ML algorithms from more inherently interpretable options such as decision-tree models and more complex ensemble learners such as eXtreme Gradient Boost may also be useful in evaluating insights across algorithms. Furthermore, ongoing work within this space will inevitably lead to the co-development of ML-derived decision-support tools trained on human-led MDT decisions. By applying Responsible Research and Innovation (RRI) frameworks we are currently engaging with and including stakeholders’ opinions (oncologists, radiologists, psychologists, computer scientists, and patient representatives) to identify strategies for optimal implementation, acceptability and usability of such an ML-derived tool (243). We are incorporating RRI principles to support multidisciplinary scientific collaboration, anticipate key future challenges and reflect on better practices for responsible data governance. The need for explainable and preferably interpretable models built on RRI principles is paramount, especially now with the potential for AI in healthcare to transform the practice of medicine in general. The approach presented here represents a route towards trust within these frameworks by first offering global insight into team-level decision-making when mirrored by ML. While the model used in this study is tailored to our local MDT, the process can be performed either on a population-level for scalability or targeted to a specific unit, to enhance data diversity and representativeness of underrepresented demographic groups or geographical areas). Understanding decision-drivers, some of which we argue are sub-conscious in practice, is invaluable for the pursuit of personalised medicine for OC patients and essential in building clinician-patient trust in future implementations of AI within OC.

4.7 Conclusion

This study applied XAI methods to highlight how significant, yet sometimes subconscious factors like age may influence treatment decisions for OC patients. While treatment choices are often framed by clinical factors, age remains salient even with functional metrics like performance status delineating patients into a fitter cohort, more likely to undergo all curative treatments, versus an unfit less-eligible group. The uniformity in predicted probabilities for curative treatments persists only until the 7th decade of life. After this, a notable rise in the probability for surgery-alone and palliative options is juxtaposed against a decline in neoadjuvant therapy (NAT) prospects. Our analysis not only emphasizes age's pivotal role amidst traditional clinical drivers but also showcases the clarity and insight achievable with ML in navigating complex treatment landscapes. This explainability is crucial for clinician-engagement and trust within future AI-based decision support tools.

4.8 Research in Context

The technical elements of this thesis have endeavoured to set the foundations for both prediction generation and explainability, both at the global level and the instance level. As described within Chapter 1, the work from Appendix F highlighted areas of alignment between MDT models as well as areas of discordance and differences in perception between what the machine deems important in OC treatment allocation and what the human in the room might believe to be important in reaching that decision. This work already showed signs that MDT personnel felt positive towards the use of AI-driven decision support within MDT settings. However, the foundations of useable and trusted AI include technical validation, transparency, stability, upgradeability and end-user buy-in. These needs must be met to allow such a CDSS to integrate, be trustable to be used sufficiently to provide value to the MDT and the wider healthcare infrastructure. The following chapter aims to draw all these threads together by firstly validating the ML models with external data for a truer sense of generalisability, then building these models into a user interface which allows users insights into why a prediction is generated, and finally integrating a program of Responsible Research and Innovation which allows key stakeholder such as MDT clinicians, patients, and patient representatives to vocalise the needs they feel must be considered in how such a CDSS evolves and is used. The outcome of this chapter is a CDSS which offers externally validated predictive performance, explainability and is cognisant of future needs for its evolution.

Chapter 5 The Oesophageal Cancer Multi-Disciplinary Tool: A responsibly co-designed, externally validated, machine learning tool for oesophageal cancer decision making

Journal: eClinicalMedicine (Lancet), 2025, Impact Factor 10.0, CiteScore 17.0

EClinicalMedicine (Lancet). 2025 Sep; 89: 103527. DOI: 10.1016/j.eclinm.2025.103527

Navamayooran Thavanesan BMBCh[1], Mohammad Naiseh PhD[2], Miguel Terol BA[1], Saqib Andrew Rahman PhD[1], Samuel Luke Hill PhD[1], Charlotte Parfitt MBBS[3], Zoë S Walters PhD[1], Sarvapali Ramchurn PhD[4], Sheraz Markar PhD[5,6], Richard Owen PhD[5,6], Nick Maynard PhD[5,6], Tayyaba Azim PhD[4], Zehor Belkatir PhD[4]. Elvira Vallejos Perez PhD[7], Mimi McCord[8], *Tim Underwood PhD[1], *Ganesh Vigneswaran PhD[1]

*These authors are jointly share Last Author for this manuscript

[1] Innovation for Translation Research Group (ITRG), School of Cancer Sciences, Faculty of Medicine, University of Southampton

[2] Department of Computing and Informatics, Bournemouth University

[3] University Hospitals Southampton NHS Foundation Trust.

[4] School of Electronics and Computer Science, University of Southampton

[5] Nuffield Department of Surgical Sciences, University of Oxford

[6] Department of Oesophagogastric Surgery, Churchill Hospital, Oxford University Hospitals NHS Trust

[7] School of Computer Science, Horizon Digital Economy Research, University of Nottingham

[8] Heartburn Cancer UK

Corresponding Author: Navamayooran Thavanesan

Address: School of Cancer Sciences, Faculty of Medicine, University of Southampton, South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, UK, SO16 6YD

Email: N.Thavanesan@soton.ac.uk

5.1 Acknowledgements

Within the scientific work presented here, I wish to acknowledge my co-authors for their contributions.

Contributions:

1) **NT was involved in the conception of this work, primary data collection, primary data analysis, UI coding, its drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy**
    - **NT performed the data collection, collation, cleaning and coding. He performed coding for the ML models and model evaluation methods in R (in collaboration with SAR for the palliative survival model) as well as the external validation methods in R. He was directly involved in conducting the included clinician interviews and RRI workshops. He provided code for the User interface in collaboration with MT. He drafted the initial manuscript based on the results he generated and made amendments to the subsequent drafts based on feedback by his co-authors on this paper. He will undertake the submission process, receive and act on reviewer comments and performed the final submissions too.**
2) Mohammad Naiseh provided invaluable qualitative data analysis (specifically thematic analyses) of the clinician interviews and RRI workshops, drafting qualitative sections of the methods and results section of the draft manuscript in conjunction with NT, final approval, and agreement of accountability for accuracy
3) Miguel Terol provided coding for the R Shiny User Interface in collaboration with NT which houses the ML models developed by NT and final manuscript approval
4) Saqib Andrew Rahman was involved in primary data analysis (specifically providing coding support for the palliative survival model training and performance evaluation in collaboration with NT) and final approval of the manuscript
5) Sam Luke Hill was involved in final approval of the manuscript
6) Charlotte Parfitt was involved in primary data collection, and final approval of the manuscript
7) Zoe Walters was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
8) Sarvapali Ramchurn was involved in the final approval of the manuscript

9) Sheraz Markar was involved in the final approval of the manuscript
10) Richard Owen was involved in the reviewing and recommending revisions for critical and important intellectual content and final approval of the manuscript.
11) Nicholas Maynard was involved in the final approval of the manuscript
12) Stephen Ash was involved in the primary data collection (Oxford data), final approval of manuscript
13) Tayyaba Azim was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval of manuscript
14) Zehor Belkhatir was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval of manuscript
15) Evira Perez Vallejos was involved guiding RRI principles adherence and in the reviewing and recommending revisions for critical and important intellectual content, final approval of manuscript
16) Timothy J Underwood was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
17) Ganesh Vigneswaran was involved in the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy

The CRediT taxonomy is as follows:

NT - conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing– review & editing.

MN - data curation, formal analysis, methodology, visualization, writing – original draft, and writing– review & editing.

MT - conceptualization, formal analysis, methodology, software, validation, visualisation, and writing– review & editing.

SAR - data curation, formal analysis, methodology, software, validation, visualisation, and writing– review & editing.

SLH - methodology, project administration, and writing– review & editing.

CP - data curation, formal analysis, and writing– review & editing.

ZSW - funding acquisition, investigation, methodology, project administration, supervision, writing – original draft, and writing– review & editing.

SR - methodology, resources, supervision, and writing– review & editing.

SM - data curation, investigation, methodology, resources, validation, and writing– review & editing.

RO - data curation, investigation, methodology, resources, validation, and writing– review & editing.

NM - data curation, investigation, methodology, resources, validation, and writing– review & editing.

TA - investigation, methodology, validation, and writing– review & editing.

ZB - investigation, methodology, supervision, validation, and writing– review & editing.

EVP - conceptualization, investigation, methodology, resources, supervision, validation, and writing– review & editing.

MM - conceptualization, data curation, investigation, methodology, resources, validation, and writing– review & editing.

TJU - conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft, and writing– review & editing.

GV - conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

5.2 Summary

The oesophageal cancer (OC) multi-disciplinary team (MDT) operates under significant pressures, handling complex decision-making. Machine learning (ML) can learn complex decision-making paradigms to improve efficiency, consistency, and cost if trained and deployed responsibly. We present an externally validated ML-based clinical decision support system (CDSS) designed to predict OC MDT treatment decisions and prognosticate palliative scenarios, co-designed using Responsible Research and Innovation (RRI) principles.

Clinicopathological data collected from 1931 patients between 4th September 2009, and 8th November 2022 were used to test and validate models trained through four ML algorithms to predict curative and palliative treatment pathways along with palliative prognosis. 953 OC cases treated at University Hospitals Southampton (UHS) were used to train ML models which were externally validated on 978 OC cases from Oxford University Hospitals (OUH). Model performance was evaluated using Area Under Curve (AUC) for treatment classifiers and calibration curves for survival models. A parallel RRI program at the University of Southampton (United Kingdom) combining clinician interviews and inter-disciplinary workshops was conducted between 16.3.23 and 23.5.24. The RRI program comprised a group of 17 domain experts comprising programmers, computer scientists, clinicians and patient representatives to allow end-users to contribute towards the co-design of the CDSS user interface.

Cohorts differed in baseline characteristics, with the external cohort (OUH) being younger, having better performance status, and a higher prevalence of pulmonary and vascular disease. Despite these differences, the primary treatment model achieved mean AUCs of 0·873 – 0·909, and 0·711-0·815 for the palliative classifier model (averaged across k=5 cross-validation). On

bootstrapped resampling, mean primary classifier AUCs ranged from 0.863-0.867 (95% CI ranging 0.862-0.868) and mean palliative classifier AUCS 0.736-0.799 (95% CI 0.734-0.800). Predicted survival probability from the palliative survival model was well calibrated over the first 12 months post-diagnosis in both cohorts. The RRI program provided a collaborative environment leading to valuable modifications to the CDSS including prediction explanations, visual aids for survival and integrated education for users producing a user-friendly and quick to use tool.

We present a novel, responsibly developed, externally validated AI CDSS trained to predict oesophageal cancer MDT decisions. It represents the foundations of a transformative application of ML, personalised, consistent and efficient MDT decision-support within OC which aligns to RRI principles.

5.3 Funding

Doctoral Studentship for NT (Institute for Life Sciences (University of Southampton) & University Hospital Southampton), UKRI TAS Pump-Priming Grant (TAS_PP_00167).

5.4 Research in Context

5.4.1 Evidence before this study

Machine learning (ML) a branch of Artificial Intelligence (AI) may offers a viable solution towards supporting clinicians however to date no externally validated models have been reported within Oesophageal cancer (OC). We searched PubMed on August 27th, 2025, without date or language restrictions for publications using the terms “Machine Learning” AND “Oesophageal cancer” AND “Multidisciplinary Team” (or “Cancer Board” or “Tumour Board”). We did not identify any additional studies beyond those previously published by this research group investigating ML as a means of predicting treatment assignment at MDT for OC.

5.4.2 Added value of this study

The machine learning algorithms used within this study are easily accessible, off-the-shelf libraries and compatible within the current digital healthcare infrastructures of many countries worldwide. The resulting CDSS, which provides both treatment classification and palliative prognostication has been externally validated using data from a separate geographical

catchment. Finally, the parallel Responsible Research and Innovation (RRI) program, has integrated early input from stakeholders in the development process.

5.4.3 Implications of all the available evidence

Our results suggest that ML can learn and predict MDT treatment decisions effectively in OC posing significant implications for future-proofing MDT operations against continued rises in caseload both within OC as well as other cancer types. Future iterations can also adapt to novel molecular markers and treatment modalities. The CDSS here provides rapid decision support for OC MDT personnel as well as a platform with which to counsel patients.

5.5 Introduction

Oesophageal cancer (OC) is the 7th commonest cause of cancer death worldwide and is a cancer of unmet need (244,245). Affected patients commonly present beyond their late 60s, are nutritionally compromised and often co-morbid. They require high-quality decision-making as treatment options have grown in number and complexity, each carrying significant survival and quality of life implications (18). Cancer multidisciplinary teams (MDTs), while greatly improving patient outcomes, face a relentless increase in caseload and clinical complexity (26). They are susceptible to pressured, inconsistent and potentially suboptimal decision-making (23,24).

In 2017, Cancer Research UK evaluated UK MDT services finding an urgent need for evolution and adaptation within their operational framework (26). Their report stressed an aging population combined with expanding treatment options had led to caseload volumes rising linearly with almost no corresponding increase in MDT resources to adapt or cope, a scenario common to many economies and countries. MDTs had on average 2-3 minutes to discuss cases, with no additional time to audit, reflect or learn from their internal decision-making. The MDT's challenges are also financial: the national cost of MDTs in the United Kingdom was estimated at £50 million in 2010, £88 million in 2011/12, approximately £150 million by 2014/2015 and £316 million as of 2024 (26,246,247). While this data is now over a decade out of date, there is nothing to indicate that the situation has improved in that time with regards to cost or case discussion time. Furthermore, assuming a starting NHS consultant salary of approximately £100,000 p.a., a 3-hour MDT would cost at minimum £7,500 per consultant present per year (with a minimum of 4-5 consultants present being typical of most MDTs). Reducing an MDT by even an hour could provide a hospital significant savings over a calendar year.

A process to streamline, prioritize, and ease MDT caseload is essential within the current economic climate of many world regions. Artificial intelligence (AI) has seen a boom in healthcare use-cases in the form of clinical decision-support systems (CDSS) (38,39,41,44). Machine learning (ML), a branch of AI which leverages advanced computational power to identify patterns within complex and multimodal data has provided one such engine for CDSSs and its potential to support OC management has been recently recognized (63,73). ML has seen increasing adoption within early detection of cancer (248–250) yet while AI platforms have been applied to MDT-style frameworks in some medical fields, OC MDTs have remained untouched in this regard (41,43,44,63). Similarly, a paucity of qualitative evidence exists on the viewpoints of clinicians and patients on the use of AI CDSSs in OC which creates a knowledge gap when design such tools for translation. Medical AI (MAI) necessitates trustworthy, ethical and responsible innovation (55). Where much of the literature has focused on proving MAI tools, there is a paucity of consideration for their implications on stakeholders from design-to-deployment (55). These include governance, handling bias, quality control, data drift detection and AI explainability (59). Responsible Research and Innovation (RRI) has developed in recent years to address this, aiming to maximise societal benefit while minimizing harm (60). The AREA framework (Anticipation, Reflection, Engagement and Action) is an example of this which integrates RRI within the life cycle of research programs (60).

Within this study we present a novel, responsibly developed, externally validated AI CDSS trained to predict oesophageal cancer MDT decisions. The tool utilizes readily accessible, off-the-shelf ML algorithms built into a user-friendly interface. The CDSS was co-designed with Patient & Public Involvement (PPI), clinicians, and computer scientists specialising in AI. By harnessing AI-based technologies in a bid to replicate and simulate OC MDT decision-making ML may be able to offer the potential to streamline, standardize and increase efficiency within the OC MDT operational framework in a manner which still aligns with Responsible AI (RAI) principles.

5.6 Methods

This was a mixed-methods study including a retrospective complete-case analysis of oesophageal cancer patients across two tertiary referral centres in the UK (University Hospital Southampton and Oxford University Hospitals) under the ethical approvals of IRAS 233065 & 319540.

5.6.1 Study cohort

5.6.1.1 Training Cohort

Oesophageal cancer patients discussed at MDT at University Hospital Southampton (UHS) between 2010 - 2023 were identified from a prospectively maintained local database and unit submission records to the UK National Oesophagogastric Audit (NOGCA). Treatment decisions were based on UK National Institute for Clinical Excellence (NICE) guidelines (63,122). Patients who present with non-metastatic disease (T0-4, N0-3, M0 disease) and fit (determined by the referring clinician and ratified by the MDT) for neoadjuvant therapies and/or surgery are filtered down curative pathways. For those with metastatic disease at presentation, or who are non-metastatic but felt too unfit for curative treatment are managed with palliative intent which may also be filtered based on their performance status (PS 0-2 patients for example, are deemed eligible for 1st-line palliative chemotherapy by NICE).

The mainstay of curative treatment for locally advanced OC is surgical resection alone (designated “Surgery”) or surgery combined with neoadjuvant therapy (NAT) (neoadjuvant chemotherapy (designated “Chemo”) or neoadjuvant chemoradiotherapy (designated “CRT”)). While a small proportion of patients detected early are eligible for endoscopic resection, their management remains controversial and entry to the MDT, nuanced meaning they could not be standardized to allow a fair comparison (251). While they were excluded from the external validation process, the results of a UHS model incorporating endoscopic resection are presented separately within the supplementary materials. Definitive CRT as monotherapy was also excluded from this study owing to insufficient training data for meaningful modelling.

In general, non-curative patients are offered one of five possible outcomes: best supportive care (designated “BSC”), palliative chemotherapy (designated “Chemo” within the palliative

models), palliative radiotherapy (designated “RTX”, typically to either the primary tumour and/or symptomatic secondary sites amenable to radiotherapy, however for the purposes of this study, RTX was defined as therapy to the primary tumour), palliative oesophageal stent alone, or with an oncological adjunct (chemotherapy or radiotherapy, and designated “Stent_Onc”).

Predictor variables for model training were derived from clinicopathological variables known to be routinely considered by the MDT. 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 (7th edition until 2017 and 8th edition thereafter). Novel molecular markers and immunotherapies which have been approved for metastatic disease in the UK since 2021 were not built into this first generation of models as these are emerging treatments and consequently there was insufficient training data for inclusion.

5.6.1.2 External validation cohort

The validation cohort were identified from a prospectively maintained clinical database (Cancer Outcomes Database Application for Upper GI or “CODA-UGI”) at Oxford University Hospitals (OUH) which was similarly submitted to NOGCA. The included patients were discussed at MDT over the same study period and underwent the same inclusion/exclusion criteria as the training cohort.

5.6.1.3 Ethics

This research (including all relevant participant informed consents) was conducted under the following ethical approvals; The United Kingdom Heath Research Authority (HRA) Integrated Research Application Systems (IRAS) 233065 & 319540 as well as under the approval of the local ethical review board: University of Southampton Ethics Research & Governance Online (ERGO) 70735. Anonymised external validation data access was granted after review by CODA-UGI data access committee, and following registration and approval via the Oxford University Hospitals governance platform (project no. 8441).

5.6.2 Statistics

5.6.2.1 Patient Sample

Sample size was dictated by the number of retrospectively recorded cases available for analysis at both centres. As a specific "effect" is not sought here from comparing treatment outcomes, a sample size calculation was not relevant to this use case. We set a historical boundary at 2010 to ensure we balanced maximising sample size while ensuring treatment paradigms remained relevant and still in-practice within the modern era.

5.6.2.2 Cohort comparison

Differences between the training and validation cohorts were assessed using Standardised Mean Difference (SMD). An SMD of 0.2 was deemed a small difference, 0.5 a medium difference and 0.8 a large difference.

Numeric performance metrics where relevant are presented as mean ± standard deviation (SD) and mean ± standard error from the mean (SEM) for the 5-fold cross-validated models. Where model performance has been tested with bootstrapped resampling, 95% confidence intervals have also been provided.

5.6.2.3 Model comparison

Differences in performance between algorithms 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)).

5.6.3 Machine Learning Model Development

5.6.3.1 Data preparation and analysis

Data analysis, model training and validation were conducted in R (version $4 \cdot 2 \cdot 2$) with relevant packages described where first used (Supplemental materials). The features used in this study (Table 5.1) are derived from a combination of domain expertise and UK national guidelines (122). Data was manually checked for quality control by NT and CP. Data entry was standardised for analysis using terminology accepted within the clinical field. As this was a complete analysis, any missing data was retrospectively extracted from hospital electronic

health records to ensure high-fidelity quality control. Age and overall survival were treated as continuous variables, while the remaining covariates were categorical (Table 5.1). Three separate decision-assistance models were developed: a primary classification model which triaged patients into either a specific curative pathway directly or triaged to a secondary, bespoke, palliative treatment classification model. A third, survival model was also trained to predict prognosis for a palliative patient from time of diagnosis when factoring in palliative treatment. Survival analysis was first undertaken using a Kaplan-Meier survival estimator ("survival" package). Median survival was stratified by treatment with a log-rank test-of-significance between curves. Overall survival was defined as survival from date of diagnosis to date of death or last recorded follow-up.

5.6.3.2 Feature selection

The features used in this study are derived from a combination of domain expertise and UK national guidelines (122). The features outlined in Table 1 are common to both the full cohort model and the palliative models except for the additional "obstructing" variable within the latter which was defined as either severe dysphagia to solids and liquids or difficulty passing the gastroscope at the time of the original diagnostic gastroscopy (while dysphagia of some degree is a hallmark of OC even in curative settings, cases which are deemed curative at diagnosis have rarely progressed to a stage where the lesion is causing severe dysphagia or an inability to pass a gastroscope which is more typically of palliative cases). The final palliative treatment allocation was then included as an extra feature within the palliative survival models. Feature selection was primarily dictated by the clinical variables routinely collected at the respective training and validation units (this was to ensure a pragmatic access to realistically accessible variables combined with domain knowledge of variables routinely discussed at MDT. Race, BMI, smoking status for instance are not routinely discussed or considered beyond exception circumstance (in situations of extremely high BMI which may make surgery more challenging or risky for instance). Similarly, while the American Society of Anaesthesiology (ASA) grading system is assessed pre-operatively in all surgical candidates, this score is not used in those not undergoing surgery or those who are palliative. As such their performance status is a more practical variable as it is considered across treatment pathways.

5.6.3.3 Machine Learning algorithms

The ML algorithms used in this study were chosen for several reasons: firstly, they allowed us to focus explainable, accessible and technically realistic ML architectures which can be implemented easily within current healthcare systems. In many world regions (including the UK) these systems are under immense financial and technological restrictions. Deep learning platforms were avoided as they are too opaque for this level of high-stakes decision-making, and too complex for easy implementation while still allowing regulators and hospital clinicians ready access to the explainability of the final decisions. Furthermore, high quality, clean, clinical data is notoriously difficult to curate at the scales needed for deep learning platforms which typically demand thousands if not tens of thousands of data points for quality learning, making standard architectures which can handle smaller datasets instantly more favourable. Finally, it is established that within tabular data structures, ML algorithms such as tree-based models outperform deep learning architectures when provided tabular data (252). Multinomial Logistic Regression, Random Forests and eXtreme Gradient Boost models were trained through "caret" package using "nnet", "RandomForest", and "xgboost" libraries respectively (108). Survival modelling used Random Survival Forests as these have been shown to outperform traditional Cox Proportional Hazard models for prognostication in OC patients post-oesophagectomy (randomForestsSRC package) (73,153).

5.6.3.4 Model Training

Classifier models were trained in the "caret" package in R using the train() (the "method =" argument was determined by the base algorithm, "metric" was set to "logloss" and the "trControl" argument applied). The trainControl() function was used with "method = cv". A 5x cross validation was set with the train and test folds from each indexed for tracking of predictions. The test fold predictions were then saved and averaged to provide individual ROC curves for each outcome class with 1x standard error of the mean (the rationale for this is described in the next section). A manual ROC for each class was generated over a single Multinomial ROC as this provided insight into which classes were best or least confidently discriminated. Additionally, internal metrics on balanced accuracy were obtained using the resamples() function ("caret" package) and averaged across the 5-fold CV models.

The palliative survival model was trained using the rfsrc() training function ("randomForestSRC" package, ntree=1000, "nodesize =" was set based on the tune() function (ntreetry = 200)).

Model hyperparameters for all final models will be provided within the Supplementary Results.

5.6.3.5 Validation and model performance

Internal validation for the treatment classifier models was by k=5-fold cross-validation ("caret" package) to provide estimated generalizability error averaged across test sets in each fold. The final model for each algorithm was then trained on the full training cohort and tested on the full OUH validation cohort (external validation). Classifier models were optimised for log loss during training and their mean-model performance assessed primarily on balanced accuracy (accuracy weighted by class size) and area under the curve (AUC of the Receiver Operator characteristic (ROC)) for each outcome class (one vs rest) using default probability thresholds set by the caret package. As 5-fold cross validation was used (to optimise a balance between sufficient diversity in the test folds without reducing training set sample size unduly) providing 5 sample metrics, 95% confidence intervals are not provided here as they assume a normal distribution (c.f. Kwak et al., 2017 (253)) and the law of large numbers and central limit theorem typically requires at least 30 samples for this to be testable. Importantly, the need for estimating generalisability error within the training set is largely obviated by a truly independent external validation set (oxford cohort) providing a direct assessment of generalisability. A standard error of the mean however is provided across these thresholds on the visual ROC plots for error estimation. To statistically test for differences in performance between classifier algorithms, AUCs were also generated over 1000 bootstraps (models were trained on the bootstrapped sample and tested on the out-of-bag cases). Mean, standard deviation, range and 95% confidence intervals are provided for the bootstrapped model AUCs. Differences in performance between algorithms 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)).

Survival forests were internally validated using bootstrapped resampling (1000 forests, ntree = 1000 per forest) with hyperparameter tuning via the Tune() function. Mean-model performance was assessed primarily on calibration, while additional metrics: Prediction error and Continuous Rank Probability Score (CRPS) are also provided.

Calibration curves were plotted both by quintile (based on survival probability at a single time point), and by event-probability at 3,6, and 12 months ("pec" package). Quintile-based survival curves were derived from mean test-set predictions averaged at each time point across all bootstrapped models and plotted against the corresponding Kaplan Meier (observed) survival probability. Cases were stratified into quintiles based on predicted 1-year survival using the RSF model with Q1 being highest risk (0-20% predicted survival) versus Q5 being lowest risk of

death at 1-year (80-100% predicted survival). The predicted survival over 5 years is then plotted for each subgroup (the x-axis) as 5-year survival is a standard survival metric within oncology. This approach is again based on Rahman et al (73). Quintile-based plots provide evaluation of the model when patients are stratified by risk at a single defined time-point, while calibration plotted at sequential time-points allow for comparison of predictions across the cohort at multiple timepoints. This combined approach offers clearer insight into the optimal operating window for the model longitudinally and by patient-risk.

Prediction error was defined as 1- Concordance (153). Here, concordance is the percentage of observation-pairs where the probability of a true event is greater than a true non-event (a perfect model error rate = 0) (154). Error rate was extracted for each bootstrapped model and averaged.

The Continuous Rank Probability Score (CRPS), (defined as Integrated Brier Score divided by time) is another measure of prediction calibration and derived from the Brier score (mean squared difference between predicted probability and observed probability (155)). In this study it was averaged across all bootstrapped models (153). A perfect model scores 0 and a perfectly inaccurate model scores 1.

Model fairness was not a primary outcome in this study however the impact of age on OC treatment allocation has been previously investigated (109).

5.6.4 Responsible Co-Design

To ensure the applicability and real-world utility of the CDSS we pursued an RRI program in parallel to the CDSS development. Heartburn Cancer UK, a leading charity for oesophageal cancer provided PPI, offering insight into the patient experience. Our approach involved early engagement with clinicians and computer scientists to ensure the tool was clinically relevant, technically sound and user-friendly. Regular RRI workshops were combined with semi-structured interviews using MDT domain experts. These are detailed within the Co-Design section of the Supplementary Methods.

5.6.5 User Interface

Using insights from our RRI program, we developed a high-fidelity prototype of the User interface (UI) using the “Shiny” R package. Trained models were uploaded with their performance metrics, Receiver Operator Characteristic (ROC) curves and a short educational

summary of the performance metrics. The Palliative Survival model is presented using treatment-specific survival curves for the recommended palliative pathway and a user-selectable alternative pathway to provide a visual comparison of the potential prognoses. For classifier models, a Local Interpretable Model-Agnostic Explanation (LIME, “LIME package”) was integrated to provide prediction explanations in real-time. LIME was used within the prototype UI as the package currently supports a diverse array of ML models through the “caret” package.

5.6.6 Role of the Funding Source

The funding sources were not involved in study design, data collection, analysis, interpretation of data or writing of this manuscript.

5.7 Results

5.7.1 Cohort demographics

A total of 1047 eligible UHS cases were identified of which 94 were excluded for endoscopic resection, leaving 953 cases for training the initial model. Within the palliative sub-group (N=439), two were excluded from the palliative-specific models as they were assigned a non-standard chemoradiotherapy regimen. As the initial model does not need to provide a specific palliative treatment however, they were eligible for inclusion within the primary model to maximise training data.

Within the validation cohort, a total of 978 eligible cases were identified and provided by OUH of which 475 palliative cases were identified for validation the palliative models. The Training Cohort (UHS) and the Validation cohort (OUH) are outlined in Table 5.1. Detailed demographic breakdown by outcome class is provided in Supplemental Table 4 while palliative cohort demographics are provided in

Supplemental Table 5.

The two cohorts differed in composition across several variables including age, performance status, cT and cN stage, tumour location, and incidence of chronic pulmonary disease, peripheral vascular disease and cerebrovascular disease (Supplemental Table 6). In summary the OUH cohort presented a typically younger, physically more active cohort despite a higher incidence of pulmonary and vascular disease. The distribution of biological gender, cM staging at presentation, tumour histology and incidence of the remaining co-morbidities were consistent in both cohorts

Table 5.1 Demographics for the Training cohort (UHS) and validation cohort (OUH). Standardized Mean Differences (SMD) are provided for the two cohorts. An SMD of 0.2 is considered a small difference, 0.5 medium and 0.8 or more, a large difference.

Pre-treatment variables	UHS (N =953) (%)	OUH (N =978) (%)	Test SMD
Gender			
Male	718 (75.3%)	744 (76.1%)	0·017
Female	235 (24.7%)	234 (23.9%)	
Median Age in years (Range)	70.0 (21.0 – 96.7)	68 (29.0 – 96.0)	0·156
Performance status			
0	371 (38.9%)	712 (72.8%)	0·726
1	329 (34.5%)	150 (15.3%)	
2	160 (16.8%)	71 (7.3%)	
3	88 (9.2%)	43 (4.4%)	
4	5 (0.5%)	2 (0.2%)	
cT stage			
0	4 (0.4%)	0	0.885
Is	3 (0.3%)	0	
1	7 (0.7%)	2 (0.2%)	
1a	1 (0.1%)	13 (1.3%)	
1b	1 (0.1%)	17 (1.7%)	
2	169 (17.7%)	196 (20.0%)	
3	557 (58.4%)	503 (51.4%)	
4	134 (14.1%)	7 (0.7%)	
4a	37 (3.9%)	138 (14.1%)	
4b	15 (1.6%)	72 (7.4%)	
X	25 (2.6%)	30 (3.1%)	
cN stage			
0	254 (26.7%)	313 (32.0%)	0·340
1	437 (45.9%)	310 (31.7%)	
2	183 (19.2%)	253 (25.9%)	
3	61 (6.4%)	97 (9.9%)	
X	18 (1.9%)	5 (0.5%)	
cM stage			
0	690 (72.4%)	712 (72.8%)	0·047
1	257 (27.0%)	263 (26.9%)	
X	6 (0.6%)	3 (0.3%)	
Tumour location			
Proximal Oesophagus	22 (2.3%)	20 (2.0%)	0·885
Mid oesophagus	102 (10.7%)	176 (18.0%)	
Distal Oesophagus	570 (59.8%)	321 (32.8%)	
Siewert 1	56 (5.9%)	256 (26.2%)	
Siewert 2	124 (13.0%)	205 (21.0%)	
Siewert 3	57 (6.0%)	0	
Siewert undefined	22 (2.3%)	0	
Tissue Histology			
Adenocarcinoma	749 (78.6%)	780 (79.8%)	0·029
Squamous Cell	204 (21.4%)	198 (20.2%)	
Co-morbidities			
Chronic pulmonary disease (CPD)	130 (13.6%)	179 (18.3%)	0·128
Peripheral vascular disease (PVD)	43 (4.5%)	23 (2.4%)	0·119
Cerebrovascular disease (CVD)	106 (11.1%)	44 (4.5%)	0·249
Uncomplicated diabetes (DM uncomp)	128 (13.4%)	155 (15.8%)	0·068
Leukaemia	4 (0.4%)	1 (0.1%)	0·062
Lymphoma	11 (1.2%)	13 (1.3%)	0·016
Renal disease	39 (4.1%)	34 (3.5%)	0·032

5.7.2 CDSS model performance

5.7.2.1 Primary treatment model classification performance

Performance of the primary treatment classifiers were first tested internally on 5-fold cross validation as well bootstrapped resamples after which the models were tested externally on the OUH cohort to confirm generalisability with a single-shot pass through each model (Table 5.2, Figure 5.1).

On internal validation (UHS cohort) all three algorithms exhibited strong classification performance within the primary model. Mean balanced accuracy calculated for multinomial logistic regression (MLR), random forests (RF) and XGBoost (XGB) algorithms were 0·780±0·008, 0·752±0·019 and 0·781±0·009 respectively. On cross-validation the XGB model performed best with a mean AUC across classes of 0·909±0·044 (MLR 0·905±0·048, RF 0·883±0·059) and over 1000 bootstrapped resamples, mean performance was comparable across the algorithms with MLR and RF slightly outperforming XGB (MLR 0·866 (95% CI 0·866 - 0·867), XGB 0·863 (0·862 - 0·864), RF 0·863 (0·867 - 0·868), (Supplemental Table 7 & Supplemental Table 8)), although the differences in mean performance remained modest.

On external validation (OUH cohort) the MLR model generalised best with a validation set mean AUC across classes of 0·894±0·056 (XGB 0·887±0·061, RF 0·891±0·059).

Model performance was separately assessed on an internally validated UHS model incorporating an additional endoscopic resection class. Model performance is provided in Supplemental Figure 10 and Supplemental Table 9.

Table 5.2 - Mean classification performance AUCs for the UHS test set versus OUH validation set. Best performance for each class within the UHS training and OUH validation sets is highlighted in bold.

UHS Model	UHS (N=953) OUH (N = 978)	Chemo	CRT	Surgery	Palliative	Mean (±SD)
MLR	UHS test set	0.893	0.868	0.884	0.976	0.905±0.048
	OUH validation set	**0.895**	0.815	**0.924**	0.943	**0.894±0.056**
XGB	UHS test set	**0.897**	**0.873**	**0.892**	**0.979**	**0.909±0.044**
	OUH validation set	0.894	**0.835**	0.849	0.970	0.887±0.061
RF	UHS test set	0.872	0.835	0.847	0.978	0.883±0.065
	OUH validation set	0.890	0.835	0.865	**0.973**	0.891±0.059

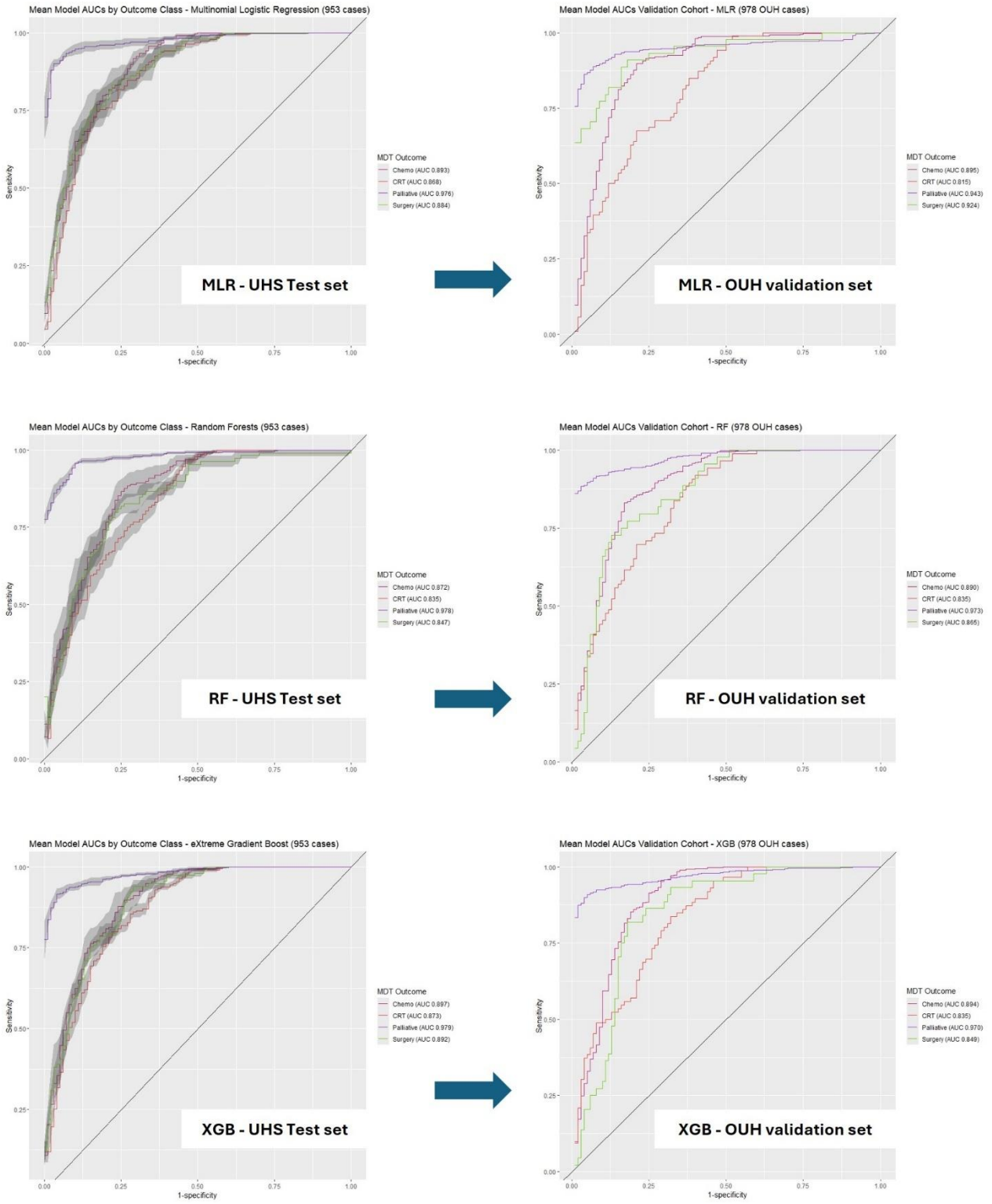


Figure 5.1 - Mean cross-validated ROC curves for each classifier algorithm (UHS vs OUH). Shaded areas represent ±1x Standard error from the Mean.

5.7.2.2 Palliative classifier performance

Palliative classifier performance was assessed in a similar manner. All algorithms performed well in classifying palliative treatment however the XGB models offered best performance on internal and external validation (Table 5.3).

On internal validation, balanced accuracy for XGB, MLR, RF and were 0·690±0·013, 0·689±0·018 and 0·683±0·028, respectively (Table 5.3, Figure 5.2). Mean AUC across classes were XGB 0·815±0·081, MLR 0·805±0·096 and RF 0·793±0·083. Over 1000 bootstraps, XGB again performed statistically best (0·799 (95% CI 0·798-0·800) versus RF 0·781 (0·778 - 0·782) and MLR 0·736 (0·734 - 0·737) (Supplemental Table 10 & Supplemental Table 11).

On external validation Mean AUCs across classes were: XGB 0·742±0·064, RF 0·730±0·077 and MLR 0·711±0·053.

Table 5.3 - Mean palliative treatment classification performance AUCs for UHS (test set) versus OUH validation set. Best performance for each class are in bold.

UHS Model	UHS (N=437) OUH (N = 475)	Chemo	BSC	RTX	Stent	Stent_Onc	Mean
MLR	UHS	**0.909**	0.690	0.730	0.889	**0.805**	0.805±0.096
	OUH Validation	0.780	0.746	0.697	0.687	0.645	0.711±0.053
XGB	UHS	**0.909**	**0.737**	**0.746**	**0.892**	0.790	**0.815±0.081**
	OUH Validation	0.817	**0.794**	0.734	**0.704**	**0.662**	**0.742±0.064**
RF	UHS	0.881	0.713	0.712	0.875	0.782	0.793±0.083
	OUH Validation	**0.821**	0.784	**0.739**	0.670	0.636	0.730±0.077

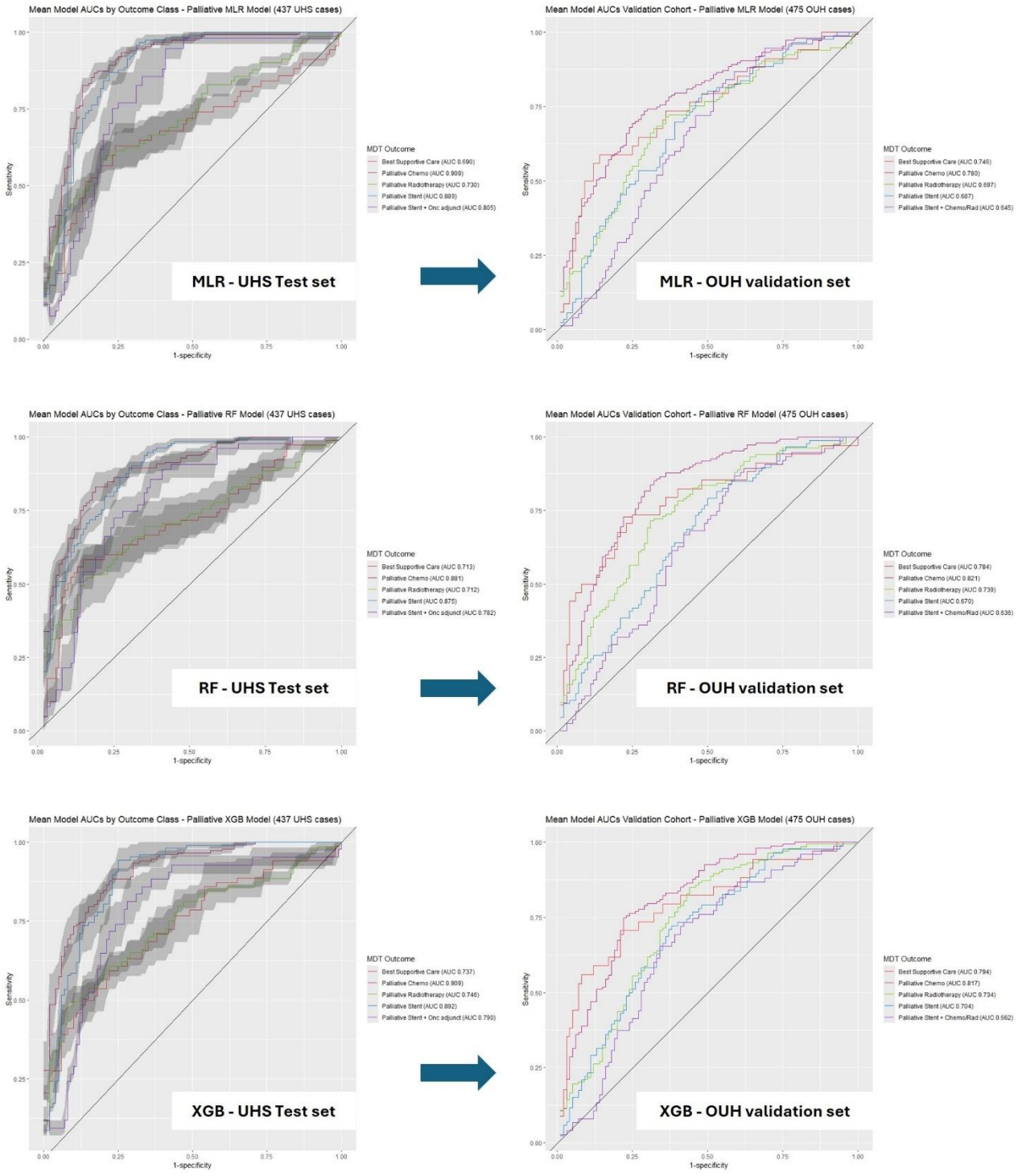


Figure 5.2 - Mean cross-validated ROC curves for each palliative classifier algorithm (UHS vs OUH). Shaded areas represent ±1x Standard error from the Mean

5.7.2.3 Palliative survival model performance

Palliative survival in both cohorts demonstrated significant survival differences between treatments (Figure 5.3). Best median survival was associated with palliative chemotherapy in each cohort (UHS: median 11.1 months (95% CI 9·7-12·2), OUH 11·2 months (9·9-12·9)) followed by radiotherapy and stent ± oncological adjunct. However, while the stent only group survived longer in the UHS cohort, they experienced poorer outcomes relative to the BSC group within the OUH cohort. Supplemental Table 12 details median survival for both cohorts by treatment.

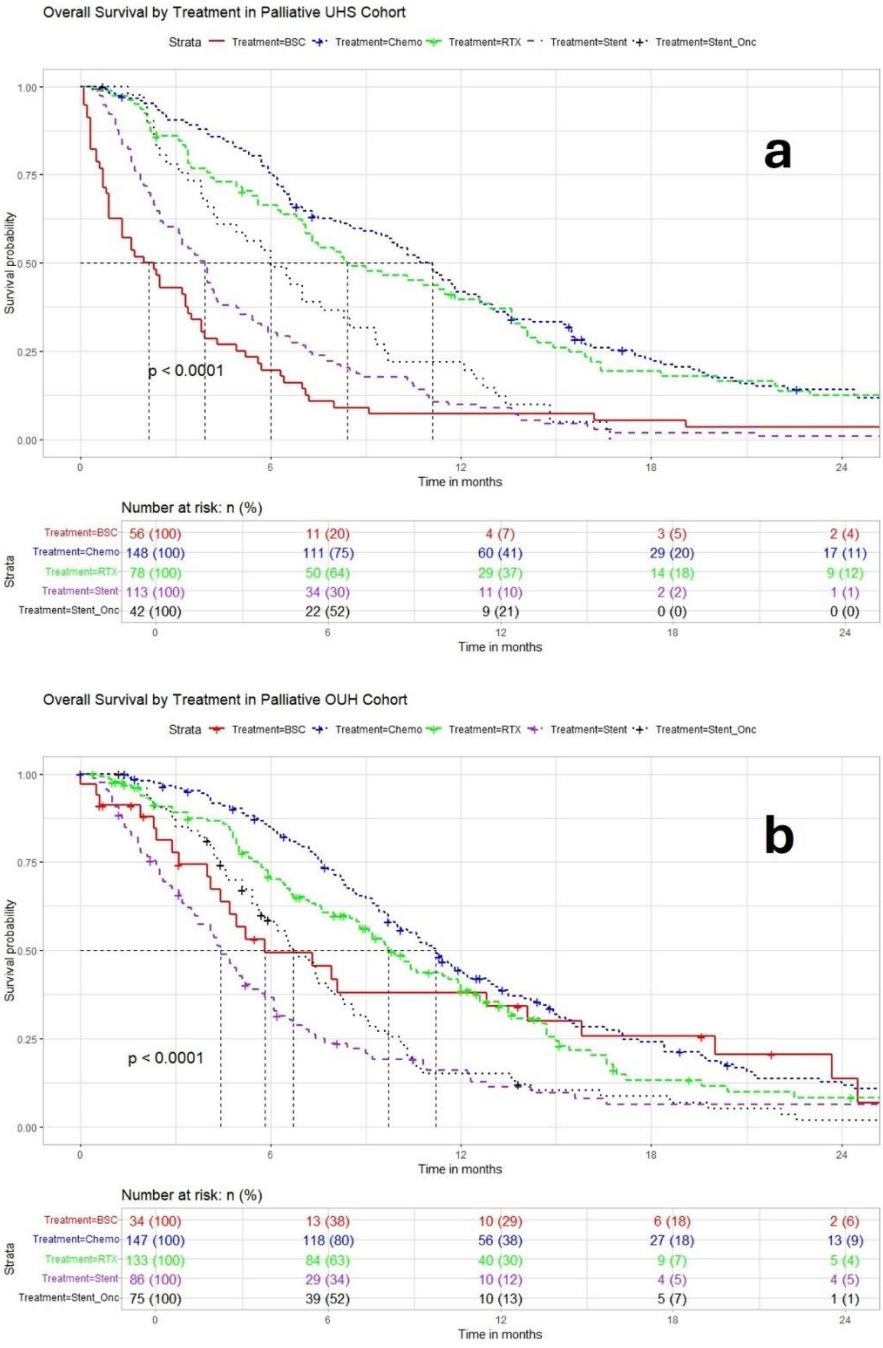


Figure 5.3 - Kaplan Meier palliative survival plots for the UHS cohort (a) and OUH cohort (b).

The final random survival forest model, trained on the full cohort after internal validation, demonstrated a prediction error of 0·331 and a continuous rank probability score (CRPS) of 0·077. On internal validation over 1000 bootstrapped models, mean prediction error was

0·334±0·018 while mean CRPS was 0·112±0·020. This was consistent with the validation cohort (Table 5.4).

Calibration curves were stratified by 1-year survival quintiles (Figure 5.4) as well as by whole-cohort survival at sequential time points where calibration was best within the first 12 months (Figure 5.5). Quintile-based analysis indicated calibration was best for the three highest-risk quintiles (Q1-3). Model predictions were pessimistic for Q4 patients but over-optimistic by Q5.

Table 5.4 - Survival model performance metrics for UHS and OUH cohorts

Metric	**Cohort**	**Score**	**Reference**	**Interpretation**
Prediction error (1-Concordance)	UHS model	0.334±0.017	0 = perfect concordance 1 = perfect non-concordance	Fair
	OUH validation set	0.354		Fair
CRPS (Integrated Brier Score/time)	UHS model	0.112±0.020	0 = perfectly accurate model 1 = perfectly inaccurate model	Very Good
	OUH validation set	0.093		Very Good

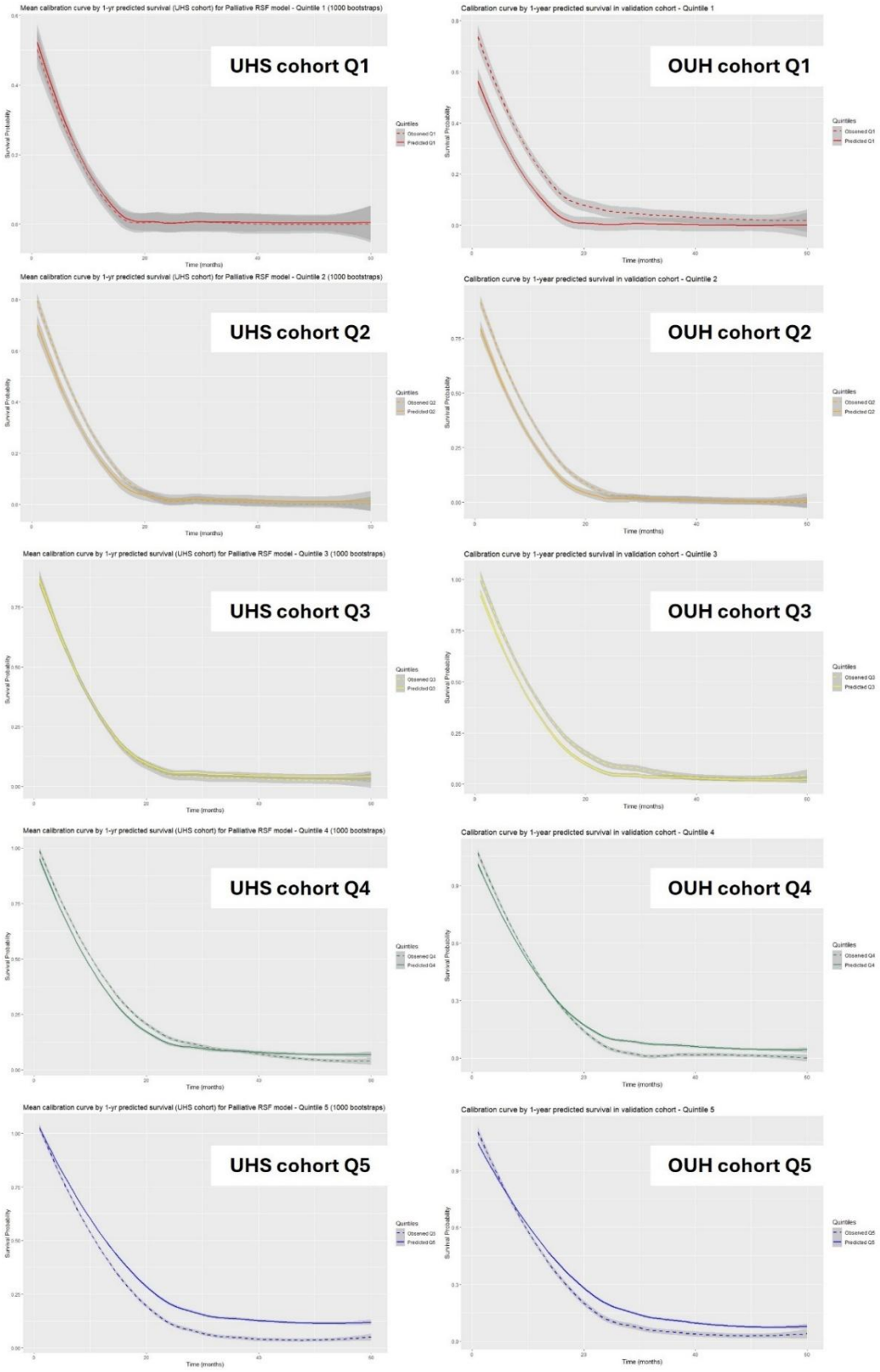


Figure 5.4 - Quintile Calibration curves for palliative survival model plotted with standard error over 60 months Quintile cases are stratified based on predicted 1-year survival probability as determined by the RSF model (Quintile 1 = 0-20% (a), Quintile 2 = 20-40% (b), Quintile 3 = 40-60% (c), Quintile 4 = 60-80% (d), Quintile 5 = 80-100%)

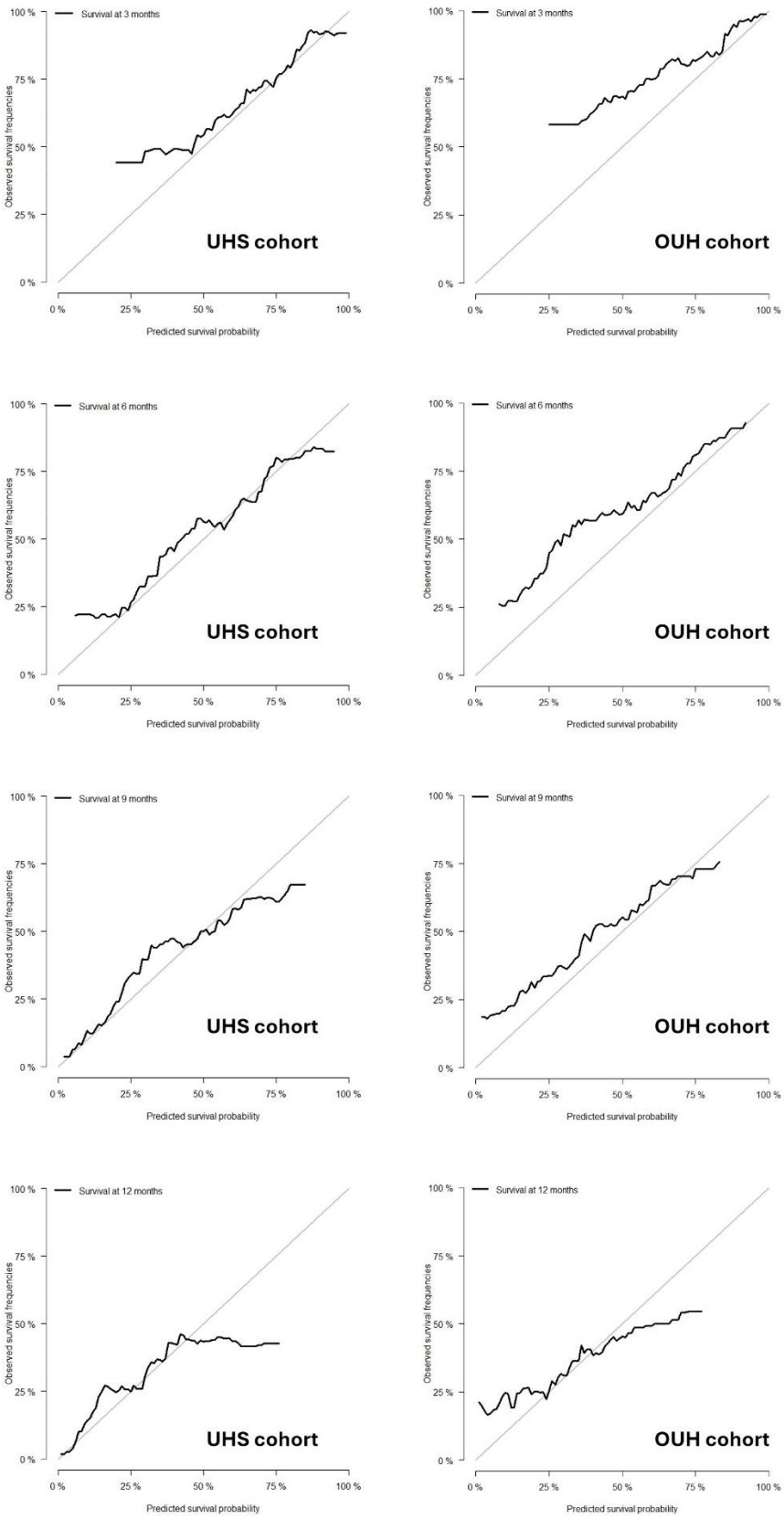


Figure 5.5 - Calibration plots for the UHS cohort versus OUH validation cohort at 3,6,9 and 12 months post-diagnosis.

5.7.2.4 OUH models

To determine if the modelling process remained robust when applied to a non-UHS cohort the same algorithms were again trained using OUH as the training centre and tested on the UHS cohort as the external validation centre. XGB models again performed best on both internal and external validation across treatment classifiers (Supplemental Table 13 & Supplemental Table 14). Similarly, a survival model trained on the OUH cohort demonstrated good calibration within the first 12 months after which predictive performance dropped away (Supplemental Figure 12, Supplemental Table 15).

5.7.3 Co-design insights

The co-design responsible research and innovation (RRI) program was set up to provide guiding insights into user-needs and concerns when implementing a clinical decision support system (CDSS). This was stimulated by discussing prompts from the RRI card deck (Supplemental Figure 13) in combination with insights drawn from our clinician interviews (Interview questions provided in Supplemental materials) and RRI workshops. It highlighted several key challenges and considerations which were factored when developing the CDSS. Themes identified included: bias within the models, data drift, unintended inequalities of access, as well as safety and accuracy from a regulatory perspective. The RRI process recognised the impact CDSSs may have on clinical training for junior clinicians as MDTs are traditionally a source of experiential learning along with a need for education in AI literacy. Explainability proved a recurring theme along with the potential ramifications of group AI interactions where multiple human actors are interacting dynamically with the AI. This in turn prompted considerations over where the ultimate decision-making responsibility lies when a CDSS is supporting high-risk decision-making within healthcare. The themes identified through the RRI process are detailed in Table 5.5 & Table 5.6 along with adjustments we gradually introduced into the tool to address these where possible.

Table 5.5 - Thematic analysis of domain expert interviews highlighting user expectations, concerns and solutions engineered into the tool in response

Clinicians' Understanding of AI and Its Role in Healthcare	Perceived Potential Benefits of AI in Multidisciplinary Teams (MDTs)	Barriers to the Adoption and Trust of AI in Healthcare
Theme 1: Conceptual Understanding and Knowledge Variability The interviews revealed a significant variability in clinicians' understanding of artificial intelligence (AI) and machine learning (ML). Some clinicians demonstrated a deep understanding of these technologies, recognizing their potential and limitations, while others exhibited a more superficial or unclear perception. This disparity in understanding is likely to influence how different clinicians interact with and trust AI tools. As one clinician noted, **"Some of us see AI as just algorithms, but the deeper layers are often misunderstood."** This variability suggests the need for targeted educational initiatives to ensure that all clinicians have a sufficient grasp of AI concepts.	**Theme 1: Improved Diagnostic Accuracy** Clinicians widely recognized AI's potential to improve diagnostic accuracy, particularly by analysing large datasets and identifying subtle patterns that may be overlooked by human clinicians. This capability was seen as a significant advantage, especially in the context of complex diseases like oesophageal cancer. **"AI can help us catch things we might otherwise miss in diagnostics,"** one clinician stated, highlighting the perceived value of AI in enhancing diagnostic precision. **Our Response:** We have ensured the tool's models have been internally and externally validated on a large training cohort	**Theme 1: Concerns About Bias and Accuracy** A significant barrier to the adoption of AI tools identified by clinicians was the concern about bias in AI algorithms and the accuracy of AI decisions. Clinicians expressed worry that AI could perpetuate or even exacerbate existing biases in healthcare, leading to unfair treatment decisions. One clinician emphasized, **"Bias in AI is a serious concern, especially if it leads to unfair treatment decisions,"** reflecting a common apprehension about the ethical implications of AI in healthcare. **Our Response:** We have identified and acknowledged the possibility of bias. The tool provides a detailed breakdown of the training cohort for user inspection. Future iterations will also include warning messages where specific clinical inputs represent cases with few data points e.g. "dementia = Y"
Theme 2: AI as an Extension of Data Analytics Several clinicians perceived AI as a powerful extension of traditional data analytics, capable of processing larger datasets and uncovering patterns that might escape conventional methods. This perception ties AI closely to evidence-based practice, where it is seen as enhancing the clinician's ability to make data-driven decisions. One clinician remarked, **"AI is like**	**Theme 2: Workload Reduction and Efficiency** Another major benefit identified was AI's potential to reduce clinicians' workload by automating routine tasks. This could allow clinicians to focus more on complex cases and patient interactions, thus improving overall efficiency in the healthcare setting. One clinician commented, **"AI could free us from repetitive tasks, giving us more time to focus on patients,"** emphasizing the role of AI in enhancing productivity.	**Theme 2: Transparency and Explainability** The need for transparency and explainability in AI decision-making processes was another critical barrier. Clinicians expressed a strong desire to understand how AI reaches its conclusions to trust and use these tools effectively. **"I need to know how AI makes its decisions before I can trust it,"** one clinician explained, underscoring the importance of AI interpretability in clinical practice.

Clinicians' Understanding of AI and Its Role in Healthcare	Perceived Potential Benefits of AI in Multidisciplinary Teams (MDTs)	Barriers to the Adoption and Trust of AI in Healthcare
data analytics on steroids; it can handle much larger datasets and pick up on trends we might miss."	**Our Response:** We have integrated the ability to upload a list of patients simultaneously and provide a generate report of predictions for each patient in real time. This will allow use of the tool between MDT meetings too	**Our Response:** The tool provides a detailed breakdown of the training cohort for user inspection; it also provides detailed performance metrics of the current models. We have integrated a LIME explanation plot that reactively updates in real time with user inputs to give a specific explanation at an instance level.
Theme 3: Mystification and Misconceptions The interviews also revealed that some clinicians hold misconceptions about AI, viewing it either as an almost omniscient entity or as an unreliable tool. This mystification of AI can lead to polarized views—some clinicians might place undue reliance on AI, while others might harbour unwarranted scepticism. As one participant explained, **"Some think AI is this magical tool that can do anything, while others don't trust it to do anything right."** These misconceptions underscore the importance of clear communication about what AI can and cannot do in clinical settings. **Our Response:** We recognise that there remains an ongoing knowledge gap for Clinicians in the MAI sphere. While this is an evolving field, we have sought to assist the AI-lay clinician using the tool by providing a section which outlines some key metrics and a guide of their interpretation to allow them to critically appraise our model performance.	**Theme 3: Enhanced Decision Support** Clinicians also saw AI as a valuable tool for enhancing decision support, particularly in complex cases where multiple variables need to be considered. The ability of AI to process and analyse data rapidly was viewed as a way to formulate more comprehensive and informed treatment plans**. "AI could be a valuable assistant in making decisions in complicated cases,"** one clinician noted, underscoring the potential for AI to augment clinical decision-making.	**Theme 3: Impact on Clinical Autonomy** Concerns were also raised about the potential impact of AI on clinical autonomy. Clinicians worried that an over-reliance on AI might diminish the role of human judgment in decision-making, leading to a reduction in their autonomy. As one clinician put it, **"I'm worried that AI might take away our decision-making power, making us too dependent on it,"** reflecting a fear of losing control over clinical decisions. **Our Response:** We recognise this is a valid risk of automating a clinical decision-making framework like the MDT. Where the tool generates reports for a group of patients at once, it orders them in order of confidence in the recommendation. A traffic light system then signposts clinicians to cases of low confidence where the human is required to assess and recommend. This keeps the human central to the process for discussing those most difficult cases first and sense-checking high-confidence cases thereafter.

Clinicians' Understanding of AI and Its Role in Healthcare	Perceived Potential Benefits of AI in Multidisciplinary Teams (MDTs)	Barriers to the Adoption and Trust of AI in Healthcare
	Theme 4: Personalized Medicine The potential of AI to advance personalized medicine was another recognized benefit. Clinicians appreciated AI's ability to tailor treatments based on individual patient data, which could lead to better outcomes. **"With AI, we could move closer to truly personalized medicine, where treatments are tailored to the individual,"** one participant remarked, highlighting the transformative potential of AI in this area.	**Theme 4: Legal and Liability Concerns** Legal and liability concerns were also prominent among clinicians. They were unsure who would be held accountable if an AI tool made a mistake—whether the responsibility would fall on the clinician using the tool or the developer who created it. **"Who's responsible if AI makes a mistake? This is a big question to ask,"** one clinician stated, highlighting the legal uncertainties surrounding AI adoption. **Our Response:** We acknowledged the implications from a ethicolegal perspective and have included a disclaimer message at first use which explains clearly that responsibility of the decision remains with the human as it is a decision-support tool. This will remain the case even if certification as a medical device is
	Theme 5: Predictive Analytics for Preventive Care Clinicians acknowledged the potential of AI in predictive analytics, particularly for preventive care. AI could be used to identify patients at risk of certain conditions, enabling early intervention and improving patient outcomes. One clinician noted**, "AI could help us predict and prevent diseases by identifying at-risk patients earlier,"** indicating the proactive role AI could play in healthcare.	**Theme 5: Fear of Losing Skills** Finally, some clinicians expressed a fear that the adoption of AI could lead to a loss of skills, particularly in routine diagnostic tasks. There was concern that AI might replace certain aspects of their work, leading to skill degradation, especially among less experienced clinicians. **"There's a fear that AI could replace us in certain tasks, which may make some juniors lose some important skills,"** one participant observed, pointing to a potential unintended consequence of AI integration.

Table 5.6 - Thematic analysis of RRI workshop outlining additional themes extracted on user expectations, concerns along with solutions engineered into the tool in response where relevant

Themes	Computer scientists	Patient and Public Involvement
	Theme 1: Explainable AI (XAI) A major focus of the discussion has been on the tool more interpretable, leading to developments in explainable AI (XAI). This ensures that models can be understood and trusted by non-experts. *Supporting Code:* "*We've built tools that provide explanations for AI decisions, making it easier for users to trust the system."*	**Theme 1: Ethical concerns** This includes issues such as data privacy, and fears of AI exacerbating inequality, might act as barriers to public acceptance. Supporting Code: "Who controls the data when AI is involved? This is my biggest concern." **Our response:** By working symbiotically with the local hospital who provides the clinical data we also ensure it is stringently protected, anonymised as soon as possible and quality checked by the clinicians within the team. The presence of clinicians within the research team also ensure that the patient is the priority even with data storage and collection.
	Theme 2: Visualizations and Diagnostic Tools Tools like saliency maps, LIME, SHAP, and attention visualization have been discussed to provide insights into why AI models make specific decisions. *Supporting Code:* "By visualizing how models interpret inputs, we make AI decisions clearer and more understandable to end users."	**Theme 2: Complexity and lack of understanding** This may prevent the general public from engaging fully with AI tools. *Supporting Code:* "People think AI is too complicated to understand, so I think they may feel uncomfortable using it." **Our response:** We have incorporated a section which briefly outlines some of the technical information in more accessible terms. While this UI is designed to be used primarily with clinicians the principle also extends to patients being shown the UI outputs and hopes to enhance Ai literacy for patients and clinicians alike
	Theme 3: Techniques for Bias Detection and Reduction Computer scientists have also suggested ways of identifying and reducing biases in AI models, particularly in sensitive areas like healthcare. Techniques such as fairness constraints and debiasing were examples. *Supporting Code:* "We may incorporate fairness constraints into the training process to mitigate biases against underrepresented groups."	
	Theme 4: Creating Diverse Datasets Recognizing that biases often stem from the data itself, there has been a push toward creating and curating more diverse datasets that better reflect the population. *Supporting Code:* "We may need to think of other datasets to ensure AI systems perform fairly across all demographics."	

5.7.4 User interface

To generate a User Interface that could be implemented clinically, we compartmentalised user interactions into three main areas using the R Shiny platform. No significant prior training is needed to allow users to generate an output, with walk-through tutorials built into the main interface and continued across each page of the tool. Users are simply required to select their preferred pre-loaded model and the clinical input data.

Patient variables are inputted for the primary treatment model in the first instance where an instant recommendation is then provided along with predicted probabilities for all potential outcomes to illustrate how confident the final recommendation is (Figure 5.6). LIME explanations are also given for the final prediction.

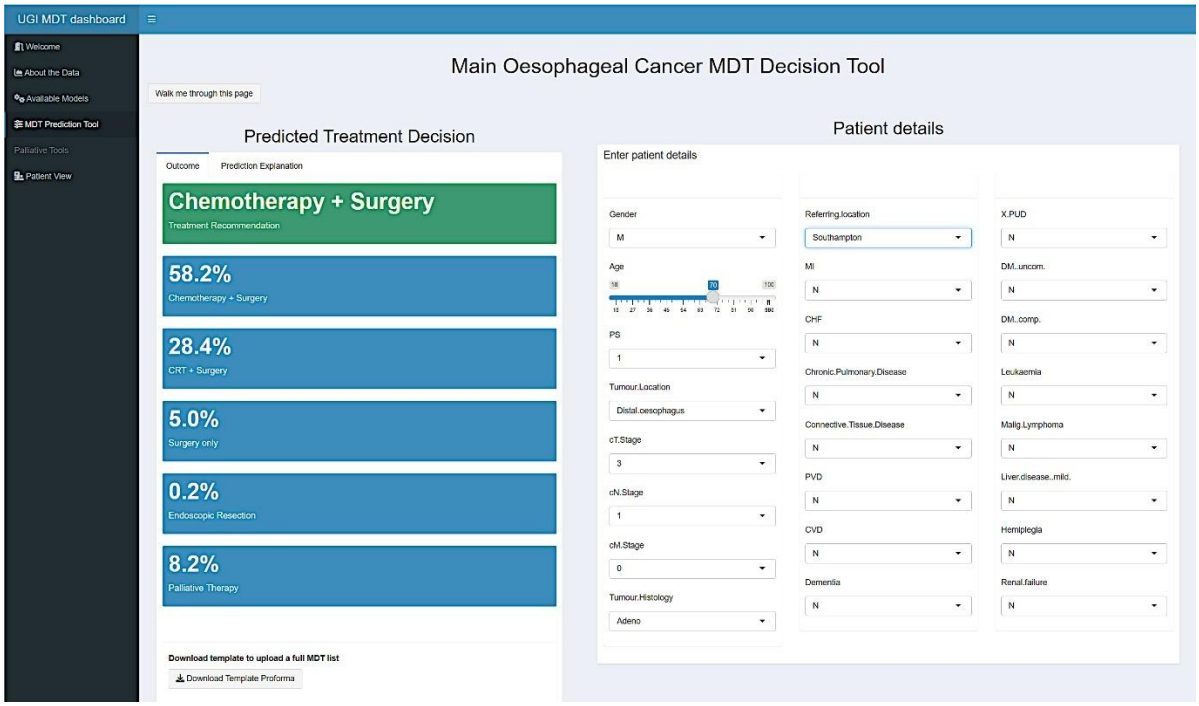


Figure 5.6 - Primary Model interface and input screen

If the outcome is "Palliative", the UI automatically carries the inputs across to the palliative treatment classifier (Figure 5.7). As with the primary model, a LIME explanation is available for the user along with performance metrics for whichever model was loaded (RF, MLR, or XGB).

For palliative treatments, the associated predicted survival curve is then automatically generated along with an option to compare survival with an alternative pathway. In the example illustrated in Figure 5.8, the original recommendation was for palliative chemotherapy which was then compared to palliative radiotherapy. The survival curves effectively personalise to the level of the treating hospital from which the training data was derived.

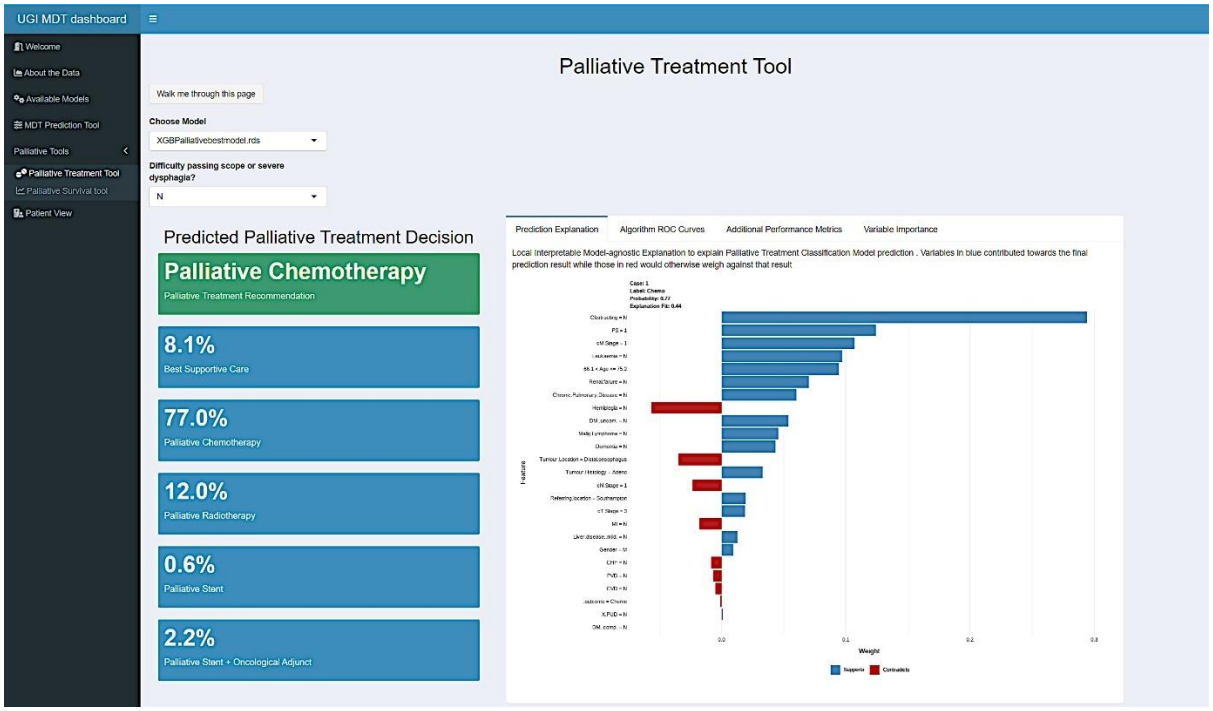


Figure 5.7 - Palliative model recommendation and associated LIME explanation

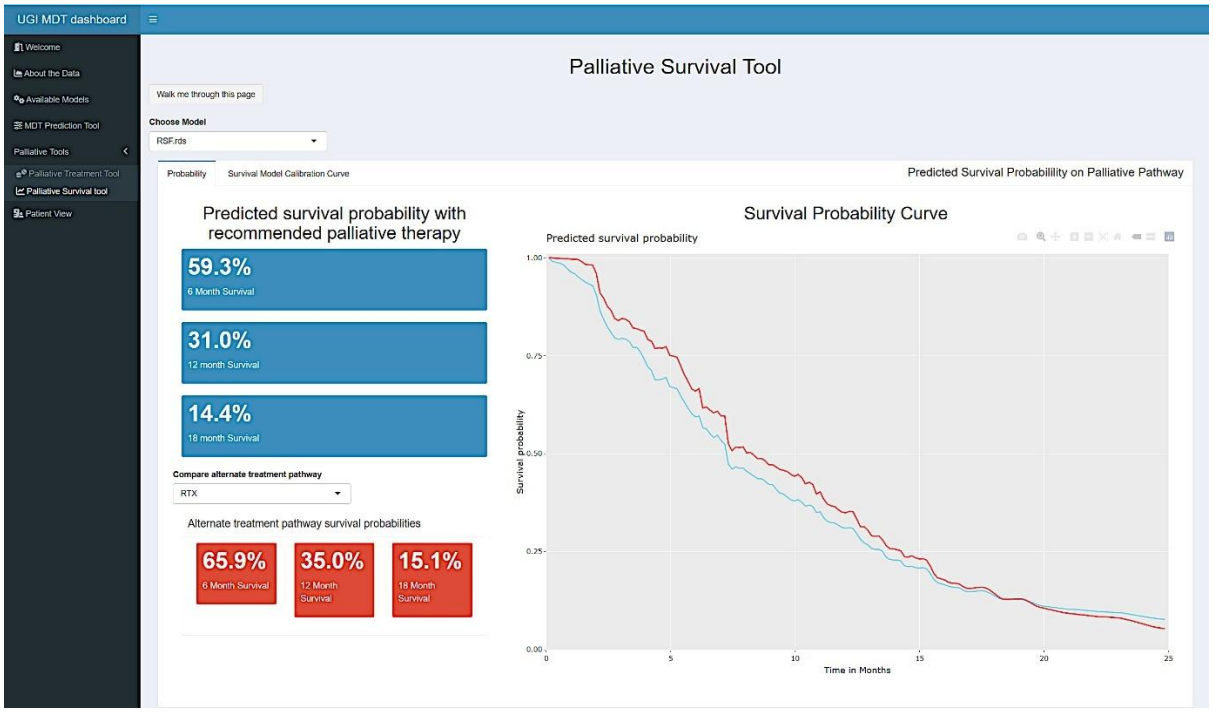


Figure 5.8 - Palliative survival curves specific to recommended or selected treatment plans

5.8 Discussion

We present the first externally validated ML CDSS co-designed using RRI principles, capable of predicting OC MDT treatment decisions early within the cancer pathway. The sequential modelling approach quickly predicts a new patient's probable treatment plan which if palliative, is accurately prognosticated within the first 12 months post-diagnosis. All algorithms performed well; however, our results particularly favour MLR and XGB models, with mean AUCs above 0·87 for the primary classifier, and above 0·711 in the palliative classifier. The RSF model performed well within the first 12 months on calibration curve analysis and CRPS scores. Furthermore, the models have shown they generalise even when faced with differing cohort demographics. This suggests that for predictions at an instance-level, the ML models appear to handle perturbations in demographics at the feature-level. The use of a parallel RRI program ensured that the design of the AI CDSS has considered stakeholders and integrated their input. Early collaboration with a diverse skill-mix and open-format workshops have produced a CDSS with immediate clinical translational potential.

The primary treatment classifier models performed consistently well in discriminating palliative pathway patients across algorithms. This is primarily driven by the large palliative subgroup within the training process, a largely binary influence of cM staging and the additional input of high PS score patients. Across algorithms the models predict the curative pathways evenly however within the validation cohort (OUH) the CRT class demonstrates lower predictive performance. This is attributable to the relatively high use of CRT at UHS versus the OUH unit which favours neoadjuvant chemotherapy preferentially outside of squamous cell cancers. Until recently, chemotherapy and CRT have been equally acceptable options however early ESOPEC trial results now suggest chemotherapy may be the long-term front runner in these non-SCC cases (11). Where misclassifications occur this discordance between units favouring chemotherapy versus chemoradiotherapy is likely to be the main source. It is also important to recognise the differential in performance between the full primary classifier, which is trained off the full cohort, versus the palliative models which are only trainable on the palliative sub-group comprising approximately 50% of the training cohort. Classes such as palliative radiotherapy and best supportive care are innately harder to predict early on (radiotherapy is most commonly utilised for symptom control (dysphagia, pain, bleeding)) while chemotherapy is most prevalent as it provides disease control. BSC is determined on a combination of disease stage, physiological reserve and most importantly, patient wishes (the latter typically only determined after the MDT meeting when patients are seen in clinic).

The findings of this study continue to support the role of ML in oncological MDT decision-making, even early within the care pathway. The preserved performance on external validation indicates that overfitting remains modest. The optimal use of the palliative survival model was localised within the first 12 months post-diagnosis, in keeping with the expected survival for this cohort on current best therapy (158,159). As median survival was 6 months across the entire palliative cohort, and maximally 11 months with best palliative oncological therapy (chemotherapy) we have to acknowledge that predictions beyond the 1-year mark will not be reliable and return to best clinical judgement for those few who survive beyond this time point. As core management of OC is well established within national guidelines, ML lends itself to modelling the UK, decision-making framework, however this principle should in theory translate internationally as well (122). Where model performance drops between the cohorts this likely reflects idiosyncrasies specific to individual centres, an observation previously made in Denmark (24).

There is a clear and urgent need to support MDT workflow. With 60% of new discussions likely to end in palliative treatment plans, rapidly predicting, and prioritising caseload is of clear clinical and financial benefit. When developed and implemented correctly they also have the capability to improve patient safety (45). Yet while AI-based CDSSs offer much to the healthcare sector, there remains a translational gap (254). This is multifactorial in nature but partly attributable to a sense of clinician superiority especially when handling nuance and uncertainty, or a lack of clinical validation (255). Furthermore, as the current boom in healthcare AI continues, the need for more responsible, co-designed and explainable AI (XAI) approaches are increasingly paramount (55,256). This study addresses this by establishing external validity of our clinical models combined with a co-designed user interface. Algorithms were chosen as either inherently interpretable (MLR) or amenable to both global and local XAI techniques.

The RRI program delineated potential challenges and barriers to the use of AI based CDSSs in clinical settings, including data bias, data governance, data drift, regulatory concerns over the role of the human-in-the-loop, and the foundations of legal and clinical responsibility (54,58). The new EU AI Act unsurprisingly classifies MAI into the “high-risk” category especially when the impact of new technologies may still be emerging far downstream of their first deployment (The “Collingridge” Dilemma) (257,258).

Our study employed one of the largest patient cohorts within the literature, providing robust internal and external validation of our models which span curative and palliative pathways. The

algorithms are off-the-shelf libraries, ensuring that scalability of implementation is not reliant on high-performance computing clusters, something the current NHS digital infrastructure cannot offer evenly across the UK. The integrated RRI program was designed to act in the best interests of patients and clinicians. Including stakeholders early we have developed a highly functional CDSS which can be rationalised on a user-specific basis. The UI allows clinicians to counsel patients while insights derived from the program also allow development of future iterations of the CDSS. Consequently, the present study presents the first, cohesive, responsibly derived ML solution to assist OC MDT workflow needs which has not been provided previously within the literature. This is also the first study to externally validate OC treatment allocation models building on our previous efforts to integrate explainability within the process (63,108,109,195). Importantly, while previous studies have highlighted the utility of ML in other conditions (41–44), yet there often lacks a clear roadmap to guide the transition from technical demonstration to active clinical application. Here we have sought to provide a working tool that can be deployed online quickly and used by clinicians not specifically trained in ML.

One limitation was that the endoscopic management of early cancers had to be excluded from the validation analysis as we could not ensure a consistent selection criteria within the external cohort. Many cases are identified through Barrett's surveillance programs, and their care is not necessarily initiated by the MDT in the first instance making consistency of case presentation difficult. Additionally, we could not include novel molecular markers or immunotherapies within this generation of models as insufficient training data was available. Future iterations will support an expanding array of systemic treatments such as Chemotherapy +/- anti HER2, anti-PD-1/PD-L1, Claudin 18·2, and immunotherapies for MMR-d/MSI-H tumours (164–166). As a newer cohort of patients emerge accruing data in these biomarkers, it is conceivable that these cases will be used to train a smaller model on just those features, the predicted probabilities of which may then be fed into a larger model leveraging the main cohort for whom those biomarkers may not have played a part in their treatment, reconciling the separate training datasets. Similarly, the recently reported ESOPEC trial may narrow down indications for NACRT, and future iterations will readily adapt to such trial outputs (11). While early curative cohort prognostication would be desirable (ideally prior to treatment initiation), the temporal effect of two separate major interventions (neoadjuvant therapy and subsequent surgery) make it extremely challenging in a single static model without post-operative inputs (73,259). It is also important to note that much of the data fed to MDTs may be recorded by non-clinical personnel or those of varying oncological experience especially in evaluating PS scores for patients. By way of example, while the OUH cohort may represent a fitter cohort, it is equally conceivable

that less fit patients were either screened out pre-MDT in this unit or assigned lower PS scores erroneously. This also extends to data input as a whole, where prediction quality is inextricably linked to the quality of this input. While algorithms such as RF and XGB are capable of handling missing data, the user interface is designed to ensure all fields are completed. Fields set to a default and if left un-touched will still allow a prediction to be generated, however, it sits with the end-user to ensure that final inputs are correct else the prediction quality may be affected. Model fairness is not directly addressed within the scope of this study. Within the feature set only gender and age are protected characteristics, the latter of which we have previously investigated (109). However, it necessary to recognise that advanced age carries risk and clinician experience may easily be confused for bias in this context (227). Gender remains vulnerable to bias in OC too, which is historically a male-centred condition (260). Assessing model fairness regarding gender however requires assessing the equitability of the predicted outcomes which was beyond the scope of this study and evaluating long-term fairness of models will require more clarity in the definition of "equity" within OC treatment allocation. Finally, we have consciously chosen to map the current MDT versus an attempt to model the "best decision". There remains no single, quantifiable metric currently agreed within OC to adequately encapsulate the myriad outcomes important to OC patients. Survival may not in every case be the most salient outcome measure, yet it is by far the most prolific in quantifying treatment "success" of oncological strategies. It is intended to be a springboard towards composite metrics which consider quality-of-life, complication rates or even resection margin status. Meanwhile, for this technology to translate to clinical use, we must first prove capable of mapping what "is" while the field attempts to agree upon what "should be".

Future work in this field will look to integrate many of the novel markers discussed previously, as well as develop additional co-designed patient-only user interfaces. Broadening external validation to additional centres will further verify the results reported in this study. Trust must be established with patients, clinicians and regulators alike, and this study now sets the foundations for prospective trials within real-life scenarios to smooth the way towards clinical implementation. With the introduction of the EU AI Act, the regulatory landscape for medical AI continues to shift. Satisfying regulatory hurdles moving forward will almost certainly involve risk management, data governance, transparency, human oversight (and override mechanisms), post-market surveillance, quality management systems and CE marking among other considerations (56). Additional work will also be required to test such CDSSs in real-time clinical application. This will provide insight into if such a tool functions best when used within the MDT meetings or if is best utilised between meetings to triage discussions and "pre-screen"

cases. Finally, an aspect lacking within the current literature is investigation into the decision-making thresholds for human agents faced with AI-based predictions in clinical settings across a range of machine confidence levels – at what confidence level is a clinician willing to accept and trust a prediction? And is this "line in the sand" equivalent for every use-case, patient or treatment? This will guide future Medical AI researchers when validating their model performances.

This is the first co-designed externally validated AI-derived CDSS targeted towards decision-making within the MDT cancer pathway for oesophageal cancer. It provides an integrated sequence of ML models which can reliably predict treatment allocation and palliative prognosis both locally and externally. The integration of an RRI program is intended to enhance user confidence that the CDSS considers individual and society risk as well as sources of potential bias within its design. Such technologies must contend with the standard challenges facing workflow integration within current digital healthcare infrastructures, as well as achieving clinician buy-in, especially where such models may adversely impact future clinician training. While future work includes prospective trials for real-world validation and regulatory approvals to address this, these models offer potential for a transformative impact on current MDT operations within the UK in OC and is both theoretically and technically transferrable to other cancer types and world regions.

5.9 Contributors

NT - conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing– review & editing.

MN - data curation, formal analysis, methodology, visualization, writing – original draft, and writing– review & editing.

MT - conceptualization, formal analysis, methodology, software, validation, visualisation, and writing– review & editing.

SAR - data curation, formal analysis, methodology, software, validation, visualisation, and writing– review & editing.

SLH - methodology, project administration, and writing– review & editing.

CP - data curation, formal analysis, and writing– review & editing.

ZSW - funding acquisition, investigation, methodology, project administration, supervision, writing – original draft, and writing– review & editing.

SR - methodology, resources, supervision, and writing– review & editing.

SM - data curation, investigation, methodology, resources, validation, and writing– review & editing.

RO - data curation, investigation, methodology, resources, validation, and writing– review & editing.

NM - data curation, investigation, methodology, resources, validation, and writing– review & editing.

TA - investigation, methodology, validation, and writing– review & editing.

ZB - investigation, methodology, supervision, validation, and writing– review & editing.

EVP - conceptualization, investigation, methodology, resources, supervision, validation, and writing– review & editing.

MM - conceptualization, data curation, investigation, methodology, resources, validation, and writing– review & editing.

TJU - conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft, and writing– review & editing.

GV - conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

Authors NT and GV had access to and verified the underlying data

All authors have read and approved the final version of the manuscript for publication

5.10 Data sharing agreement

Data Availability: The data included within this study relates to sensitive intellectual property and cannot be shared freely at the present time

Code Availability: The code used within this study includes sensitive intellectual property and cannot be shared freely at the present time

5.11 Declaration of interest

The Authors have no conflicts of interest to declare.

5.12 Acknowledgements

NT receives a joint studentship from the Institute for Life Sciences (University of Southampton) and University Hospital Southampton. The project receives additional funding from the UKRI Trustworthy Autonomous Systems Hub (TAS Hub) Pump Priming Fund. The funding sources were not involved in study design, data collection, analysis, interpretation of data or writing of this manuscript

Chapter 6 Discussion of findings

Within this body of work, I have shown that machine learning is capable of replicating and automating team-based oncological decision-making. While it is naturally dependent on the quality and quantity of training data, the recording and databasing of cancer patients within the current UK health system is a standard process and thus a feasible route for acquiring this data. For the technology to translate and be implementable within current practice it is essential however that it can be shown to be reliable, accurate and at the very least explainable. The algorithms explored within my work are computationally inexpensive relative to advanced deep learning platforms, able to handle smaller datasets and relatively inexpensive as any real-world clinical assistive decision tool needs to operate within current electronic healthcare infrastructure.

6.1 Classification performance of curative and palliative MDT models

Structured tabular datasets comprising simple clinicopathological data are suitable for classification tasks even within relatively modest datasets as demonstrated by my initial study trialling treatment classification in curative oesophageal cancer patients. Within this pilot cohort, the data was found to suit multinomial logistic regression, a simpler, interpretable technique, even favouring it above more advanced classifiers such as random forests or XGboost. While advanced algorithms are typically expected to outperform simpler algorithms in complex tasks, where data is high quality and well-structured well even basic classifiers may perform comparably with advanced algorithms (261).

We observed consistently similar receiver operator characteristic (ROC) curves regardless of the parent algorithm, reflecting an underlying trend within the patient cohort itself. This was best seen within the predictive performance for the Neoadjuvant Chemotherapy (NACT) + Surgery and Neoadjuvant Chemoradiotherapy (NACRT) + Surgery classes. The observed drop in performance discriminating these groups was attributable to several factors such as the limited training data, absence of non-curative outcomes with which to compare scenarios during learning, and finally: uncertainty within the clinical field itself. Evidence for the survival benefit of neoadjuvant therapy in locally advanced oesophageal cancer is well established (9,127–129) yet the superior neoadjuvant modality (for adenocarcinoma in particular) remains contentious

(11,104). It follows that historical clinical equipoise in the field may also be picked up on during model training however going forward it is likely that this will change with the long-term outcome data from the ESOPEC and MATTERHORN trials (the latter combining perioperative FLOT chemotherapy and Durvalumab, an immune checkpoint inhibitor) (223,262).

Within palliative treatment classification tasks, the dataset suited XGboost models best, however, again as with the curative cohort, simpler logistic regression-based models still performed comparably. All models discriminate palliative chemotherapy, stent-alone or stent + oncological adjunct effectively while best supportive care and palliative radiotherapy proved challenging by comparison. Palliative radiotherapy in non-curable oesophageal cancer is primarily indicated for disease-control, bleeding from the primary tumour, or relief of severe dysphagia (263). While severe dysphagia was factored into these models (and infers a need for disease control), I was unable to collect reliable data on tumour bleeding which may account for the difficulty in predicting this outcome class at the same level. Predicting those intended for best supportive care is yet more challenging as this outcome is typically reached through a combination of clinician experience and patient preferences. Patients deemed unlikely to cope with the rigors of palliative therapies or who simply do not wish to subject themselves to this course of action are then offered symptom management. Reaching this decision requires a face-to-face review with the clinician or specialist nurse in charge of their care to assess the patient's physical fitness and preferences. As this typically occurs after MDT deliberations this information is often unavailable early, translating to uncertainty within ML models intended for use at this stage of the care pathway.

6.2 Palliative survival modelling

While treatment assignment is obviously a key MDT function, these decisions are intimately tied with clinical outcomes. Survival is not the only metric of patient outcome within oncology but is perhaps the best known and most prevalent within the literature. The ability to predict survival in advance carries obvious benefit in informing difficult clinical decisions. Predicting curative survival at the pre-treatment time point remains hugely difficult for curative oesophageal cancer patients as we must account for the separate effects of both neoadjuvant therapies and complex surgeries. Incorporating post-operative data within the model as was employed by Rahman et al., addresses this very well however such models cannot be used prior to the patient completing their treatments (73). This becomes less troublesome however within palliative settings as only one major treatment paradigm is tested in relation to the

outcome. A key benefit of leveraging the random survival forests architecture within my palliative survival model is its powerful ability to predict survival probabilities over a range of time-points allowing the user to decide and set their preferred forecasting time frame.

When we combine risk-based calibration with standard calibration analysis at defined time points we find that my model's forecasting was most reliable prognosticating over the first 12 months post-diagnosis and within the highest-risk quintiles (Q1-3). The model was 'pessimistic' for Q4 and overly 'optimistic' in Q5 patients which may be explained by the limited training cohort combined with difficulty forecasting low-risk patients within a traditionally aggressive disease process. Limiting predictions to the first 12 months is pragmatic given a median survival of 6.3 months, with only 6% of patients alive at 2 years. It will be of material benefit for clinicians counselling patients on their prognosis as it also tailored to their own cancer unit. Historically prognosis has been based primarily off national statistics which are averaged over the entire national population. This cannot adjust for nuance in oncological practice or procedural technique within individual units (148). Furthermore, as model training incorporates treatment type, it becomes possible to generate different survival forecasts for the same patient by varying their planned treatment allowing them to observe the consequences of the "road not taken".

6.3 Model explainability using XAI.

The influence exerted by individual variables within ML models was a key interest for this research because it speaks to the importance of matching the ground-truth validity for any models attempting to reproduce human-based decision-making paradigms especially in the clinical domain. Focussing on models which are either interpretable or explainable are an effective tool towards building trust within the intended human-AI interactions (55).

Variable importance measures have offered one such means of deriving global insight into my models. For the curative pathway classifiers, significant importance was naturally assigned to T-stage, N-stage, performance status, tumour histology and tumour location in all models. This corresponds with a-priori domain expertise and national guidance on the management of oesophageal cancer. All tree-based palliative classifiers also valued age and cM-stage, with XGBoost and random forest models, recognising the importance of 'obstructive' clinical signs and Performance Status (PS) scores greater or equal to 3. NICE guidance (NG83) recommends stenting for luminal obstruction and relief of dysphagia, as well as combination chemotherapy in those with advanced oesophagogastric cancer, minimal comorbidity, and a PS score of 0-2

which aligns with findings (122). For survival models, treatment choice proved most important. As survival benefit here is typically a function of both patient selection and treatment effect, this further illustrates the benefit of non-linear ML-based methodologies such as tree-based models that can account for these interactions (157).

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 disparity most likely stems from the relatively few instances of co-morbidities such as haematological/hepatic and connective tissue disorders meaning that within a linear model they may exert a disproportionately high influence even with shrinkage of the parameter weights through L2 regularization. The higher instances of training cases for pulmonary and diabetic disorders by comparison provides for more credible weighting of these within the tree-based models. When we examine the palliative classifiers, chronic pulmonary disease again, featured more prominently along with dementia and renal failure. Notably, this did not translate to the palliative survival models suggesting that while comorbidities hold some influence in assessing treatment eligibility, these may exert a more modest effect on prognosis within the palliative cohort.

A key benefit for applying explainable AI to oesophageal cancer management is its ability to challenge pre-conceived notions about decision-making dynamics through hidden patterns within the data. Identifying a noticeable role age played within my models early offered me an opportunity to focus more closely on delineating how age interacts with the probability of a given treatment plan both in isolation and in relation to other key variables. It ranked most significant on variable importance measures to all tree-based curative classifiers and ranked either 1st or 2nd within the palliative classifiers. Unexpectedly, it proved much less important to the logistic regression models in curative and palliative settings despite performing highly among the tested algorithms. More surprisingly still, when age is removed from the curative feature-set, all algorithms (including logistic regression) demonstrated a reduction in performance for discriminating those offered surgery-alone or combined with NACT though the decision to offer NACRT with surgery appeared largely unaffected by age (with the sole exception of decision tree modelling where it actively increased performance for NACRT). These results seem at odds at first inspection however it is useful to consider the process by which variables are handled and featured within the different modelling approaches to better understand the likely cause. The logistic regression models used in my work were penalised with L2 regularization. Here, variables weights for features which are felt to contribute less to

the overall model are shrunk in the final equation to avoid a disproportionate influence on the predicted class probability. However, while LASSO regularization may shrink unhelpful features down to zero if needed (and in doing so remove them entirely), L2 or “ridge regression” preserves all original variables even if they are minimised, on the basis that all variables may provide useful information to the final model. Tree-based ensemble learners by comparison typically use a smaller subset of features to train individually weak learners which on aggregate produce well performing models (and by randomizing the feature subset in each tree they tend to minimise overfitting more than the linear models). Consequently, linear and tree-based models may adopt different routes to arrive at the same conclusion and yet still achieve similar discriminative performance in the process. Age has historically been associated with increased frailty (measured in part by performance status) and higher rates of co-morbidity (264). It is not unreasonable therefore to theorise that tree-based models may attempt to leverage age as a more efficient “composite” feature as a surrogate for frailty and comorbidity. We observe evidence of this effect within the variable importance plots where the logistic regression model takes in information from most of the feature pool, in comparison to the tree-based models where the bulk of importance is distributed within the top 5-7 features.

If we accept the weighting of age within the tree-based models to be a true reflection of the current MDT, why then might it be ranked so high? It is tempting for many clinicians to believe age does not play a role within their practice. Yet it was this very same historical bias toward elderly patients which previously led a 2012 UK Department of Health initiative, demanding personalised treatment decisions on “physiological” rather than “chronological” age for cancer management (131,132). Patients offered surgery-alone within my curative cohort demonstrate a higher median age versus those offered neoadjuvant therapy (NAT) prior to surgery. With tree-based modelling I can even narrow down the exact age at which probabilities for curative pathways begin to shift in earnest - approximately 77 years of age. There is a well-recognised risk of deconditioning frail patients after neoadjuvant therapy and potentially rendering them unfit for surgery (14). A single attempt may be their only chance at cure which neoadjuvant therapy may accidentally compromise. NACRT prediction was comparatively unperturbed by age which may reflect a previously held view that pre-operative chemoradiotherapy (CROSS-style) is more tolerable compared to modern chemotherapy regimens (i.e. FLOT, Fluorouracil, Folinic Acid, Oxaloplatin, Docetaxel) (10,14,217–220). This has been somewhat contradicted in the recent German ESOPEC trial which reported an identical proportion of patients successfully reaching surgery in both NACT and NACRT arms (11). While median age in both NACT+S and NACRT+S groups were similar, a larger proportion of NACRT+S patients recorded better

performance status scores versus the NACT+S group. Within an already fitter cohort, chronological age may then have proved less consequential when determining suitability for multimodal neoadjuvant therapy.

PD analysis also allowed me to visualise scenarios vulnerable to variability in practice. This is beneficial in examining for health inequality and treatment bias. For surgery-only strategies the greatest variability in decision-making (depicted by a broad range of partial dependencies) was observed in those with cT2N0-1 disease. This is a patient group historically problematic – using potentially toxic neoadjuvant therapy to pre-empt disease control (here we presume nodal disease to be present and simply not yet detectable) at the risk of deconditioning patients out of surgical fitness for potentially little additional survival advantage (215,216). While these decisions are less troublesome in younger patients, for those over 75 my findings show that my local clinicians become more risk averse.

The inclusion of time-period in the form of an epoch as a test variable allowed me to account for shifts in practice over time. While Southampton offers both NACT and NACRT, the trends illustrated by Figure 4.1 over time indicates shifts in practice in response to the dissemination of key trial data. However, while trial epochs were incorporated within a separate classifier for explainability intent, they were not incorporated within my treatment prediction models (as they would not be a particularly useful feature when predicting forwards in time). Nevertheless, their inclusion within the model for partial dependence analysis allows for a retrospective accounting of changes in historic practice and in the context of a tool for self-auditing the MDT, is a relevant and useful variable. (10,223).

Performance status is used commonly by oncologists and remains widely utilised (albeit highly subjectively) within UK MDTs despite the existence of over 20 different frailty measures (264,265). It is recognised increasingly that performance status may be too unidimensional for personalised decision-making when assessing for treatment eligibility, yet it continues to be used prolifically, primarily due to ease of use, simplicity and standardization in clinical trial data (266). Nevertheless, through PD analysis I was able to group the Southampton cohort into two dominant clusters independently of age: those with PS 0-1 and those with PS 2-4. Prognostic significance maybe attached to pre-treatment patient physical activity, and example of which as described in 4.6.4 are MET scores of 4 or less (228). Patients whose baseline exercise tolerance match such scores equate roughly to PS scores of ≥2. Here again we can see the ability of XAI techniques to isolate trends within historic behaviours and link them to prognostic rationales. This can be used to audit and validate clinical practices within oesophageal cancer

MDTs. While national guidance on stratifying PS in curative oesophageal cancer cases is not immutable and in future, the field may leave PS measures behind in favour of more validated scoring systems, techniques like PD analysis allows for ML-driven benchmarking, a concept explored in other surgical specialties (229).

6.4 Healthcare professionals' perception of AI CDSSs and barriers to adoption

The integration of stake-holder input is a vital aspect of any healthcare innovation to ensure that the final product matches the needs of its intended users. AI, particularly, is a divisive technology within this sector and it was necessary to gauge the sentiments of clinicians who routinely attended MDTs. This allowed me to sense-check my models against the human paradigm, assess for discordance or disagreements therein, highlight potential areas of unconscious bias, and finally, pre-empt the barriers which might prevent more widespread adoption of these technologies.

The national survey (Appendix F) highlighted that firstly, most respondents were positively minded towards the concept of AI CDSSs indicating that there was an appetite for AI-assistance in MDTs. Secondly, there were clear, rational, concerns about trusting them. Respondents requested that novel molecular markers be considered within AI CDSS inputs as well as more granular information regarding the primary tumour (such as size, length and location). Symptom-related factors such as dysphagia, vomiting and quality of life were felt to be beneficial to model training, as well as nutrition-related features such as weight(loss), BMI and nutritional markers.

They also raised specific concerns about their willingness to use AI for decision making such as how to balance their innate assumptions of human superiority with embracing digital automation. How would innovators be able to address the need to maintain patient individuality and thus the influence of patient wishes on final treatment plans? Where would transparency and safeguarding be integrated into the AI pipeline and how? Could we prove that the system would actually improve current operations and improve upon the human paradigm? Finally, how would model inputs and model training also allow for changes within the field and advances in management from clinical trials.

The survey results have consequently provided a valuable, transferrable benchmark of both user-needs and regulator-requirement. These will prove a useful guide for the wider scientific

and commercial community working on AI innovation across sectors. The results directly led to several modifications and design elements for the subsequent MDT tool described in Chapter 5 such as the inclusion of local explanations, the inclusion of model performance and metrics for user-information, the inclusion of severe dysphagia within the palliative models as a key feature in providing oesophageal stents, as well as stimulating future planning of next-generation model inputs and plans for model life cycles to allow updates in training.

6.5 External validation of the MDT models

The external validation of my MDT models has confirmed the validity of this research proposal I put forward at the beginning of this thesis. I have shown through both internal and external validation processes that machine learning can leverage simple tabular clinicopathological data that is routinely collected by MDTs to successful learn and map MDT decision-making processes.

A key reason for the success of this process is in the core clinical guidelines that all UK units will inevitably follow as part of gold-standard clinical management for oesophageal cancer. This provides for a relatively consistent "core" stability of the MDT models both at a single centre and across centres. However, where ML shows its true benefit is in enhancing performance by being able to incorporate and integrate the nuance associated with an individual centre beyond the skeleton framework provided by clinical guidelines.

The external validation cohort demonstrated significant differences in their demographic make-up (the Oxford patients being younger, fitter and slightly earlier staged disease on average) yet the models remained stable in their predictive performance in spite of this confirming that the models could be applied, if so chosen, to their unit even when trained on Southampton patients. As the Oxford-trained models demonstrated, this generalisability was not solely down to chance as the process was successful in both directions.

With nearly 2/3rd of new referrals to the MDT ending in palliative pathways, I have also shown that this process is robust in rapidly predicting, and prognosticating for those not eligible for curative pathways, which is of clear clinical and financial benefit (34). When developed and implemented correctly these models may also have the capability to improve patient safety (45). While AI-based Clinical Decision Support Systems (CDSSs) may offer much to the healthcare sector, there remains a translational gap to successful implementation (254). There are undoubtedly numerous reasons for this, but it is partly attributable to a sense of clinician

superiority especially when handling nuance and uncertainty, or a lack of clinical validation (254,255,267). While there is clearly always a need for the human to remain in the loop even with my MDT models implemented, the process may also be leveraged into re-prioritising cases and ranked by order of computer-certainty (i.e. the machine knowing when it doesn't know). This may be a more palatable middle ground for the human agents within MDTs in the short term as the training datasets amass over time in both volume and granularity.

While the predictive performance of my models for early cancers eligible for endoscopic resection could not be validated in this study, this group accounts for a relatively small proportion of new oesophageal cancer cases, and many will have arrived at MDT discussion through routes that subvert the usual referral pathways. This makes for a challenging validation process as cases need to be standardised for fair comparison and modelling, however I believe the consistent performance seen within the other outcome classes on internal and external validation is likely to also be preserved within the endoscopic resection cohort too. The recently concluded CONGRESS study which looked to evaluate the best treatment strategies for early oesophagogastric cancers has recently amassed over 1600 cases and may prove a potential means of upscaling the data in this relatively under-represented cohort (268).

6.6 Machine learning versus traditional statistical approaches in this research

As discussed in Chapter 1, there is often a question of what Machine Learning offers over and above traditional statistical modelling. For many clinicians, traditional modelling has been a cornerstone of medical academia, providing tools for new predictive models, nomograms, and insights into how variables inter-relate. In the face of a long-established discipline, ML is a relatively new contender for most clinicians in practice today, needing to prove its worth. However, in the domain of predictive performance, ML takes the lead. ML models do not care as much about how variables might relate, simply that its predictions are accurate and reliable time after time. In the context of tools looking to alleviate workload pressures, it is clear that a tool such as ML devised primarily to achieve this with efficiency may be more beneficial in day-to-day practical settings, while traditional statistics may provide more benefit within purely research settings. However, to bridge the gap between the two disciplines, logistic regression models were utilised to significant effect within this work. Logistic regression models reside in both worlds, representing a classic and well-known technique within traditional statistics, capable of providing direct insight into how models all relate to each other and the final

outcome, as well as being able to generate new predictions. Furthermore, the ability to tune and feature select which is available through regularisation further allows prediction optimization over a static model. It also provides a direct comparison of a single type of traditional statistics with more commonplace ML techniques. Further still, MLR models acquitted themselves with significant success, however their utility is slightly hampered by the need for reliably linear relationships within their predictors and outcomes. This is something uncommonly seen in medical data and as such their potential benefits must be tempered with caution. The tree-based techniques used in this work allow for more non-linear relationships to be captured and provide more versatility in that regard.

6.7 Decision-making through experience versus cognitive bias

6.7.1 The Dual Process Theory

The fundamental drive in the performance of my ML models remains with the clinical decision-making at the human level. As such it is helpful to understand the underpinnings of this process especially when evaluating for sources of bias.

A popular model for human decision-making comprises two parallel pathways for information processing: known as the Dual Process Theory (DPT). Type 1 processes are rapid, intuitive and unconscious. Type 2 processes are by comparison slower, deliberate, more thought out and demand larger cognitive resources (269). In healthcare settings, as within many other domains defined by largely experiential learning, technical decision-making at more experienced levels is dominated by Type 1 processes, which stem from pattern recognition (190).

A relevant example of both processes would be to think “this 80-year-old male is unlikely to tolerate aggressive neoadjuvant chemotherapy given his age in my experience” (Type 1 processing) versus “This 80-year-old male, on paper seems a concern for tolerating neoadjuvant chemotherapy, when aggregating his comorbidity, poor exercise tolerance and intermediate Cardio-pulmonary exercise test, I wonder if he would tolerate it” (Type 2). Repeated processing in Type 2 will eventually allow skill gains and a shift into Type 1. It also provides for an “executive override” of Type 1 heuristics (“Normally we avoid neoadjuvant chemotherapy in someone of his age, but his prior conditioning is better than the average 80-year-old, on delving deeper he exercises regularly and is very active, I wonder therefore if…”) (269).

Type 1 processes have further been broken down into those hard-wired (biological driven), emotionally learned (hardwired or socially acquired), over-learned (embedded through frequent exposure and over-learning such as the "frequent flyer" attendee in the emergency department) and implicitly learned (learning without explicit intent or conscious awareness, such as observing and adopting biased behavioural traits from ones environment) (269,270). The latter two scenarios are particularly relevant to clinician decision-making where future decisions may be based off repeated negative experiences (such as a series of back-to-back cases which have failed to reach surgery after neoadjuvant chemotherapy within the elderly), or through subconsciously learning to equate age with poor outcomes from watching more senior colleagues make such judgements repeatedly in MDTs.

6.7.2 Cognitive bias

While bias can of course be introduced into both Type 1 and 2 processes it is fundamentally more likely in the former as the rapid, instinctive, unconscious nature of a Type 1 pathway is also uncorrected (190,271). Systematic error may be propagated in the absence of a negative feedback pathway. As much of our time is spent functioning in Type 1 processing especially within high-pressure, time-constrained clinical environments, cognitive bias can infiltrate MDTs in myriad ways from diagnostic momentum (reinforcing an already firmly held belief within the group), anchoring bias (focusing on early information within a case history such as advanced age, even when pertinent information later is available e.g. age identified first before hearing of a reasonable performance status), through to confirmation bias (accepting only information which supports a pre-conceived conclusion such as a borderline cardiopulmonary exercise test (CPET) result) (190).

My work in Chapter 4 investigated the importance of age in the accessibility and availability of specific treatment for OC patients. As discussed in that chapter we see a significant bias away from offering patients aggressive curative combination therapies once they hit their late seventies. We know from the literature that peri-operative risk increases in the elderly population and thus informs both our Type 1 and Type 2 processes via experiential learning both from the literature and from direct experience (225,226). It follows then that over time an implicit bias may form leading clinicians away from automatically considering aggressive therapies and instead towards automatically avoiding them unless provided a good reason not to (executable override).

How then to interpret the findings of my research when considering how to differentiate problematic “bias” from sound clinical “experience”? Which biases are “good”, and which are “bad”? While bias as a concept is traditionally sought out for eradication, especially in computational settings such as ML, not all bias is necessarily problematic. Pot and colleagues have previously argued that within ML, some biases may in fact be beneficial provided they do not propagate inequity (inequality through unjust cause) (227). While they spoke more directly towards the bias within training data of machine learning models and techniques which seek to diversify training data, I argue that the sentiment is also appropriate for human bias at the cognitive level. Appreciating that elderly patients are commonly co-morbid, decondition with major surgery and may suffer higher rates of complications is well established and thus commonly taught from and to clinicians. At a surface glance then, the bias towards the elderly may in fact be a willingness to spare them potentially a significantly worse quality of life. We may regard this as inequality of treatment access but not necessarily inequity as the cause is arguably just.

Nevertheless, this does not undermine the need to highlight and make such biases explicit. This provides fundamental insight into the human thought process followed by the machine’s logic. Explainability is paramount to satisfying trust from patients and regulators alike but also provides MDTs a route back towards self-reflection even if it simply to reaffirm their rationales.

6.8 Value and limitations of the MDT models

The value in these models may be considered from several different scenarios. The first is the benefit towards satellite units which feed into a central specialist MDT as is typically the case within the UK health service. The Southampton MDT for example is fed by case referrals across Hampshire and the Channel Islands including Winchester, Basingstoke, Isle of Wight, Jersey, Guernsey, Falklands with additional cases from Salisbury (which are partly filtered through Bournemouth) and a few bespoke referrals further afield. It is unlikely that a significant reduction in case-referrals would occur immediately based off such an MDT tool. The current medicolegal climate incentivizes the offloading of oncological decision-making responsibility from any single individual and towards the MDT. It is more conceivable that referring clinicians from those units may use the tool to pre-counsel patients on the most probable treatment options based off the MDT model ahead of formal discussions as a way to manage patient expectations and aid counselling.

Secondly, an interactive real-time dashboard offers an opportunity to compare the human decision making against the AI within the MDT meeting itself. The ability to provide real-time explanations for these predictions also provides an opportunity to discuss whether those reasons are still salient (or not).

Finally, a huge benefit from the computational power driving ML models remains their ability to not only learn from large data volume, but to also generate predictions at-scale, almost instantaneously. This means the capability to upscale workflow significantly to meet increasing caseload demands if this is required, or alternatively to re-prioritize and offload more clear-cut decisions to a separate "pre-MDT" filtering process and shorten the list (149).

With this in mind, I am mindful that my MDT models are not without their limitations. The average balanced accuracy from my models currently sits between 70-80% depending on the model indicating that 1 in 5 predictions will need reviewing. As time moves on and data volume increases the expectation would be to close this gap further, especially with additional and more targeted features such as molecular and radiological inputs. Nevertheless, the underlying nature of ML models remains unchanged – models seek to closely re-produce real-world scenarios and paradigms. Closely, but almost never perfectly. This is aptly demonstrated in the high but not perfect ROC curves, balanced accuracies, calibration curves and concordance rates of the models presented in this thesis. Some decisions are likely to change following contact with patients in clinic, re-discussion with additional MDT colleagues and sometimes simply because a different clinician or surgeon happened to be attending a given meeting. Consequently, such ML models will need to be applied in conjunction with the MDT rather than seek to replace elements of its operation.

6.9 How these models may influence OC management and learning at the MDT level

I envisage the models being used at present for decision-support as opposed to decision automation. While ML is undoubtedly extremely capable, it is not yet at a stage of advancement to completely replace the high-level nuance that human decision-making executes in an almost unconscious fashion (c.f. section 6.6). Consequently, the models and their application interface will need to be used in the short term to develop trust with the MDT while team members acclimatize to its presence, using them early on to sense-check MDT plans. Over

time this provides confidence in the consistency and safety of the tool to allow either an increase in cases per week or offloading simpler cases into a pre-MDT triage.

There will be a natural concern for the risk of automation bias (the gradual dependence and over-reliance on automated system recommendations even in the face of contradictory human expertise) (190). If MDTs find the AI models do indeed mirror their own practices, they may develop an increased reliance on the system as a whole. While this is not inherently an automatically negative consequence (and in fact long term is the hope of any MAI tool that it can offload some of the workflow burden), caution must remain with MDT personnel as no AI tool is infallible and the models are trained on historic human decisions (some of which may be suboptimal in hindsight) and there will always be a risk to patients where an AI is contributing to healthcare (56,57).

As described in section 5.7.3, the MDT is also traditionally a source of learning for junior and senior clinicians alike, benefiting from exposure to diverse domain expertise. The partial or even full automation of MDTs in the future creates risk of losing that source of decision-making or indeed normalising to a mean rather than the optimal. For these reasons it will be necessary to preserve and reassure clinicians that while ML models offer workflow benefits, the optimal model will be a human-AI collaboration with iteratively updates to the model based on updated practices in the human domain (following clinical trials and new research) with use of the AI to target either a sense check or to streamline the discussion lists.

Finally, there will be understandable concern over the ethical implications of embedding both decisions which were made historically but which later may prove to have been sub-optimal as well as propagating historical biases (such as age-influenced decision-making). Both of these are valid concerns. Taking the first scenario – what makes a good decision? As described in section 3.6, the first major hurdle is the lack of a well-used and standard single outcome measure that is quantifiable and routinely used universally. While “textbook” outcomes do exist for oesophageal cancer surgery, they ae currently multidimensional without a single final output metric (272). And while they are recommended for quality assurance assessment in oesophageal cancer surgery, this naturally excludes non-operative cases. Furthermore, where such outcomes are used, they are often compressed into a binary yes/no outcome (this is not in itself a downside for ML, however it requires that all sub-domains of the outcome are completely met, rather than accounting for partial and graded success) (273). What remains needed is a metric which remains a single measure but can account for variable success within multiple domains such as complication rate, survival, and quality of life, applicable for all

treatment modalities and routinely used in MDT across units. Until that is possible, I am restricted to predicting what currently occurs with little recourse for shifting towards predicting "success" of a given decision. The additional concern over propagating historical biases remains a valid issue within my models for the same reason – they are based on historical practice and decision-making. While the broader study period aims to average out some of the variability, this is none-the-less a challenge difficult to circumvent in this proof-of-concept research. Overtime, as recognition of such biases drives changes in practice, the inherent life-span limitations of deterministic models (those that remain static in predicting the same output for the same input criteria) should ensure that retraining on newer more updated data will also gradually remove some of these biases along with the continual and ongoing integration of RRI practices which are designed to ensure that such biases are acknowledged where they exist, and mitigated where they can be.

6.10 Pathway to deployment

The outcome of this research has led to a prototype decision support system in a useable form. However, the landscape of medical device regulation has evolved significantly, especially with the advent of medical AI systems and the EU AI Act in 2025. The current version of this tool could be released into the world as a research tool without need for additional regulatory input however for medico-legally sound deployment it would need to address the regulatory requirements of a medical device.

Regulatory pathway steps include the following: defining a device's intended use, assessing its risk classification (within the EU AI act, this tool's risk would be considered "high risk"), the implementation of a quality management system (QMS), generation of clinical evidence, undergo a conformity assessment (engage with an approved regulatory body to check technical documentation etc), secure regulatory approval and finally ensure adequate post-market surveillance. Risk classification is dictated by intended use; be it diagnosis, treatment, monitoring etc. Software as medical device (SaMD) classification ranges from class I (low risk, patient records, scheduling etc), class IIa (low to medium risk, providing information for downstream decision making), class IIb (medium to high risk, informing clinical decisions with risk of harm) and class III (highest risk, informing decisions with risk of irreversible harm or death) (274). Based on current classification, this tool would sit at least at class IIb. The generation of clinical evidence will typically involve model evidence (Internal validation), generalizability (external validation), performance metrics, clinical utility studies (testing how

the tool may enhance workflows in real life, whether it provides a quantifiable benefit to day to day operations), prospective studies (including trials in a "shadow mode" within a real clinical workflow such as an NHS MDT where predictions are run in parallel with the MDT but not used to influence outcomes at this stage) and where relevant, the use of randomized clinical trials. This thesis effectively addresses the first three elements of this already. Quality Management Systems are frameworks designed to assess, monitor, document and provide processes for auditing medical AI, including quality assurance, risk and risk mitigation strategies as well long-term life cycle monitoring (275). It is an important aspect of quality control as well as a means of addressing downstream issues and risk or bias development in medical AI tools. Finally, post-market surveillance ensure that the tool remains fit for purposes as time moves on, accounting for shifts in the target demographic, identifying data drift, and also forms part of the Good Machine Learning Principles for Medical Device Development guide produced by the US FDA, Health Canada and the UK MHRA (276). Once these steps are achieved, the tool can then be introduced into a real NHS workflow to test if it can provide tangible benefits to MDT operations such as caseload screening or prioritization, or in the form of time and cost savings before being potentially adopted full-time into the NHS digital ecosystem.

6.11 Study limitations and future directions

The initial pilot study in curative patients was limited by the study sample size. Despite a cohort spanning approximately 10 years within a tertiary referral centre which serves a catchment of over 3 million people along the south coast and Channel Islands, the unit sees approximately 40-60 curative cases a year of which some will represent rarer histological subtypes not included within the current models (277). By utilising supervised-learning techniques which tolerate smaller datasets in conjunction with cross-validation techniques I have sought mitigate some of the overfitting associated with smaller training sets and provide a reasonable estimate of the likely generalisability error associated with my models. A similar issue affected the palliative cohort who ordinarily account for the majority of MDT cases. However, once cases were filtered to ensure appropriate histology and eligible treatment this again left a cohort of just over 430 cases. Datasets of this size are common to healthcare scenarios where access, adequate recording and of course patient consent may limit how much data can be curated. For a cohort who rarely survive beyond 2 years, prognostication beyond this for palliative cohorts is extremely challenging, especially for those few patients who prove outliers in survival. This again circles back to the challenges in modelling limited training cohorts although the survival within our local cohort matches those seen nationally and consequently a larger

dataset may not change the distribution of survival probabilities significantly. Within machine learning, no exact guidance is agreed upon within domain experts on sample size, though it is not uncommon to curate datasets comprising 1000s of observations, especially for deep learning platforms and tasks requiring complex data such as images and audio. A typical rule of thumb recommended by multiple domain experts for tabular data is to ensure that the number of observations is an order of magnitude greater than the number of features (278).

A key choice in this research from the outset was the choice to focus on tabular data structure with clinicopathological data. As described in section 1.1.7 there have already been numerous studies reporting positive outcomes from applying ML to imaging and histopathological data types. However, in the face of striving for a novel CDSS for MDT decisions in oesophageal cancer it was important to start more simply and determine proof-of-concept. Tabular MDT data is readily available, easy to read for statistical computing software and rapidly analysed even in real time. Additionally, as described previously, within more restricted datasets, tabular data when analysed by more classical ML algorithms can outperform more complex deep learning models (252). Consequently, while it is possible that future iterations may benefit from integrating multimodal data types at the time of prediction (and indeed will yet more closely represent the true process of the modern MDT), simpler more structured data types are necessary for grass-roots validation of concept and lend themselves far more to explainability.

The typical criticism of single-centre training data does raise a philosophical point unique to this research – how to balance generalisability (and thus “validity” in clinical circles) with modelling a nuanced and individualised team-based process (and thus “personalised”). Multi-centre datasets offer larger training sets which shore up statistical power but simultaneously introduce more variability within the dataset. MDTs often reach a natural equilibrium in composition over time and team-members become acquainted with their own team dynamics leading to regional idiosyncrasies. How then to balance the need for statistical power with the risk of reducing overall accuracy and reliability of predictions locally when applied to a single MDT? Patients are, after all managed by one MDT and almost never discussed nationally (although more recent “national MDTs” have started to make an appearance within oesophageal cancer, generally from an academic or interesting-case perspective). A “one-size-fits-all" model that generalises nationally would be computationally cheaper and less labour intensive to implement but fails to capture the nuanced geographical biases which inform patient management locally at a particular MDT. Instead, we risk normalising to a central “mean” nationally where models would continue to perform “well” but there would remain a

perpetual gap in performance not accounted for. A counter argument to this would of course be that standardization of care is a key outcome from this research over time, why then is this a problem? Simply put the first step in championing the benefits of AI in this process would be to standardize decisions at the local level. Standardizing practice nationally will only come once a standard of care is established both in practice and in intended outcome; the latter currently remains absent. Once our oncological outcomes are agreed upon, we can then work towards standardization towards the best practice nationally.

The predictor variables I selected to test were aimed to achieve a balance between sufficient features to achieve a diversity of clinical inputs clinicians may consider during deliberation while simultaneously being realistically available to all cases for the MDT to access at the time of determining a treatment plan. As described in section E.3, cardiopulmonary exercise testing could not be used in this work as it was only available and applicable in practice to those offered surgery and is not yet widely used for non-curative patients even as a means of judging treatment tolerance. Nevertheless, the Eastern Cooperative Oncology Group (ECOG) performance status has offered a relatively accessible alternative which my models have found of material benefit nonetheless.

The model features were curated using a priori knowledge and national guidelines. While the latter are not subject to very frequent changes nationally, several novel molecular markers and immunotherapy-based targets have come to the fore during the period of this research. The lead-time on their clinical adoption means these variables cannot be reflected within the present models. This highlights another difficulty in curating training datasets for ML modelling. As clinical trials are released, these tend to produce shifts in practice which occur gradually but must then be provided a period of bedding in to acquire sufficient training data, lending itself to the criticism that ML-based CDSSs may prove slow to adapt. Future iterations may also be able to integrate variables such as lifestyle risk factors, BMI, shifts in oncological practice (e.g., neoadjuvant chemotherapy 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. The notable exclusion of immunotherapies and other novel systemic modalities will naturally limit model performance for new cases in the immunotherapy era. Future modelling will need to incorporate an expanding array of systemic

treatments such as Chemotherapy +/- anti HER2, anti-PD-1/PD-L1, Claudin 18.2, and MMR-d/MSI-H (163–166).

It is also worth noting in this regard that while survival is by far the most prolific outcome measure in oncological studies, it is not always the most salient from a patient's perspective. In the palliative setting, quality of life (QOL) is often key yet frequently overlooked. QOL data is not yet routinely collected within our MDT thus not modelled within the present study but will be essential in future models. At present these models also do not factor visual data (radiological and histopathological imaging), nor key social/ human factors (the last of which, previous studies have found inconsistent in MDT environments) (117,118). However, as this is the first time machine learning has been applied to the modelling of MDT decisions for oesophageal cancer patients, it was incumbent on me to start with a more simplistic data structure to produce proof-of-concept. Tabular clinicopathological data was therefore the first approach to pursue and once this could be achieved this will set a scientific rationale for upscaling to more complex datatypes such as medical imaging and histological samples in the future.

This model explainability study using XAI techniques benefitted from an uplifted dataset combining both curative and palliative cases over a slightly longer period (893 OC patients over a 13-year period). It remained a single-centre analysis over a study period when several shifts in oncological practice have undoubtedly occurred in both NACT regimens as well as the emerging option of immunotherapy. I focussed primarily on global explainability techniques within this study (variable importance and partial dependence) however, local techniques were not employed on this occasion. The development of this work within the oesophageal cancer space will allow the development of ML-derived decision-support tools which can be used by MDT personnel. The need for explainable and ideally interpretable ML models and thus the approach presented here represents a route towards clinician trust by first offering global insight into team-level decision-making when mirrored by the machine. As useable tools then developed, local XAI techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHappley Additive exPlanations (SHAP) can then offer instance-level explanations too to focus the explanations to the individual patient (209,210).

The models within this research are deterministic (i.e. for a given set of inputs, they will always generate the same outputs every time). Consequently, the algorithms do not learn and adapt in real-time to new data when it is available, requiring instead, to be re-trained periodically with new data to ensure the predictions remain consistent with the medical field. The benefit here is in ensuring new data is not liable to introduce errors in training, either through corrupted input

data or reporting bias. However, this comes at the price of labour intensive (and onerous) data cleaning and quality monitoring. Nevertheless, deterministic models have for the reasons above been historically favoured by regulators such as the FDA though it has been recognised for some time that limiting the use of continually learning AI also withholds one of the unique benefits AI has to offer (279). It is anticipated that over time regulators will develop action plans to accommodate continually learning models which would minimise human workload provides a means of quality control monitoring was in place.

The use of a Responsible Research and Innovation program will also be essential to ensure that as far as possible the design of the AI CDSS has considered the stakeholders concerned with its intended application and acted to mitigate and assess risk to individuals along with society as a whole. Early collaboration with a diverse skill-mix and open-format workshops will lead to a CDSS with immediate clinical translational potential. RRI frameworks such as the AREA 4Ps framework (280) will help delineate potential challenges and barriers to the use of AI based CDSSs in clinical settings, including the need to mitigate against training data bias, data management and governance, the adaptability of models to future shifts in practice and more regulatory concerns such as the role of the human within the loop, the foundations of legal and clinical responsibility (54,58,281). With the advent of the newly released EU AI act in July 2024, this is of salience as healthcare related AI will inevitably fall into the high-risk category of AI applications especially when the impact of new technologies may still be emerging far downstream of their first deployment (The “Collingridge Dilemma”) (257,258,282).

6.12 Conclusions

I have demonstrated ML – based predictive models trained on pre-treatment clinicopathological variables can predict oesophageal cancer MDT treatment decisions with good accuracy including palliative patients who account for a majority of MDT caseload yet are historically overlooked in decision-support resources (39). I have demonstrated 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 (143–146). 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.

Appendix A Supplemental Materials for Chapter 2

A.1 Supplemental Tables

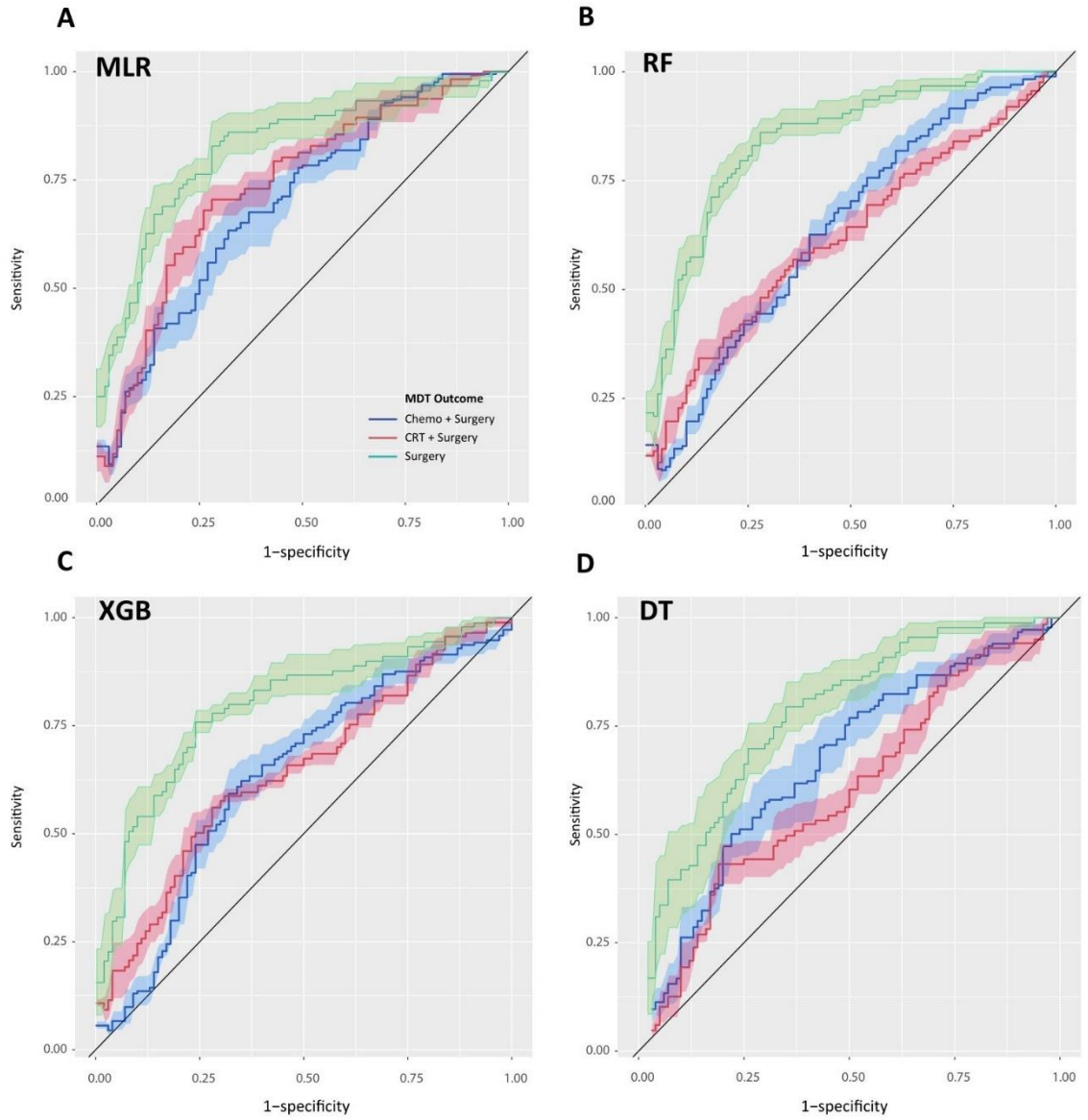
Supplemental Table 1 - Inter-algorithmic comparison of performance. P values for Kruskal-Wallis analysis of variance (ANOVA) of AUROC for outcome classification by algorithm (top left). Pairwise Wilcoxon Rank Sum Test provided for pairwise comparison (Asterisk denotes statistically significance * = P < 0.05, ** = P < 0.01, ***P < 0.001).

Chemo vs Others (P < 0.001)* **	**MLR**	**RF**	**XGB**	**DT**
MLR	No data	**0.002 ** **	**<0.001 *** **	**0.143**
RF	No data	No data	0.380	0.372
XGB	No data	No data	No data	0.152
DT	No data	No data	No data	No data
CRT vs Others (P < 0.001) * **	**MLR**	**RF**	**XGB**	**DT**
MLR	No data	**<0.001 *** **	**<0.001 *** **	**<0.001 *** **
RF	No data	No data	0.620	**<0.001 *** **
XGB	No data	No data	No data	**<0.001 *** **
DT	No data	No data	No data	No data
Surgery vs Others (P < 0.001) * **	**MLR**	**RF**	**XGB**	**DT**
MLR	No data	0.134	**0.001 ** **	**<0.001 *** **
RF	No data	No data	0.051	**<0.001 *** **
XGB	No data	No data	No data	**<0.001 *** **
DT	No data	No data	No data	No data

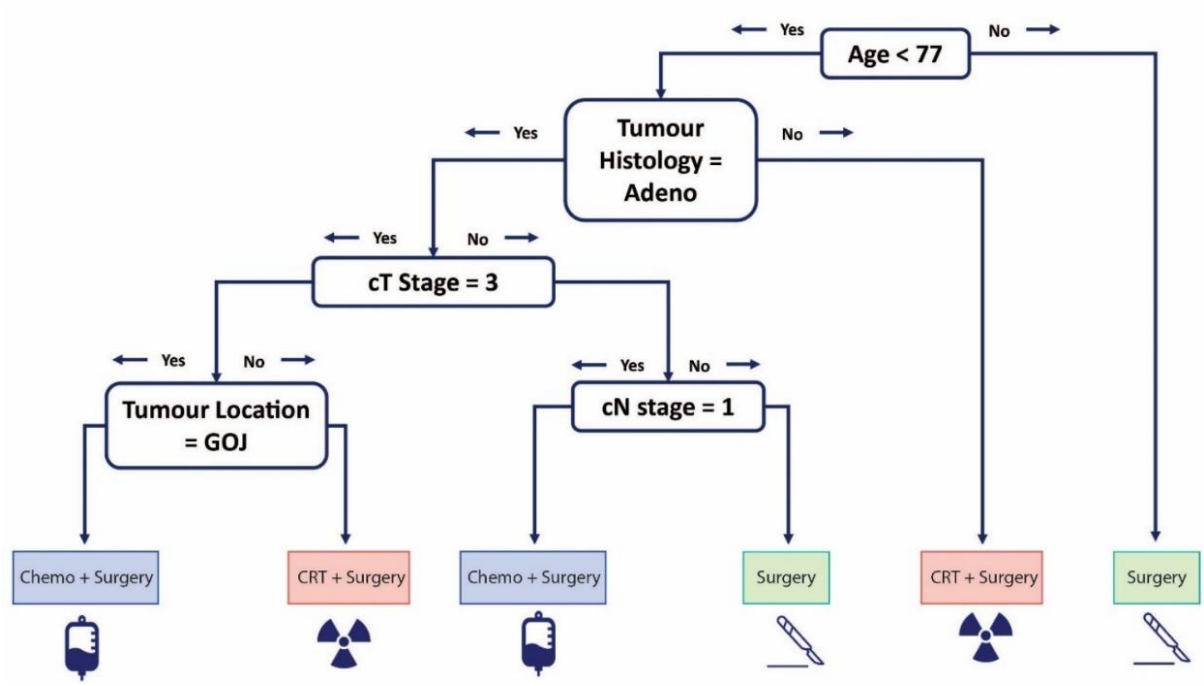
Supplemental Table 2 - Intra-algorithmic comparison of performance. - P values for Kruskal-Wallis Analysis of AUROC for intra-algorithm classification performance (top left). Pairwise Wilcoxon Rank Sum Test provided for pairwise comparison (Asterisk denotes statistically significant * = P of 0.05, ** = P < 0.05, * for P <0.001)**

MLR vs Others (P < 0.001) * **	**NACT+S**	**NACRT+S**	**Surgery**
Chemo	No data	**< 0.001*** **	**< 0.001*** **
CRT	No data	No data	**< 0.001*** **
Surgery	No data	No data	No data
RF vs Others (P < 0.001) * **	**NACT+S**	**NACRT+S**	**Surgery**
Chemo	No data	**< 0.001*** **	**< 0.001*** **
CRT	No data	No data	**< 0.001*** **
Surgery	No data	No data	No data
XGB vs Others (P < 0.001) * **	**NACT+S**	**NACRT+S**	**Surgery**
Chemo	No data	**< 0.001*** **	**< 0.001*** **
CRT	No data	No data	**< 0.001*** **
Surgery	No data	No data	No data
DT vs Others (P < 0.001) * **	**NACT+S**	**NACRT+S**	**Surgery**
Chemo	No data	**0.004** **	**0.001** **
CRT	No data	No data	**< 0.001*** **
Surgery	No data	No data	No data

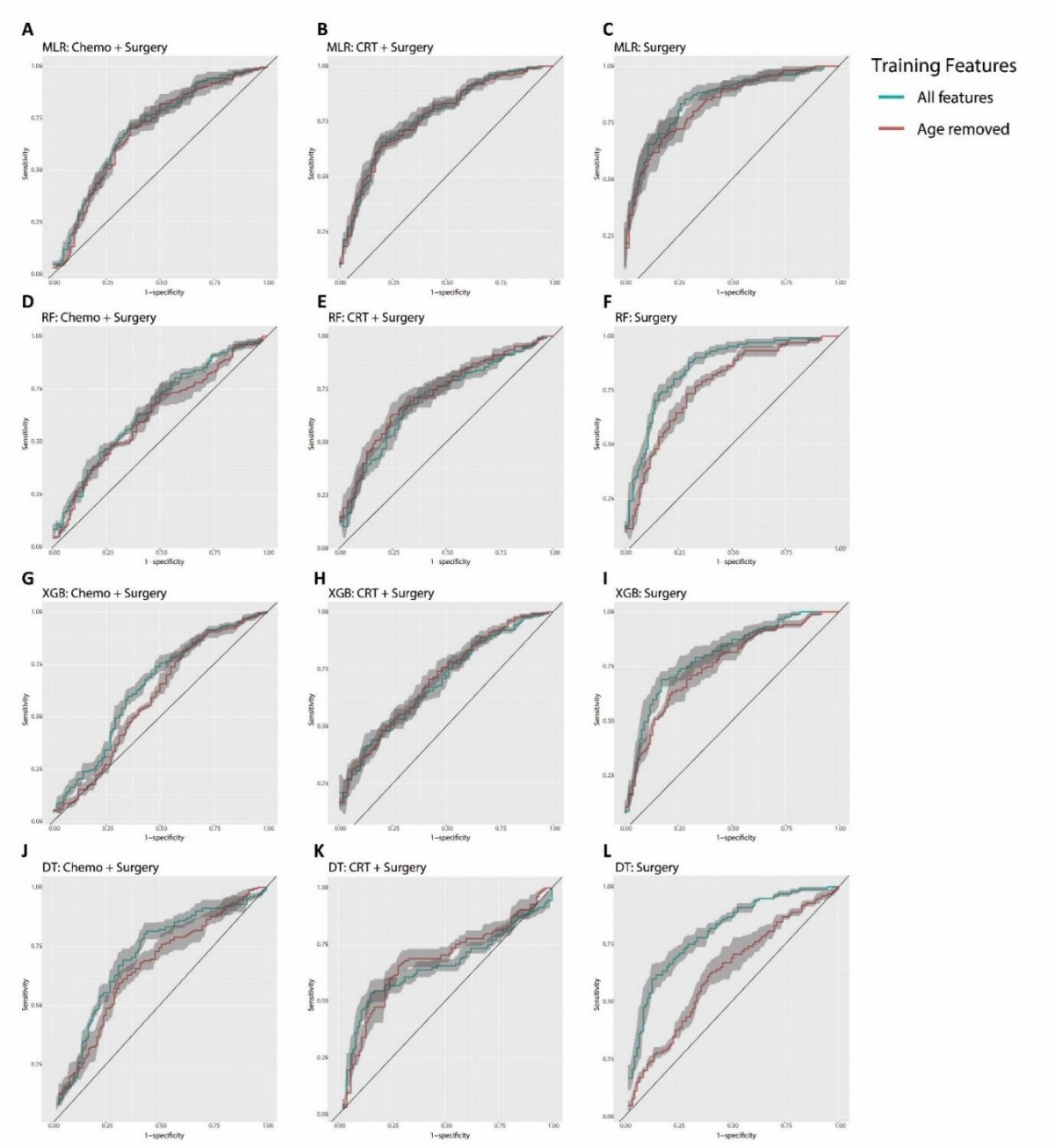
A.2 Supplemental Figures



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



Supplemental Figure 2 - Visualised Decision Tree analysis of OC MDT decision making framework (best trained model)

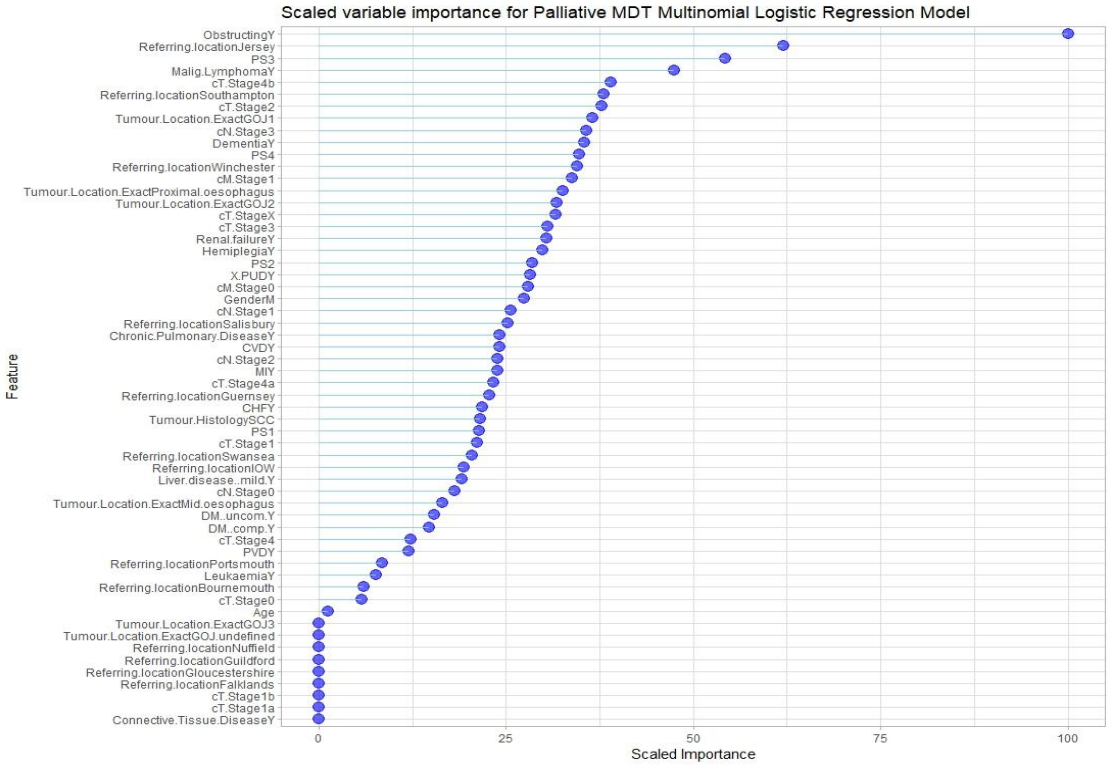


Supplemental Figure 3 - Comparative ROC analysis for all algorithms when age is included (green) and removed (red) from models trained to predict MDT treatment decisions (Statistically significant drop denoted by red asterisk). A = MLR C + Surgery B = MLR CRT + Surgery, C = MLR Surgery, D = RF C + Surgery, E = RF CRT + Surgery, F = RF Surgery, G = XGB C + Surgery, H = XGB CRT + Surgery, I = XGB Surgery, J = DT C + Surgery, K = DT CRT + Surgery, L = DT Surgery

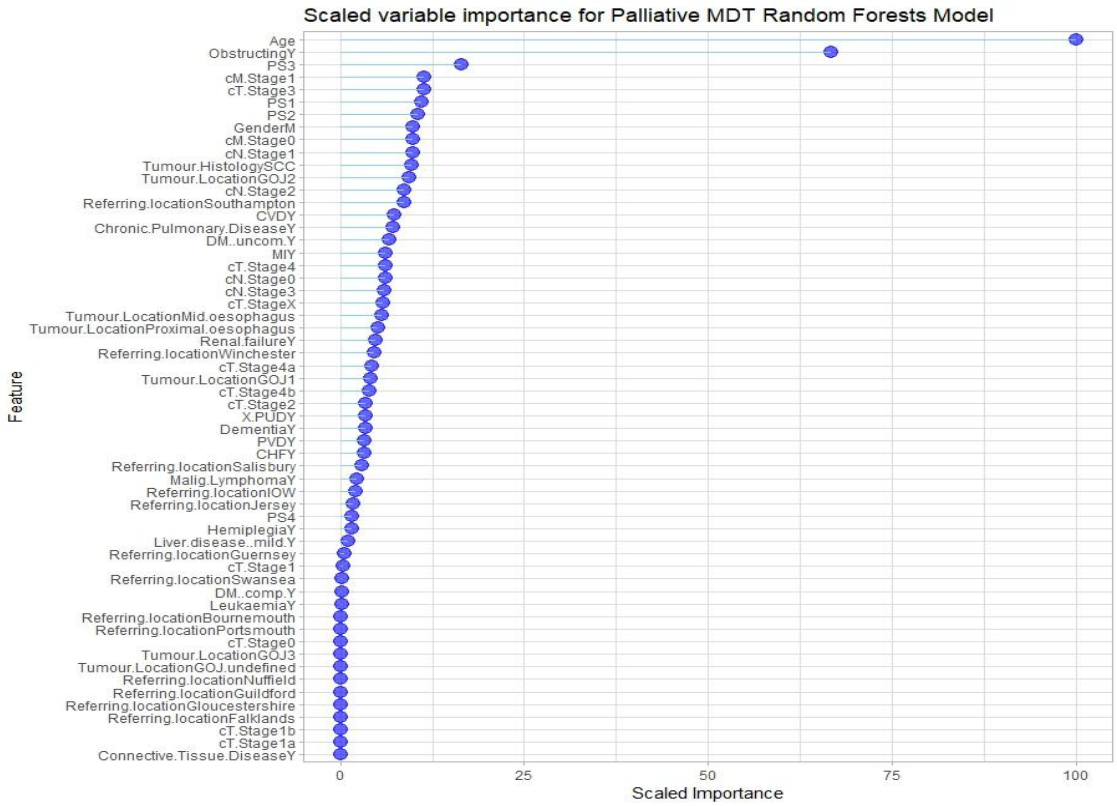
Appendix B Supplemental Materials for Chapter 3

B.1 Supplemental Figures

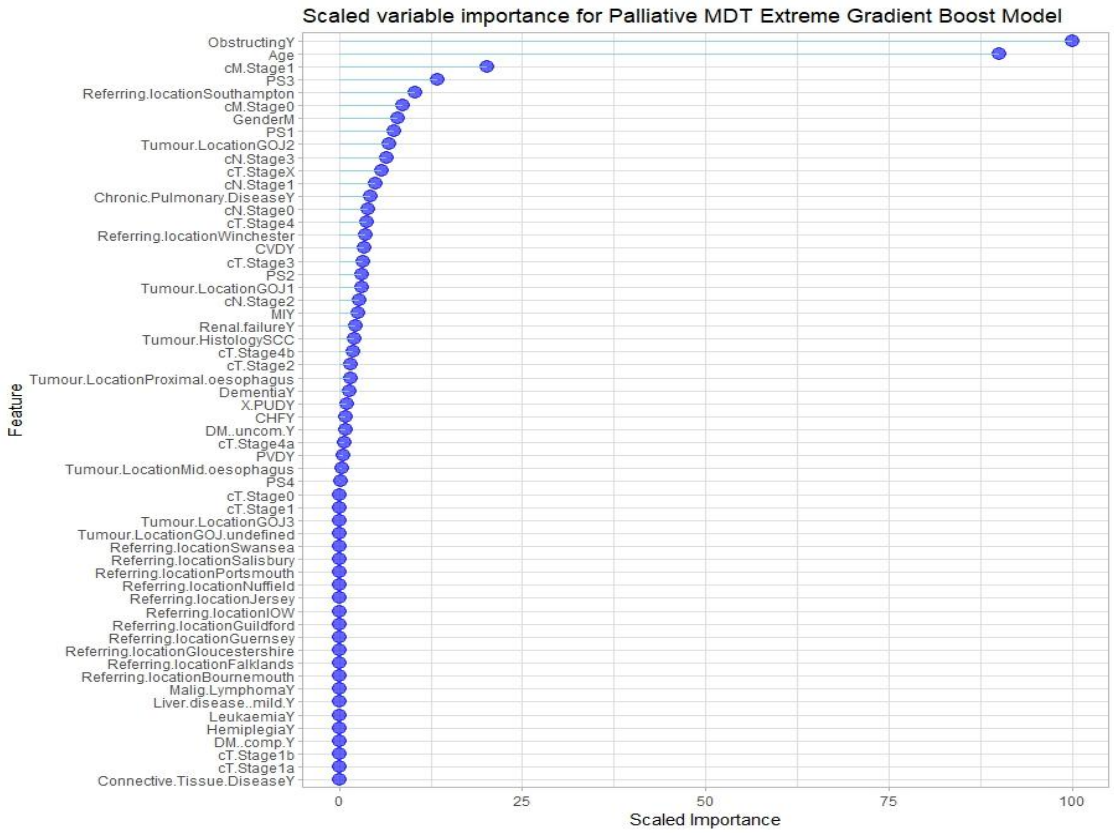
B.1.1 High-resolution figures



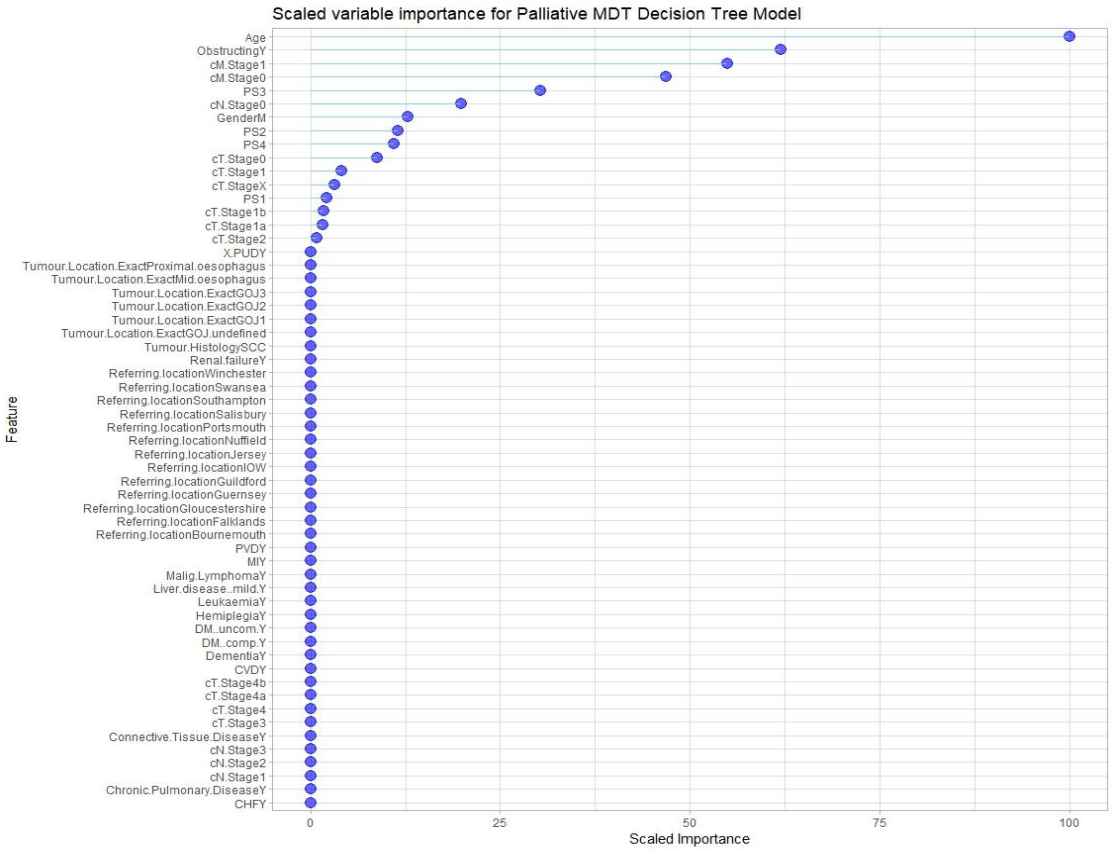
Supplemental Figure 4 - Variable importance plot for the Palliative MLR classifier model (Large version)



Supplemental Figure 5 - Variable importance plot for the Palliative RF classifier model (Large version)



Supplemental Figure 6 - Variable importance plot for the Palliative XGB classifier model (Large version)



Supplemental Figure 7 - Variable importance plot for the Palliative DT classifier model (Large version)

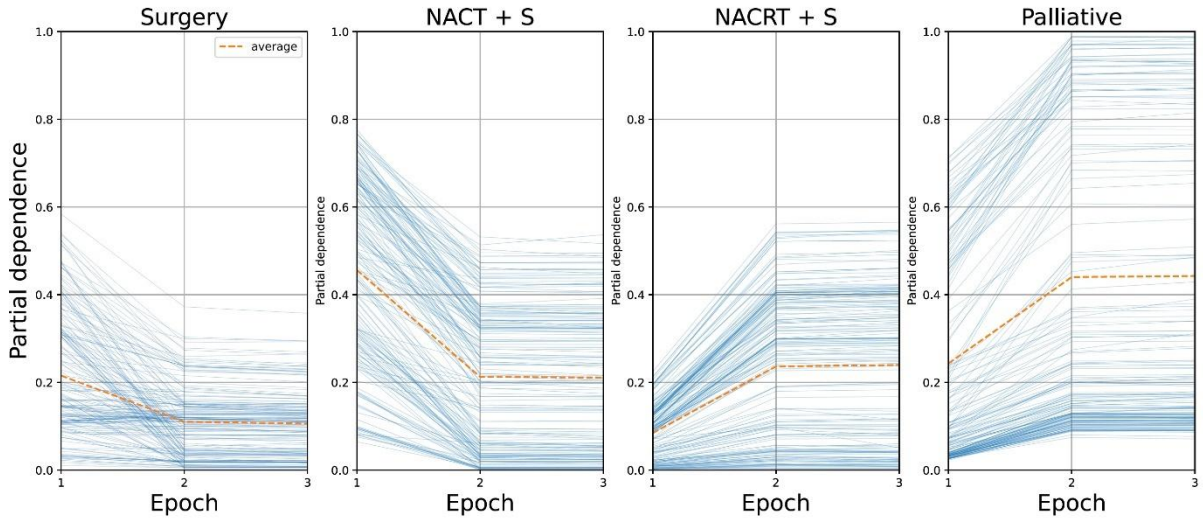
Appendix C Supplemental Materials for Chapter 4

C.1 Supplemental Tables

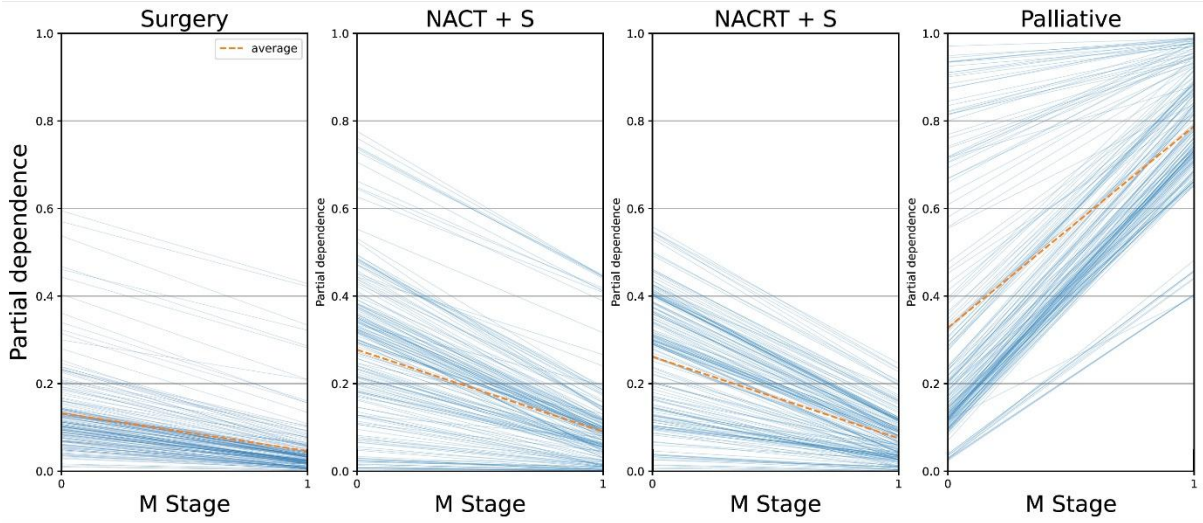
Supplemental Table 3 - Breakdown of cases by referral unit for the study cohort

Referral Unit	"Chemo" (N = 209) (%)	"CRT" (N = 196) (%)	"Surgery" (N = 102) (%)	"Palliative" (N=386) (%)	Total (N = 893) (%)
Basingstoke	14 (6.7%)	30 (15.4%)	6 (5.9%)	12 (3.1%)	62 (6.9%)
Bournemouth	2 (1.0%)	1 (0.5%)	1 (1%)	1 (0.3%)	5 (0.6%)
Falklands	0 (0%)	1 (0.5%)	0 (0%)	0 (0%)	1 (0.1%)
Gloucestershire	0 (0%)	1 (0.5%)	0 (0%)	0 (0%)	1 (0.1%)
Guernsey	1 (0.5%)	1 (0.5%)	0 (0%)	1 (0.3%)	3 (0.3%)
Guildford	2 (1.0%)	0 (0%)	0 (0%)	0 (0%)	2 (0.2%)
Isle of Wight	39 (18.7%)	28 (14.3%)	20 (19.6%)	15 (3.9%)	102 (11.4%)
Jersey	15 (7.2%)	7 (3.6%)	10 (9.8%)	6 (1.5%)	38 (4.3%)
Nuffield	0 (0%)	0 (0%)	1 (1.0%)	0 (0%)	1 (0.1%)
Portsmouth	0 (0%)	0 (0%)	1 (1.0%)	1 (0.3%)	2 (0.2%)
Salisbury	5 (2.4%	2 (1.0%)	0 (0%)	12 (3.1%)	19 (2.1%)
Southampton	80 (38.3%)	94 (48.0%)	40 (39.2%)	314 (81.3%)	528 (59.1%)
Swansea	0 (0%)	0	0 (0%)	1 (0.3%)	1 (0.1%)
Winchester	51 (24.4%)	31 (15.8%)	23 (22.5%)	23 (6.0%)	128 (14.3%)

C.2 Supplemental Figures



Supplemental Figure 8 - Individual conditional expectation plot of time epoch on predicted probability of treatment pathways (orange line = average line). Epoch 1 = time preceding release of CROSS trial (approx. 2012), epoch 2 relates to time between CROSS and release of FLOT 4 trial (2012 – 2017) and epoch 3 relates to time since FLOT4 (2017 onwards).



Supplemental Figure 9 - Individual conditional plot for influence of cM stage in isolation on predicted probability of treatment pathways (orange line = average line)

Appendix D Supplemental Materials for Chapter 5

D.1 Supplemental Methods: Responsible Co-Design

The RRI program used in this study looked to approach the development of the ML-based CDSS in a number of separate but related strands. We utilised a semi-structured multidisciplinary workshop framework to discuss a series of RRI prompts from the RRI card deck devised by Horizon Digital Economy Hub (University of Nottingham). To obtain end-user insights we also conducted semi-structured interviews with MDT clinicians to understand the clinical variables they looked at during MDTs as well as how they perceived the use of AI-driven tools both in clinical medicine as a field as well as OC more specifically. They were also shown iterations of the evolving CDSS tool. The feedback and insights generated through this process was used to make changes and updates to the CDSS.

D.1.1 RRI prompts

To ensure the CDSS development considered ethical, societal, and governance implications we utilised specially designed RRI card decks which prompts users to reflect, discuss and act on several aspects of RRI such as risk, bias, fairness, transparency (60,283). These also allow users to look ahead and anticipate potential consequences. This was used to inform the intended applications, target outcomes to predict and mechanisms through which to increase transparency and explainability of the CDSS at the point of use. The RRI card decks used within this program were designed by the University of Nottingham's Horizon Digital Economy Hub (Supplemental Figure 4)(60,283) . Each card is categorised under for one of the 4 principles of the AREA framework (Anticipate, Reflect, Engage, Act). A card will provide a specific prompt, as well as some example actions the research team can then use to action solutions. The cards cover a range of potential risks and issues including unintended consequences, sustainability, potential conflicts. Discussions were recorded and transcribed for later review and analysis.

D.1.2 Co-design workshops

We conducted regular workshops attended by diverse members of the research group including digital health experts, XAI computer scientists, clinician scientists, Heartburn Cancer UK

representatives, lay public and oesophageal cancer patients. These open-format workshops provided an arena for the research group to share thoughts on the evolution of the CDSS design while discussing validity of the models, clarity of the prediction explanations and the types of information patients may wish to know at the time of clinic review from their treating clinicians. RRI cards were again used here as stimulus prompts to consider new updates and evolutions of the CDSS between meetings and how these changes might mitigate or propagate risks associated with AI innovation. These workshops took place via teleconference over a series of months running in parallel with the UI development as well as validation of the ML models discussed within this study.

D.1.3 Clinician Interviews

The development of the machine learning (ML) model and its associated user interface (UI) for our CDSS was also guided by semi-structed interviews with clinicians which were analysed by a thematic analysis (interview questions are included at the end of this section) (284). Key themes were identified, relating to their understanding, perceptions, and concerns about the use of AI in healthcare, specifically in the MDT of oesophageal cancer. The interviews were transcribed using Microsoft Teams software, and an iterative coding process was applied. Initially, open coding was used to identify significant statements and concepts. These codes were then grouped into broader themes through axial coding, which allowed for the identification of patterns and relationships within the data. The final themes were refined through selective coding, ensuring that they accurately represented the clinicians' perspectives and insights.

The interviews were structured into two key stages with the clinicians, focusing first on their general perceptions of AI in healthcare and later, specific feedback after interacting with the prototype tool. The semi-structured interviews engaged with six clinicians, including oncologists, radiologists, specialist nurses and surgeons specializing in oesophageal cancer. All participants are UK-based clinicians, and they have varied levels of experience in MDT ranging from 5 to 20 years. These participants were selected for their expertise and their roles in the multidisciplinary team of oesophageal cancer. Their involvement was crucial in ensuring the tool was aligned with the needs and workflows of end-users.

Stage 1: Initial Interviews and Needs Assessment

The first stage involved semi-structured interviews with the clinicians to explore their general understanding of AI and their perceptions of its potential role in clinical practice. The interviews aimed to identify clinicians' expectations, concerns, and perceived barriers to adopting AI tools

in their workflow. Clinicians were asked about their familiarity with AI, their previous experiences (if any) with AI tools, and their views on the trustworthiness and reliability of AI in clinical decision-making. This stage provided critical insights into the clinicians' mental models regarding AI, which informed the initial design of the ML model and UI. Understanding the clinicians' baseline perceptions was essential for addressing any misconceptions and ensuring the tool was designed with their concerns in mind, particularly regarding transparency, explainability, and trust.

Stage 2: Prototype Demonstration and Feedback

In the second stage, after the initial design of the ML model and UI was completed, the clinicians were invited to participate in the second stage of interviews. During this stage, the clinicians were presented with a prototype of the tool, which included a preliminary version of the ML model integrated into a user interface. The clinicians were asked to interact with the tool and provide specific feedback on its functionality, usability, and clinical relevance. This feedback focused on the following areas:

- Usability: The ease of navigation within the UI, clarity of information presentation, and overall user experience.
- Functionality: The relevance and accuracy of the ML model's predictions, as well as the usefulness of the tool in supporting clinical decision-making.
- Integration: How well the tool could be integrated into existing clinical workflows.
- Trust and Transparency: The clarity of the model's explanations for its predictions and the degree to which clinicians felt they could trust and rely on the tool.

The insights gathered during these interviews were used to refine both the ML model and the UI. Specific changes were made to enhance the interpretability of the model's predictions, improve the clarity of the user interface, and ensure that the tool met the clinicians' practical needs in a clinical setting.

D.2 Exploratory Clinician Interview materials (Questions)

ERGO number: 70375

IRAS: 319540

Appendix D

Clinician Interview primary goal: To explore, clarify and consolidate the clinical variables expert clinicians intuitively utilise in determining clinical treatment pathways for the management of Oesophageal cancers (OC). We can subdivide these factors into clinical/ histopathological/ radiological and social variables.

Clinician interview secondary goal: Explore clinician perception, sentiment, and reservations for the inclusion of digital, Machine Learning (ML) and Artificial Intelligence (AI) - based clinical decision tools in health care settings, specifically the assessment and management of oesophageal cancer.

We aim to interview expert clinicians involved routinely in the management of oesophageal cancer patients. The inclusion criteria for our participants that they are current practicing consultant clinicians (Surgeons/ Gastroenterologists/ Oncologists) at the University Hospitals Southampton and experienced contributors to the Upper Gastrointestinal (UGI) Surgery Department multidisciplinary team (MDT) who are willing and agree to participate.

Section 1: Each interview will discuss the following questions:

Q1: What factors/clinical features do participants consider crucial to their decision-making? [Key decision-variables for downstream hybrid ML model]

Q1.1: What clinical features of the tumour complex would this include?

Q1.2: Are there any specific radiological factors that inform their decision-making process?

Q1.3: Do any specific histological features matter to the participant

Q1.4: Do participants consider any social or human factors routinely in choosing a treatment pathway?

Q1.4.1: If so, which? And why?

Q1.5: Of the factors discussed above, do participants attribute equal weighting to these or do they feel some factors are more important than others?

Q1.5.1: How would they rank such factors?

Q2: What do the participants understand by Machine Learning and Artificial Intelligence-based clinical decision tools? [pre-deployment perception]

Q3: What feelings/sentiments do participants possess towards these digital tools? [Current beliefs]

Q3.1 Do they currently use any automated tools outside UGI cancer clinics?

Q3.2What are their perception toward these tools?

Q4 What do they think AI could enable them to do in the clinic? [Perceived utility]

Q4.1 Do they currently feel they know enough to trust these tools in a healthcare setting?

Q5: What are the main barriers towards developing or possessing trust in such tools? [Barriers and expectations]

Q5.1 How advance do they think AI tools will be?

Q6: If such tools were available and scientifically validated what further safe-guards or measures would participants wish to have in place to feel willing to use such tools within their own practice? [Clinician confidence, motivation]

Q6.1 What motivates them to use AI tools in the clinic?

Q6.2 To what degree would clinicians be interested or willing to be part of the design and integration process?

D.3 Supplemental Tables

Supplemental Table 4 - Patient demographic by model feature and primary model outcome classes

Pre-treatment variables	UHS "Chemo" (N =210) (%)	OUH "Chemo" (N =373) (%)	UHS "CRT" (N =196) (%)	OUH "CRT" (N =86) (%)	UHS "Surgery" (N =108) (%)	OUH "Surgery" (N = 44) (%)	UHS Palliative (N =439) (%)	OUH Palliative (N =475) (%)	UHS Total (N = 953) (%)	Total (N = 978) (%)
Gender										
Male	179 (85·2%)	303 (81·2%)	137 (69·9 %)	52 (60·5 %)	82 (75·9%)	34 (77·3%)	320 (72·9%)	355 (74·7%)	718 (75·3 %)	744 (76·1 %)
Female	313 (14·8%)	70 (18·8%)	59 (30·1 %)	34 (39·5 %)	26 (24·1%)	10 (22·7%	119 (27·1%)	120 (25·3%)	235 (24·7 %)	234 (23·9 %)
Median Age in years (Range)	65·5 (21 – 81·2)	65·0 (31·0 – 80·0)	66·6 (40·0 – 81·0)	65·0 (29·0 – 77·0)	73·4 (33·7 – 83·0)	67·5 (37·0 – 83·0)	75·2 (29·8 – 96·7)	72·0 (33·0 – 96·0)	70·0 (21·0 – 96·7)	68 (29·0 – 96·0)
Performance status										
0	120 (57·1%)	350 (93·8%)	138 (70·4 %)	83 (96·5 %)	38 (35·2%)	39 (88·6%)	75 (17·1%)	240 (50·5%)	371 (38·9 %)	712 (72·8 %)
1	85 (40·5%)	22 (5·9%)	54 (27·6 %)	3 (3·5%)	53·7%)	4 (9·1%)	132 (30·1%)	121 (25·5%)	329 (34·5 %)	150 (15·3 %)
2	5 (2·4%)	1 (0·3%)	3 (1·5%)	0	12 (11·1%)	1 (2·3%)	140 (31·9%)	69 (14·5%)	160 (16·8 %)	71 (7·3%)
3	0	0	1 (0·5%)	0	0	0	87 (19·8%)	43 (9·1%)	88 (9·2%)	43 (4·4%)
4	0	0	0	0	0	0	5 (1·1%)	2 (0·4%)	5 (0·5%)	2 (0·2%)
cT stage										
0	1 (0·5%)	0	0	0	2 (1·9%)	0	1 (0·2%)	0	4 (0·4%)	0
Is	0	0	0	0	3 (2·8%)	0	0	0	3 (0·3%)	0
1	0	1 (0·3%)	0	0	6 (5·6%)	1 (2·3%)	1 (0·2%)	0	7 (0·7%)	2 (0·2%)
1a	0	0	0	0	1 (0·9%)	12 (27·3%)	0	1 (0·2%)	1 (0·1%)	13 (1·3%)
1b	0	3 (0·8%)	0	0	1 (0·9%)	13 (29·5%)	0	1 (0·2%)	1 (0·1%)	17 (1·7%)
2	35 (16·7%)	107 (28·7%)	44 (22·4 %)	19 (22·1 %)	49 (45·4%)	11 (25·0%)	41 (9·3%)	59 (12·4%)	169 (17·7 %)	196 (20·0 %)
3	149 (71·0%)	211 (56·6%)	138 (70·4 %)	53 (61·6 %)	43 (39·8%)	4 (9·1%)	227 (51·7%)	235 (49·5%)	557 (58·4 %)	503 (51·4 %)
4	19 (9·0%)	1 (0·3%)	11 (5·6%)	0	2 (1·9%)	0	102 (23·2%)	6 (1·3%)	134 (14·1 %)	7 (0·7%)
4a	6 (2·9%)	47 (12·6%)	3 (1·5%)	13 (15·1 %)	0	1 (2·3%)	28 (6·4%)	77 (16·2%)	37 (3·9%)	138 (14·1 %)
4b	0	1 (0·3%)	0	0	0	0	15 (3·4%)	71 (14·9%)	15 (1·6%)	72 (7·4%)
X	0	2 (0·5%)	0	1 (1·2%)	1 (0·9%)	2 (4·5%)	24 (5·5%)	25 (5·3%)	25 (2·6%)	30 (3·1%)

Appendix D

Pre-treatment variables	UHS "Chemo" (N =210) (%)	OUH "Chemo" (N =373) (%)	UHS "CRT" (N =196) (%)	OUH "CRT" (N =86) (%)	UHS "Surgery" (N =108) (%)	OUH "Surgery" (N = 44) (%)	UHS Palliative (N =439) (%)	OUH Palliative (N =475) (%)	UHS Total (N = 953) (%)	Total (N = 978) (%)
cN stage										
0	41 (19·5%)	140 (37·5%)	64 (32·7 %)	32 (37·2 %)	59 (54·6%)	1 (2·3%)	90 (20·5%)	108 (22·7%)	254 (26·7 %)	313 (32·0 %)
1	138 (65·7%)	144 (38·6%)	112 (57·1 %)	27 (31·4 %)	42 (39·9%)	33 (75·0%)	145 (33·0%)	131 (27·6%)	437 (45·9 %)	310 (31·7 %)
2	31 (14·8%)	76 (20·4%)	19 (9·7%)	25 (29·1 %)	6 (5·6%)	8 (18·2%)	127 (28·9%)	1551 (31·8%)	183 (19·2 %)	253 (25·9 %)
3	0	12 (3·2%)	3 (1·5%)	2 (2·3%)	1 (0·9%)	1 (2·3%)	59 (13·4%)	82 (17·3%)	61 (6·4%)	97 (9·9%)
X	0	1 (0·3%)	0	0	0	1 (2·3%)	18 (4·1%)	3 (0·6%)	18 (1·9%)	5 (0·5%)
cM stage										
0	210 (100%)	373 (100%)	196 (100%)	86 (100%)	108 (100%)	44 (100%)	176 (40·1%)	209 (44·0%)	690 (72·4 %)	712 (72·8 %)
1	0	0	0	0	0	0	257 (58·5%)	263 (55·4%)	257 (27·0 %)	263 (26·9 %)
X	0	0	0	0	0	0	6 (1·4%)	3 (0·6%)	6 (0·6%)	3 (0·3%)
Tumour location										
Proximal Oesophagus	0	1 (0·3%)	3 (1.5%)	0	0	0	19 (4·3%)	19 (4·0%)	22 (2·3%)	20 (2·0%)
Mid oesophagus	6 (2·9%)	28 (7·5%)	22 (11·2 %)	24 (27·9 %)	10 (9·3%)	6 (13·6%)	64 (14·6%)	118 (24·8%)	102 (10·7 %)	176 (18·0 %)
Distal Oesophagus	120 (57·1%)	124 (33·2%)	148 (75·5 %)	33 (38·4 %)	67 (62·0%)	16 (36·4%)	252 (57·4%)	148 (31·2%)	570 (59·8 %)	321 (32·8 %)
Siewert 1	24 (11·4%)	142 (38·1%)	8 (4·1%)	19 (22·1 %)	4 (3·7%)	9 (20·5%)	20 (4·6%)	86 (18·1%)	56 (5·9%)	256 (26·2 %)
Siewert 2	39 (18·6%)	78 (20·9%)	10 (5·1%)	10 (11·6 %)	19 (17·6%)	13 (29·5%)	56 (12·8%)	104 (21·9%)	124 (13·0 %)	205 (21·0 %)
Siewert 3	23 (11·0%)	0	1 (0·5%)	0	5 (4·6%)	0	28 (6·4%)	0	57 (6·0%)	0
Siewert undefined	15 (7·1%)	0	4 (2·0%)	0	3 (2·8%)	0	0	0	22 (2·3%)	0
Tissue Histology										
Adenocarcinoma	197 (93·8%)	343 (92·0%)	134 (68·4 %)	45 (52·3 %)	96 (88·9%)	40 (90·9%)	322 (73·3%)	352 (74·1%)	749 (78·6 %)	780 (79·8 %)
Squamous Cell	13 (6·2%)	30 (8·0%)	62 (31·6 %)	41 (47·7 %)	12 (11·1%)	4 (9·1%)	117 (26·7%)	123 (25·9%)	204 (21·4 %)	198 (20·2 %)
Dysplasia	0	0	0	0	0	0	0	0	0	0

Appendix D

Pre-treatment variables	UHS "Chemo" (N =210) (%)	OUH "Chemo" (N =373) (%)	UHS "CRT" (N =196) (%)	OUH "CRT" (N =86) (%)	UHS "Surgery" (N =108) (%)	OUH "Surgery" (N = 44) (%)	UHS Palliative (N =439) (%)	OUH Palliative (N =475) (%)	UHS Total (N = 953) (%)	Total (N = 978) (%)
Co-morbidities										
Chronic pulmonary disease (CPD)	26 (12·4%)	63 (16·9%)	28 (14·3 %)	15 (17·4 %)	19 (17·6%)	10 (22·7%)	57 (13·0%)	91 (19·2%)	130 (13·6 %)	179 (18·3 %)
Peripheral vascular disease (PVD)	6 (2·9%)	6 (1·6%)	7 (3·6%)	2 (2·3%)	5 (4·6%)	0	25 (5·7%)	15 (3·2%)	43 (4·5%)	23 (2·4%)
Cerebrovascular disease (CVD)	8 (3·8%)	8 (2·1%)	6 (3·1%)	3 (3·5%)	8 (7·4%)	0	84 (19·1%)	33 (6·9%)	106 (11·1 %)	44 (4·5%)
Uncomplicated diabetes (DM uncomp)	21 (10·0%)	57 (15·3%)	20 (10·2 %)	11 (12·8 %)	16 (14·8%)	10 (22·7%)	71 (16·2%)	77 (16·2%)	128 (13·4 %)	155 (15·8 %)
Leukaemia	0	1 (0·3%)	0	0	3 (2·8%)	0	7 (1·6%)	0	4 (0·4%)	1 (0·1%)
Lymphoma	1 (0·5%)	8 (2·1%)	2 (1·0%)	2 (2·3%)	3 (2·8%)	0	5 (1·1%)	3 (0·6%)	11 (1·2%)	13 (1·3%)
Renal disease	0	6 (1·6%)	1 (0·5%)	2 (2·3%)	3 (2·8%)	3 (6·8%)	35 (8·0%)	23 (4·8%)	39 (4·1%)	34 (3·5%)

Supplemental Table 5 **Demographics for the Palliative training cohort (UHS) and validation cohort (OUH). Standardized Mean Differences (SMD) are provided for the two cohorts. An SMD of 0.2 is considered a small difference, 0.5 medium and 0.8 or more, a large difference**

Pre-treatment variables (Palliative)	**UHS (N =437) (%)**	**OUH (N =475) (%)**	**SMD**
Gender			
Male	318 (72·8%)	355 (74·7%)	0·045
Female	119 (27·2%)	120 (25·3%)	
Median Age in years (Range)	75·2 (29·8 – 96·7)	72·0 (33·0-96·0)	0·172
Performance status			
0	74 (16·9%)	240 (50·5%)	0·808
1	131 (30·0%)	121 (25·5%)	
2	140 (32·0%)	69 (14·5%)	
3	87 (19·9%)	43 (9·1%)	
4	5 (1·1%)	2 (0·4%)	
cT stage			
0	1 (0·2%)	0	0·891
Is	0	0	
1	1(0·2%)	0	
1a	0	1 (0·2%)	
1b	0	1 (0·2%)	
2	40 (9·2%)	59 (12·4%)	
3	227 (51·9%)	235 (49·5%)	
4	102 (23·3%)	6 (1·3%)	
4a	28 (6·4%)	77 (16·2%)	
4b	14 (3·2%)	71 (14·9%)	
X	24 (5·5%)	0	
cN stage			
0	90 (20·6%)	108 (22·7%)	0·274
1	143 (32·7%)	131 (27·6%)	
2	127 (29·1%)	151 (31·8%)	
3	59 (13·5%)	82 (17·3%)	
X	18 (4·1%)	3 (0·6%)	

Appendix D

Pre-treatment variables (Palliative)	**UHS (N =437) (%)**	**OUH (N =475) (%)**	**SMD**
cM stage			
0	176 (40·3%)	209 (44·0%)	0·102
1	255 (58·4%)	263 (55·4%)	
X	6 (1·4%)	3 (0·6%)	
Tumour location			
Proximal Oesophagus	18 (4·1%)	19 (4·0%)	0·799
Mid oesophagus	63 (14·4%)	118 (24·8%)	
Distal Oesophagus	252 (57·7%)	148 (31·6%)	
Siewert 1	20 (4·6%)	86 (18·1%)	
Siewert 2	56 (12·8%)	104 (21·9%)	
Siewert 3	28 (6·4%)	0	
Siewert undefined	0	0	
Tissue Histology			
Adenocarcinoma	322 (73·7%)	352 (74·1%)	0·010
Squamous Cell	115 (26·3%)	123 (25·9%)	
Difficulty passing gastroscope and/or severe dysphagia			
Yes	222 (50·8%)	289 (60·8%)	0·203
Co-morbidities			
Chronic pulmonary disease (CPD)	57 (13·0%)	91 (19·1%)	0·167
Peripheral vascular disease (PVD)	25 (5·7%)	15 (3·2%)	0·125
Cerebrovascular disease (CVD)	84 (19·2%)	33 (6·9%)	0·370
Uncomplicated diabetes (DM uncomp)	71 (16·2%)	77 (16·2%)	0·001
Leukaemia	1 (0·2%)	0	0·068
Lymphoma	5 (1·1%)	3 (0·6%)	0·055
Renal disease	34 (7·8%)	23 (4·8%)	0·121

Supplemental Table 6 - Comparison of cohort composition between UHS and OUH patients

	UHS (Training)	OUH (Validation)
Gender	Approximately 3:1 male: female distribution in both cohorts	
Age	UHS cohort slightly older (median Age 70 yrs) versus the Oxford cohort (median age 68)	
Performance status	PS scores were broadly distributed across PS0-2 in UHS cohort	PS scores were heavily weighted towards the PS0 in the OUH cohort, indicating a generally fitter population at presentation
cT stage	The majority of UHS and OUH cases presented with T3 disease with similar distributions across T2-4 however more UHS cases were coded as cT4(unspecified) while more OUH cases were specifically designated cT4a/b.	
cN stage	Similar distribution of N0-2 disease in both cohorts but a higher prevalence of N1 staging in UHS cohort.	
cM stage	Both cohorts presented with comparable distributions of cM0 vs cM1 disease	
Tumour location	The majority of UHS tumours were distal oesophageal	OUH tumours were evenly spread across both the distal oesophagus and the GOJ.
Tumour Histology	Distribution of histology was comparable in both cohorts with an approximately 80% of tumours OAC versus 20% OSCC	
Co-morbidities	Higher prevalence of cerebrovascular disease	Lower prevalence of Cerebrovascular disease.

Supplemental Table 7 - Primary classifier model performance over 1000 bootstraps

AUCs	Mean	Range	SD	95% CI (one sample t-test)
MLR	0.8665	0.8176 - 0.9149	0.0145	0.8656 - 0.8674
RF	0.8674	0.8288 - 0.9105	0.0125	0.8666 - 0.8682
XGB	0.8627	0.8087 - 0.9023	0.0132	0.8619 - 0.8636

Supplemental Table 8 - Statistical comparison of primary classifier model performance on Kruskal-Wallis analysis

Overall p < 0.001	MLR	RF
RF	p = 0.31	-
XGB	p <0.001	p <0.001

Supplemental Table 9 - Mean classification AUCs for UHS model trained on 1047 cases with endoscopic resection class included (N = 94). Best performance for each class is highlighted in bold

UHS Model	Chemo	CRT	Surgery	Endo	Palliative	Mean
MLR	**0·906**	0·886	0·859	0·992	**0·984**	0·925±0·060
XGB	**0·906**	0·874	**0·889**	**0·993**	**0·984**	**0·929±0·055**
RF	0·893	0·856	0·860	0·981	0·980	0·914±0·062

Supplemental Table 10 - Palliative classifier model performance over 1000 bootstraps

AUCs	Mean	Range	SD	95% CI (one sample t-test)
MLR	0.7355	0.6516 - 0.8284	0.0282	0.7338 - 0.7373
RF	0.7808	0.7159 - 0.8402	0.0197	0.7780 - 0.7821
XGB	0.7989	0.7339 - 0.8806	0.0202	0.7976 - 0.8001

Supplemental Table 11 - Statistical comparison of palliative classifier model performance on Kruskal-Wallis analysis

Overall p < 0.001	MLR	RF
RF	<0.0001	-
XGB	<0.0001	<0.0001

Supplemental Table 12 - Kaplan Meier survival estimator for the palliative UHS and OUH cohorts. Hazard ratios based on Cox's Proportional Hazards provided with statistically significant differences in p values denoted by * if P <0.05, ** if P <0.01 & * if P <0.001**

UHS					
Treatment	**N**	**Events**	**Median Survival (months)**	**95% CI**	**HR (95% CI) (BSC as reference group)**
BSC	56	56	2·15	1·3 – 3·5	-
Chemo	148	134	11·1	9·7 – 12·2	0.27 (0.20-0.37)***
RTX	78	72	8·4	7·1 – 12·9	0.31 (0.22-0.44)***
Stent	113	113	3·9	3·1 – 4·2	0.77 (0.56-1.06)
Stent_Onc	42	41	6·0	4·3 – 8·5	0.56 (0.37-0.84)**
OUH					
Treatment	**N**	**Events**	**Median Survival (months)**	**95% CI**	**HR (95% CI) (BSC as reference group)**
BSC	34	25	5·8	4·4 – 15·8	-
Chemo	147	122	11·2	9·9 – 12·9	0.73 (0.47-1.12)
RTX	133	98	9·7	8·7 – 12·0	0.89 (0.57-1.39)
Stent	86	76	4·4	3·7 – 5·7	0.56 (1.12-2.80)*
Stent_Onc	75	68	6·7	5·7 – 8·2	1.46 (0.92-2.31)

Supplemental Table 13 - Mean classification AUCs for primary model using OUH as the training cohort and validating on UHS patients. Best performances by class are highlighted in bold both locally and in the validation cohort

Oxford model		"Chemo"	"CRT"	"Surgery"	"Palliative"	Mean (±SD)
MLR	OUH	**0·922**	**0·834**	0·894	0·972	0·906±0·058
	UHS Validation	0·853	**0·848**	0·782	0·970	0·863±0·078
XGB	OUH	0·917	**0·834**	**0·924**	0·975	**0·913±0·058**
	UHS Validation	**0·862**	0·829	**0·819**	**0·975**	**0·871±0·072**
RF	OUH	0·910	0·810	0·851	**0·976**	0·887±0·072
	UHS Validation	0·853	0·780	0·800	**0·975**	0·852±0·086

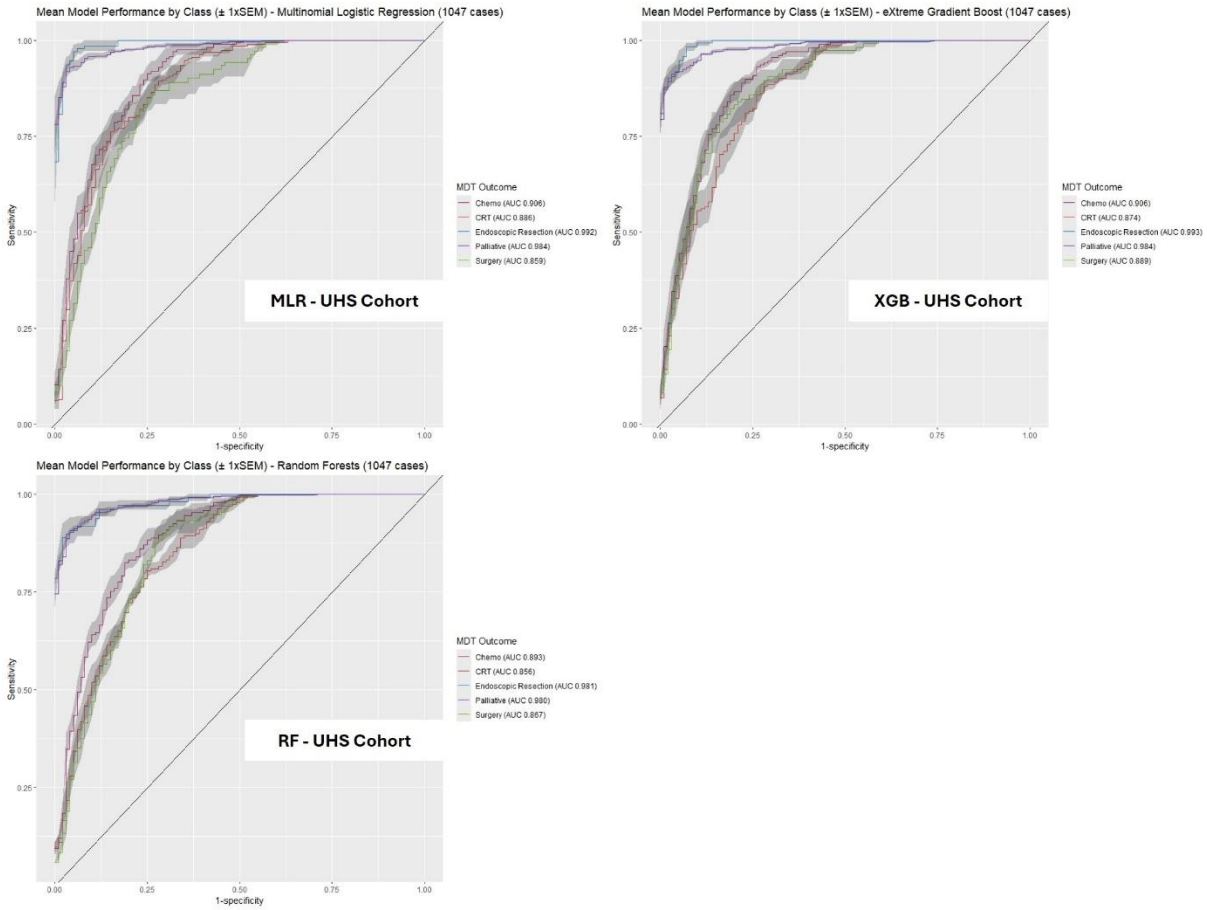
Supplemental Table 14 - Mean classification AUCS for palliative classifier model using OUH as the training cohort and validating on UHS patients. Best performances by class are highlighted in bold both locally and in the validation cohort

Oxford model		"Chemo"	"BSC"	"RTX"	"Stent"	"Stent_Onc"	Mean (±SD)
MLR	OUH	0·829	**0·803**	0·776	**0·707**	0·640	0·751±0·077
	UHS Validation	**0·881**	0·659	0·668	0·712	0·685	0·721±0·092
XGB	OUH	**0·831**	0·777	**0·787**	0·696	**0·661**	**0·750±0·070**
	UHS Validation	0·872	**0·698**	0·722	**0·837**	**0·731**	**0·772±0·077**
RF	OUH	0·818	0·776	0·779	0·690	0·603	0·733±0·087
	UHS Validation	0·844	0·683	**0·725**	0·751	0·646	0·730±0·075

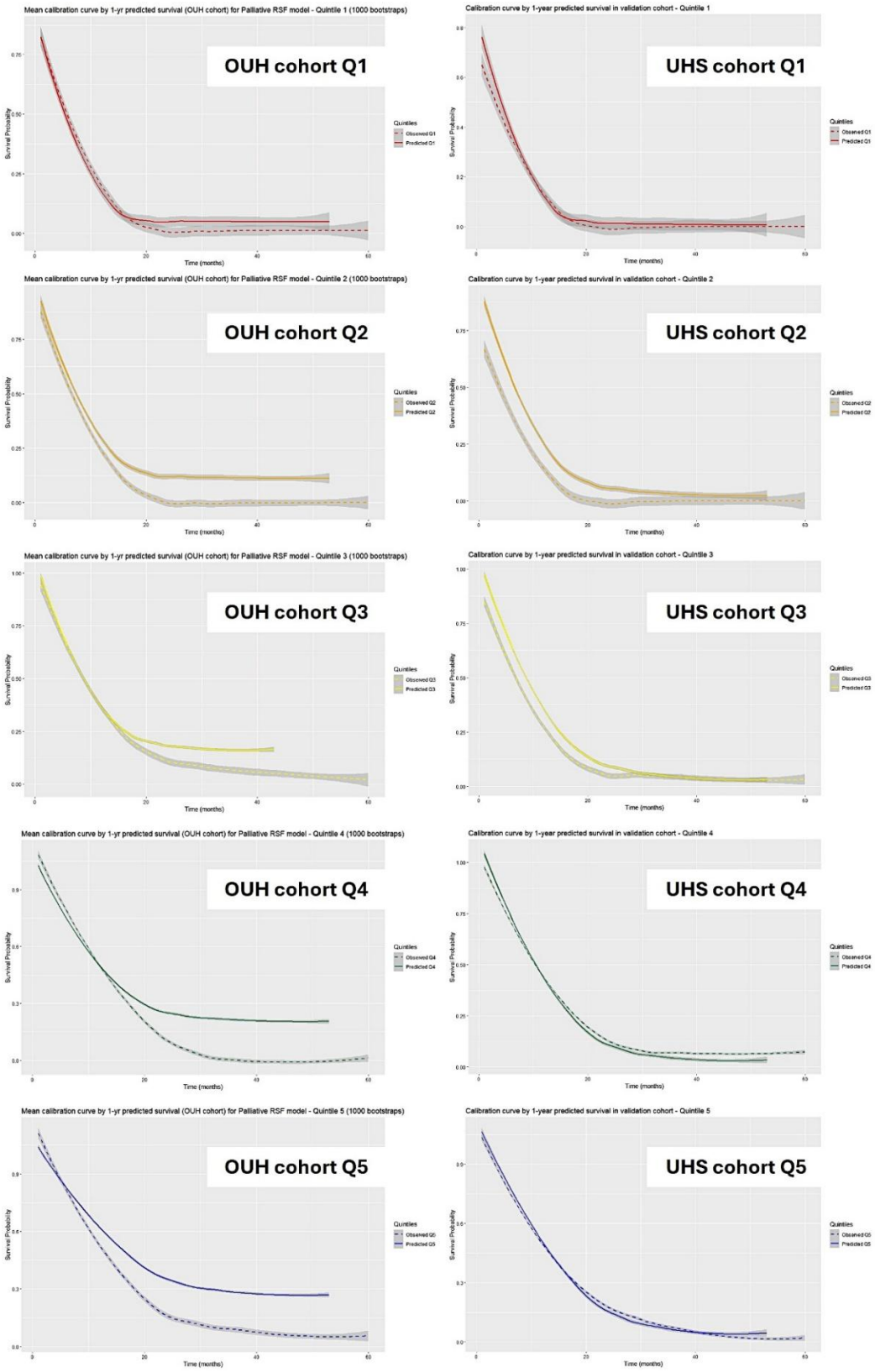
Supplemental Table 15 - Survival model performance metrics for OUH model and UHS validation cohorts

Metric	Cohort	Score	Reference	Interpretation
Prediction error (1-Concordance)	OUH model	0·336±0·021	0 = perfect concordance 1 = perfect non-concordance	Fair
	UHS validation set	0·340		Fair
CRPS (Integrated Brier Score/time)	OUH model	0·146±0·017	0 = perfectly accurate model 1 = perfectly inaccurate model	Very Good
	UHS validation set	0·101		Very Good

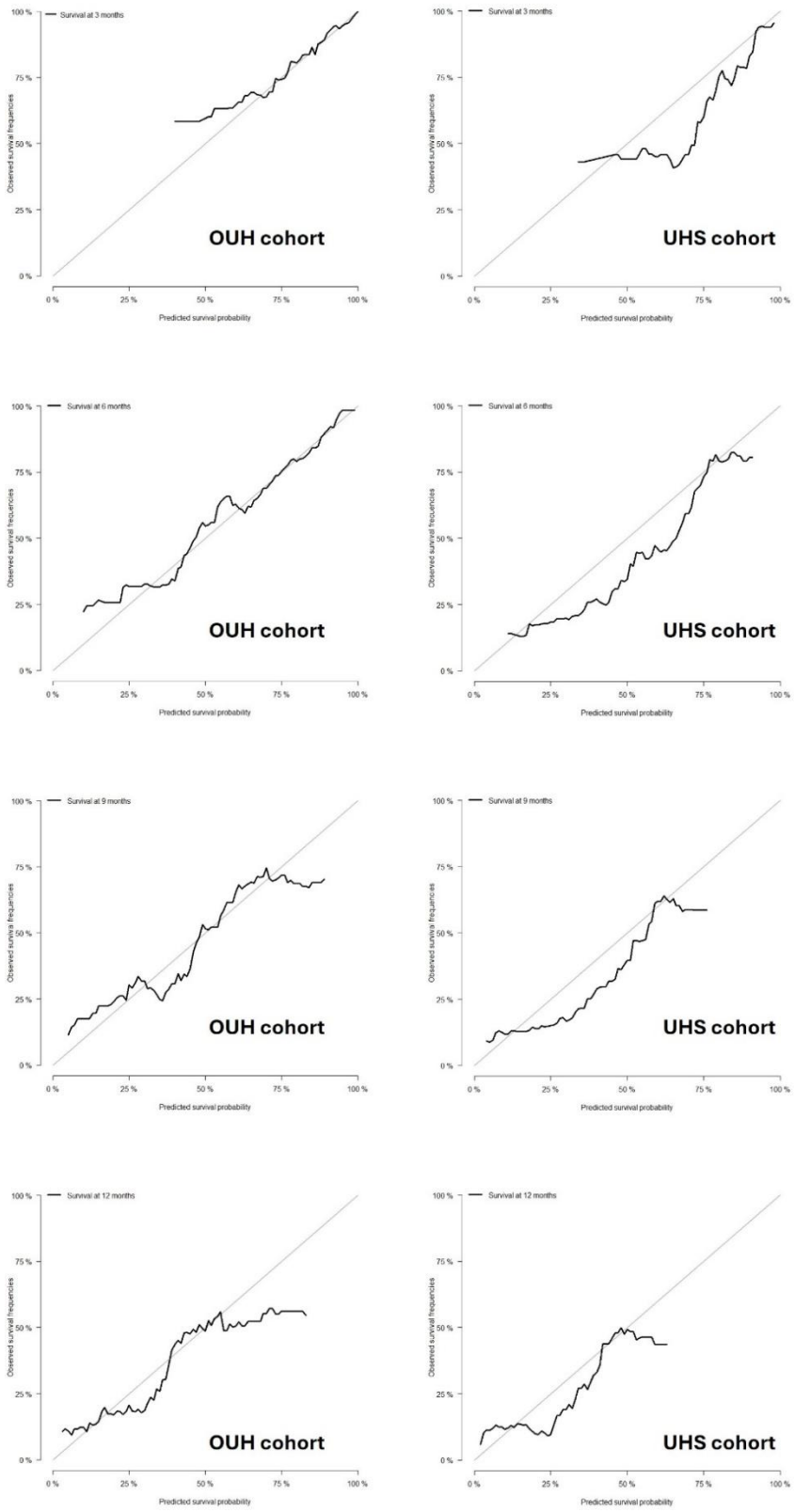
D.4 Supplemental Figures



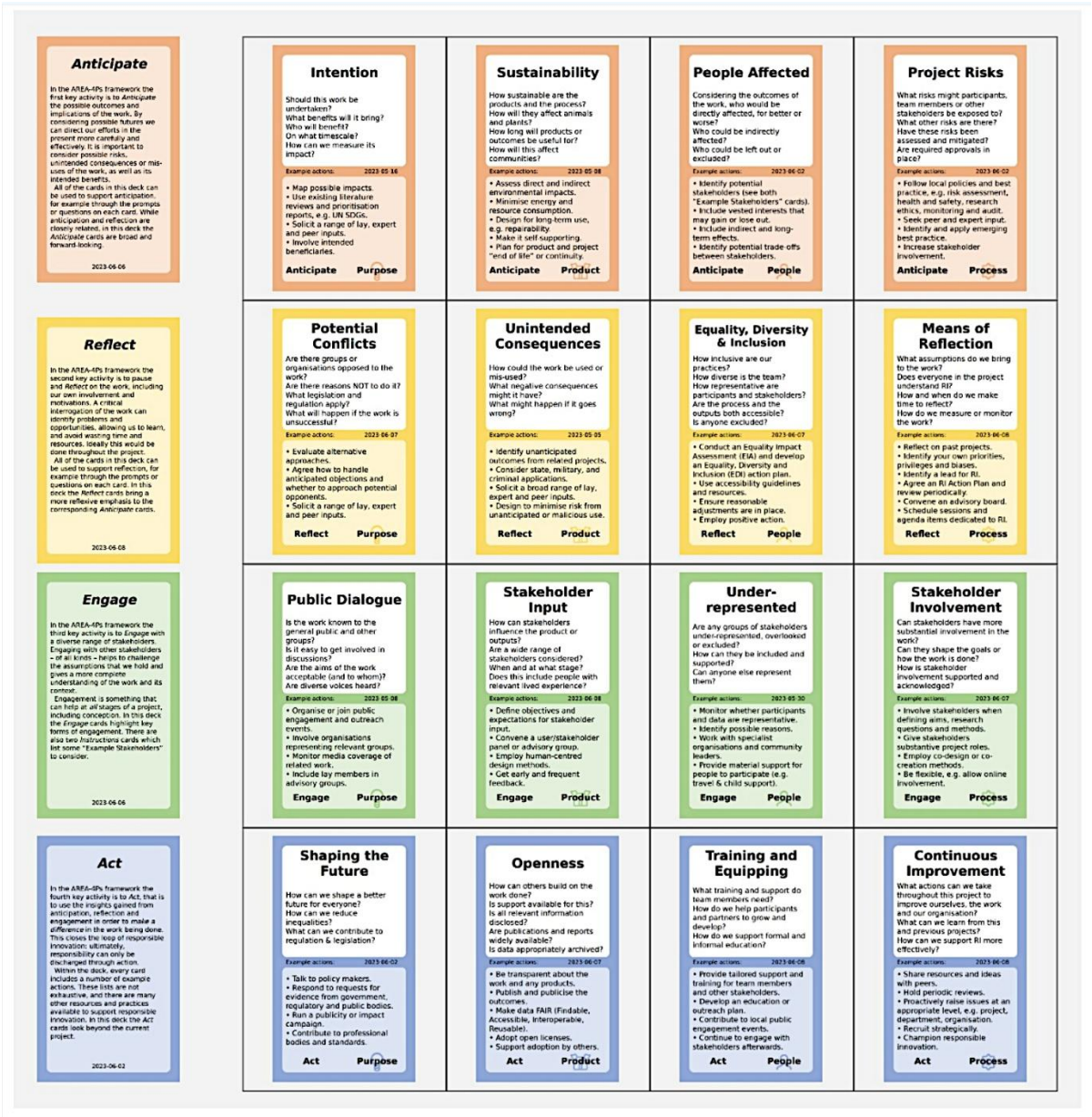
Supplemental Figure 10 - Mean cross-validated ROC curves for each classifier algorithm (UHS cohort, 1047 cases) when model incorporates endoscopic resection class (N = 94). Shaded areas represent ±1x Standard error from the Mean.



Supplemental Figure 11 - Quintile Calibration curves for OUH survival model vs UHS validation cohort, plotted with standard error over 60 months with cases stratified by predicted 1-year survival probability (Quintile 1 = 0-20% (a), Quintile 2 = 20-40% (b), Quintile 3 = 40-60% (c), Quintile 4 = 60-80% (d), Quintile 5 = 80-100%)



Supplemental Figure 12 - Calibration plots for the OUH survival model versus UHS validation cohort at 3,6,9 and 12 months post-diagnosis.



Supplemental Figure 13 - RRI Card deck. Print friendly version available at http://doi.org/10.17639/nott.7353

D.5 Final model hyperparameters

Primary classifier algorithm	Hyperparameters
MLR	Weight decay: 0
RF	mtry: 6, ntree: 500
XGB	nrounds: 50, max_depth: 2, eta: 0.3, gamma: 0, colsample_bytree: 0.6, min_child_weight: 1, subsample: 0.5

Palliative classifier algorithm	Hyperparameters
MLR	Weight decay: 0
RF	mtry: 6, ntree: 500
XGB	nrounds: 50, max_depth: 1, eta: 0.3, gamma: 0, colsample_bytree: 0.6, min_child_weight: 1, subsample: 0.75

Palliative classifier algorithm	Hyperparameters
RSF	Ntree: 1000, mtry: 17, nodesize: 2, nsplit: 10, splitrule: logrank

Appendix E Methodologies used within this research

E.1 Summary

Traditional statistical models produce outputs based on pre-determined set of conditions or "rules" in addition to input data. Whenever new data is then provided, these models continue to assume that the same rules apply, generating an output accordingly. Machine learning tasks advance this process significantly. ML algorithms comprise a sequential set of operations followed by the machine to examine large, complex, input datasets (and in some situations the "final answers") to learn the structural rules that allow it to produce a relevant model which best explains the original data while predicting on new input data. ML models aim to represent complex real-world events statistically, to a close approximation rather than a perfect replication. Multiple algorithms have been developed globally to solve similar issues, and while some algorithms perform better in specific tasks, no single algorithm will handle all tasks or data superlatively (285). A significant aspect of ML consequently involves testing multiple algorithms to determine the one best suited to a given problem or task (286).

E.2 Categories of Machine Learning

Machine Learning is broadly categorised into supervised, unsupervised learning, semi-supervised learning, and reinforcement learning (287). Within this work, the availability of labelled data allowed for the focus to rest on a supervised approach.

E.2.1 Supervised Learning

Supervised learning requires labelled data (the ground truth is known at the time of model training). It may be further sub-classified into classification tasks (where the outcome variable is typically comprised of discrete groups), regression tasks (the outcome is a continuous variable) and deep learning (which makes use of neural networks to find patterns with data) (287).

Classification tasks may be binary in nature such as predicting a yes/no outcome for a given survey question, or multinomial such as grouping images of cattle breeds, or indeed choosing a

medical treatment from a list of defined options. Regression models can predict the share prices of a commercial company or the age of a prospective new customer for a business.

Deep Learning is commonly applied to more complex data types such as audio, visual, and textual data. These models are designed to mimic human neural tissue networks creating layers of inter-connected nodes through which to pass data. They can produce immensely well-performing models however they are not without their drawbacks. They are computationally expensive taking anywhere from hours to days for model training. They require immense volumes of training data (100s to 1000s of observations) to perform in comparison to "shallower" methods which often handle and excel in smaller datasets (286). The models are always "black-box" in nature, as they typically favour performance over interpretability.

E.3 Feature selection and exclusion of cardiopulmonary exercise testing

Features were selected through a combination of a priori domain expertise and established features from current UK clinical guidelines for OC management (122,179,186,191,192,201).

CPET is a regular aspect of the UHS OC MDT process and is considered the gold-standard test for physical fitness (288,289). The testing process generates a significant amount of data which maps the complex interplay between pulmonary, cardiac and skeletal muscle physiology during exercise. However, owing to a combination of resource-intensity, complexity and availability it is not universally utilised (290). Patients who are intended for oesophagectomy typically undergo CPET evaluation following which they are assigned a risk profile based on the CPET clinician's interpretation of the data. The risk is stratified into "LOW", "INTERMEDIATE" and "HIGH" which is then returned to the MDT for inclusion within their deliberations. Anecdotally however, the process is vulnerable to clinician bias as the weighting of these recommendations are not always used in a consistent fashion. Additionally, some evidence indicates CPET data is more adept at forecasting patients not-at-risk of complications than it is at predicting those at-risk (291). Consequently, while CPET offers the potential for more granular physiological data which is salient to MDT decisions, its inclusion within the modelling process is restricted by its own limited application: namely to those only planned for oesophagectomy. As a result, I opted not to include it within this pilot study as a core predictor variable as the longer-term expansion of this process would inevitably need to be applicable to

those on non-curative pathways and those on curative pathways which do not include oesophagectomy (definitive CRT in selected cases being an example of this).

E.4 Machine Learning Algorithms

E.4.1 Considerations during model training

(1) Sample size estimation

Within this work, a sample size estimation was not carried out. This is primarily as sample size estimation is a statistical principle typically used for hypothesis testing and in healthcare settings for methodologies such as randomized controlled trials with an effect size to detect (292). Sample size calculations are a key means in healthcare studies to ensure that sufficient data is available to detect and corroborate a reported outcome and to minimise the risk of Type 1 and 2 errors (more so perhaps in the latter). Small datasets can lead to misrepresenting the target population, creating uncertainty of predictor effect or importance, mis-calibrated predictions and ultimately under confidence in the model's overall validity (293). In ML models there is typically no specific hypothesis to test nor an "effect size" to detect although it is recognised that sample size calculations may become more relevant in AI based studies over time where "risk" of an outcome is considered, specifically to ensure the risk predicted within an AI model is commensurate with the risk a patient is exposed to in the real world.

In this work, the primary goal was to maximise predictive performance across a range of measures and algorithms with no ceiling set for this (as we would aim for the greatest performance possible). While an infinitely increasing sample size may experience a plateau in performance beyond which no additional samples provide information gain, within healthcare settings this is rarely achieved and so not a realistic consideration. Additionally, for treatment allocation, this outcome variable is not an example of "hazard/risk" and so sample size calculation would not be relevant. For prognosis however, this would become relevant within this research, but as the datasets used throughout this research were already the maximum available (and eligible) at any given time, a sample size calculation remain redundant as the sample size ceiling was already reached and dictated by data availability (in other words, within reason, all possible data was already captured). Consequently, within this work a pragmatic aim was set to leave the sample size unrestricted in the interests of maximising training data (294).

(2) Parameters and hyperparameters

During model training, some information can be learned by the algorithm from the data and some information must be provided to it. Parameters are an example of the former, values learned from the data during training and determine how the trained model makes predictions when taking in new data. Parameters are consequently internal to the trained model, learned during the training phase. Taking a simple linear equation for instance:

$$y = mx + c$$

The gradient "m" here would be considered a parameter. By comparison, the process by which an algorithm learns from the training data may be adjusted or controlled externally by the user. Hyperparameters are the "settings" used to achieve this control during training. They vary in nature depending on the algorithm being used and may include settings such as the level of complexity a tree-based model should be allowed to reach, the type of cost function an algorithm uses during training or even simply the ratio of data-splitting between training and testing sets. Hyperparameters are thus integral to the process but do not form part of the final model itself.

(3) Bias-Variance Trade-off

The objective of model training is to produce a mathematical framework capable of understanding the underlying patterns within our original sample data which can be used to produce accurate predictions when new data from the wider population are presented to it. An underlying assumption is that the original sample is exactly representative of the larger population. In practice however this is rarely so and consequently, we must subject models to testing using data not seen during the training phase (typically termed "training" sets and "testing" or "validation" sets). This aims to give us a more realistic view of how our models will handle new data in when deployed in real world scenarios.

Models which are too simplistic, perhaps using too few predictor variables to adequately map the training data will typically perform poorly regardless of whether the data passed to it was from the original training set or a test set. This is termed "underfitting" and is associated with a high "bias" (an analogy here would be a marksman firing at a target; the grouping of shots

maybe close however if there is an error in setting up the gun's sights, the shots may all end up too far to the left).

Other models may by comparison perform exceedingly well on training data yet struggle to perform well when shown new data, often because these models are overly complex, using too many predictors for the size of the data being handled and result in models which have learned too much from the training sets (including any noise found within the data). The model generalises poorly onto new data and if we take our marksman analogy again here, the shots are now more central over the bullseye, but the overall grouping is poor – this is termed "variance", and the model is considered "overfitted". In practice it is very rare to achieve models with low bias AND low variance, so the objective is to aim for a middle ground – hence the term "trade off" (295).

E.4.2 Linear models

(1) Linear Regression

Linear regression models are used to examine variables linked by a linear relationship. A condition of using these models is that the outcome variable (y) is a continuous variable such as age, weight, height etc. The predictor variables however may be continuous or categorical and the relationship will resemble:

$$y = mx + c$$

A scatter plot of the data in these scenarios will usually show if a relationship is linear and thus appropriate for a linear regression model.

(2) Logistic regression

In situations where a relationship is suspected to be linear, but the dependent variable is categorical, we may use a logistic regression model (296). These models are useful in classification tasks where we wish to predict the probability of a defined set of possible outcomes whether binary or multi-class. While of a similar concept to linear regression models, it is more complex in nature and harder to assess graphically. While the ultimate objective

remains to fit a line with an intercept to the data (as for linear regression models), logistic regression models utilise a logistic function (an equation which calculate the probability of a case belong to a given outcome category or "class").

Taking binary logistic regression models as an example, we can see readily that with an outcome of only two possible types, translating the data into a scatter plot to fit a regression line is problematic. Instead, we can apply a mathematical transformation of the outcome variable (a "logit" or "log odds"). As the odds of an event is calculated by the ratio of probability of an event occurring versus not occurring, the log odds here is calculated by taking the natural log of this ratio (and is thus directly related to the original outcome):

$$\text{logit} = \ln\left(p/(1-p)\right)$$

The value of the logit can extend from positive infinity to negative infinity. When interpreting the values: a positive value indicates an event is more likely to occur, a negative value indicates the event is less likely to occur and zero means the probability is equally likely to occur as not to occur. If we are handling multiple predictor variables, these can be added linearly within the equation. Eventually we may derive the following equation for a given outcome class:

$$z = \ln\left(p/(1-p)\right) = c + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots \beta_k x_k$$

Here c is the "y" intercept, each β is a co-efficient, of a predictor variable "x". Beta-coefficients are hard to conceptualise but can be used as weightings of relative risk by exponentiating the coefficient. This provides the individual odds ratio for that parameter independent of the other predictors (296). This provides useful insights into the relationship between variables and if they are positive or negative predictors for the given outcome. While the odds ratios for continuous variables are calculated similarly, it is worth noting here that the resulting odds ratio reflects the increase in odds per unit increase in the variable itself.

As we are likely to wish to know the given probability of an outcome class, we can then take the resulting logit value from the equation above and convert it to the predicted probability of that outcome class using the following:

$$p = 1/\left(1 + e^{-z}\right)$$

In a binary classification scenario, if our p value is > 0.5 it can then be classified as the positive class or negative class if <0.5. The numerical difference between the predicted outcome and the observed is termed the "residual", which if plotted, can also be a helpful way of evaluating models.

(a) Advantages of logistic regression models

These models are typically easy to train, interpret and implement. They are one of the most explainable and interpretable forms of machine learning. The process can be extended to multiple classes (this is discussed in Section (3)) and quantifies the influence a given predictor possesses over the outcome both in magnitude (beta coefficient) and direction (positive versus negative). Of practical benefit, these models are also fast to train and make predictions as well as accurate in linearly separable datasets. Finally logistic regression models can be assessed on Receiver Operator Characteristic (ROC) curves allowing for evaluating performance in discriminating between classes (this will be discussed further in section E.5.5). Finally, they allow for adjustment of confounders by allowing the inclusion of any such features within the model (278).

(b) Disadvantages of logistic regression models

By nature, linear models are dependent on the presence of linear relationships within the dataset and are limited to predicting categorical outcomes. In scenarios with higher order relationships between variables these models are not appropriate and are typically outperformed easily by more complex algorithms such as neural networks. High-dimensionality datasets (where the number of predictor variables may even outnumber the sample size) are also prone to dilution of individual feature significance and larger standard errors, needing careful feature selection when developing the model (278,297). The use of odds ratios may be misleading as the chosen units for the continuous predictors can distort the size of the resulting odds ratio.

(3) Multinomial Logistic Regression

Multinomial logistic regression models extend the process of the logistic regression models described in the previous section to more than two classes. Again, the outcome variable remains categorical and requires that at least one predictor is also categorical. To extend the

process to multiple (K) classes, for a single predictor, we fit K-1 lines, with K-1 intercepts and K-1 slopes. Multinomial logistic regression models make several assumptions. Firstly, that the data contains linear relationships between the predictor and the outcomes. It also assumes that the predictors do not contain significant outliers and that all cases are independent of each other (for example, multiple data points do not originate from the same source). Finally, we assume that there is no multi-collinearity (i.e. that there is no correlation between two or more of the predictors) as this lessens the confidence within the resulting regression coefficients and significance values.

The modelling process is alike to running multiple binary logistic regression models however with multinomial regression the equations are fitted in parallel, and the outcome class probabilities will sum to 1. Where we have more than two classes, rather than estimating a single logit for each new case as described in section (2), we instead estimate a separate logit for each possible outcome class for a given case. These are then passed to an equation called a SoftMax function which will transform these logits into probabilities that sum to 1. The outcome class with the highest predicted probability will then be considered the winning class. When evaluating model performance using a ROC curve, we must extend the process to a ROC "surface" instead.

(a) Advantages

Advantages remain similar to those for binary logistic regression discussed in section (2)(a). The immediate benefit of a multinomial regression here is the avoiding the need to coerce multiple classes to a binary classification problem which would risk loss of information and statistical power (298). Furthermore, we are also able to apply this technique to outcome classes whether they are ordered or not.

(b) Disadvantages

As described above, these models require linear problems to solve, are vulnerable to outliers and do not handle multi-collinearity within the predictor variables well.

E.4.3 Tree-based models

(1) Decision Trees

Decision trees are the most basic form of a tree-based model and are commonly used within business and finance for mapping out complex decisions and potential outcomes. They may be considered effectively as flow charts which take variable inputs at successive branch points to produce a final outcome. The structure of a decision tree comprises a "root" node (representing the first, most important variable from which the route through the tree splits), followed by a series of "branch" nodes comprising the remaining variables used during model training (299). Finally, the tree ends at "leaf" nodes, where no further branching is considered beneficial to the tree and so the decision-making process typically terminates here (300). The tree is considered complete when all branch nodes lead to a leaf node. Decision trees can be used in both regression and classification tasks, and with multiple variables available, several different trees may be produced when varying the root node. Determining the best single tree is thus a key part of the training process.

Decision trees may be grown in several different ways. The "greedy" approach utilises heuristic problem-solving (the process of making the optimal choice locally at each node as a strategy of approximating the best choices overall). At each node therefore we choose our best variable with which to split the data. The data is then split based on this test condition and the process repeats until a leaf node is reached. The challenge with this method is determining the root node initially and this is based on a combination of entropy and information gain. Entropy relates to the homogeneity of the data (if there is complete homogeneity we have an entropy of "0" versus a data with a 50/50 split where entropy is "1"). Information gain therefore is the change in entropy when a branch or root node is split. During tree growth we are aiming to maximise information gain, and this determines the variables chosen at each branching point (301).

Gini impurity by comparison measures the disorder within a set of objects and is calculated as the probability of mis-classifying an object if it was randomly classified based on the distribution of those objects within the dataset. The impurity of a node is calculated as:

Appendix E

$$\text{Gini impurity} = 1 - (p(A)^2 + p(B)^2)$$

Here p(A) is the probability of class A and p(B) the probability of class B within the node. The impurity across the split is then calculated as:

$$\text{Gini impurity}_{split} = GI_{left} \text{ x proportion of cases from parent node} + GI_{right} \text{ x proportion of cases from parent}$$

Finally, we can calculate the Gini gain by:

$$\text{Gini gain} = \text{Gini impurity}_{parent\ node} - \text{Gini impurity}_{split}$$

The algorithm works to minimise the Gini impurity after each split, testing multiple variables for the next branch node, and selecting the one that leads to subsequent node containing cases which are as homogenous as possible (once there are only cases of a single class left (and thus a leaf node) this is termed “pure”).

The choice of Gini impurity versus information gain as a method for growing trees is largely equivalent as they agree in 98% of cases (302). Where entropy calculation for information gain may arguably be fractionally slower computationally speaking is in the need to compute a logarithmic function which is not required in Gini impurity.

Once a tree is grown it may benefit from “pruning” – a process which acts to trim down the complexity of the tree and minimise overfitting. This is turn is aimed at preserving generalisability.

(a) Advantages

Decision trees are quick to grow and easy to understand even for lay users, grouping them within the class of explainable machine learning algorithms. They can handle categorical and continuous variables and do not make assumptions about the underlying data distribution and can also handle missing data (299).

(b) Disadvantages

They are not ideal for regression tasks and may miss-classify more frequently in scenarios of high dimensionality datasets. In some situations, they can be computationally expensive to train as at each node, sorting of each branch node candidate variable is necessary to produce the best split (286). Pruning algorithms are also computationally intensive as many candidate sub-trees must be produced and tested. Finally individual trees while intuitive are often weak learners and prone to over-fitting.

(2) Random Forests

A random forest is a tree-based ensemble algorithm (one that learns from combining outputs from individual learners to produce higher accuracy and lower variance performance). As the name indicates random forest models are developed by growing many decision trees (301). They are popular as they are both high performing by nature and applicable to both classification and regression tasks.

To begin, the original training set is used to produce a pre-specified number of bootstrapped samples (this is resampling WITH replacement and will be discussed further in section E.6.1). from each bootstrapped dataset, an unpruned decision tree is then grown in parallel or “ntree”. Importantly, each tree is independent of each other as a separate training dataset is linked to that tree. Where all predictors are included in the growth of each tree, this is termed “bootstrapped aggregation” or “bagging” for short. In random forest models however, each tree in the forest selects a random subset of the predictor variable pool (“mtry”) from which to generate branching nodes. Thus, while individual trees may be underfitted, taken collectively, this minimises variance and overfitting. Finally, when new data is passed through the forest, in classification tasks, the outputs of all trees are considered and a “majority vote” taken (124). For regression tasks, an average of the results is calculated. When evaluating performance, the Out of Bag (OOB) error may be used. For classification, this is essentially the accuracy of predictions on the cases not selected during bootstrapped resampling (and thus not seen by the machine during training). For regression the R-Square and Root Mean Squared Error is calculated.

(a) Advantages

Random forest algorithms are generally well regarded as they minimise variance, can handle high dimensionality data and can also assist with feature selection tasks. As the hyperparameters are relatively few, they are also fairly user friendly. If the user hardware supports parallel-processing, computation time may also be rapid. They can also handle class imbalance well (156).

(b) Disadvantages

Random forest models can minimise variance, but they do not necessarily minimise bias (286). As the number of trees is typically in the order of many hundreds, they are not particularly interpretable or explainable either.

(3) eXtreme Gradient Boost (XGBoost)

An important aspect of RF and bagging algorithms is that models produced remain independent of each other (and can thus be executed in parallel). Boosting by comparison takes a sequential and dependent approach (303). Two main subtypes include adaptive and gradient boosting. In adaptive boosting training data is not simulated in parallel sample groups, but rather the original training data is bootstrapped with the probability of each selected case receiving equal weighting to train a preliminary model. At this point the algorithm generates a new dataset which is weighted towards observations incorrectly classified in the first iteration. The sequence repeats with a new models trained on iteratively bootstrapped datasets k-fold times with updated case and model weights through sequential voting with each iteration. Once the kth model is trained, we have our final ensemble where each tree vote is given, weighted by the individual tree's weighting. As each new model introduces a bias towards misclassified data from the previous model, this process effectively lets the algorithm "learn" from its mistakes with the final prediction for each test case taking the form of a linear combination of the weighted results. Gradient boosting is a similar process, with the main difference being how the loss function is evaluated (286). With adaptive boosting, the overall accuracy determines the next set of weights, with gradient boosting, it is defined by the residual values - the difference between the true observed value and the predicted value, (for instance if the predicted probability of an event in a binary outcome is 1, and the predicted probability was 0.79, the residual would be 0.21). The eXtreme Gradient Boost (XGBoost) model is one of the most well-known and well-regarded forms of gradient boosting, capable of building branches of different

tree in parallel, handling data which is missing and utilises regularisation (the process by which predictor variables are weighted based on how much information they contribute to the model, and less useful variables are either minimised or removed entirely) (125). XGBoost models have shown exceptional performance in competition (286), however processing-in-series does require longer computation time.

(a) Advantages

These models leverage multiple weak learners which provides excellent accuracy without expending additional computation on utilising more complex "strong" learners within the ensemble. As each new model is correcting mistakes from previous models, we reduce bias while bootstrapping also reduces variance. XGBoost in particular, handles class imbalance, data which is missing and utilises regularisation (the process by which predictor variables are weighted based on how much information they contribute to the model, and less useful variables are either minimised or removed entirely) to further optimise models (156,286).

(b) Disadvantages

Even with parallel processing, ensemble training is computationally expensive while the black box nature of the ensemble models again removes interpretability and explainability. Adaptive boosting techniques are sensitive to outliers and struggle with noisy data.

E.4.4 Final selection of ML algorithms within this work

I selected four established classifier ML algorithms which were implementable via the "caret" package; Multinomial Logistic Regression (MLR) (123), Random Forests (RF) (124), Extreme Gradient Boost (XGB) (125), and Decision Tree (DT) (126). I also selected a fifth, Random Forests based survival algorithm (Random Survival Forests, RSF) for the survival modelling (153).

MLR is a quick, efficient, simple to implement and inherently interpretable algorithm which bridges statistical and machine learning spheres. It is capable of handling multiple outcome classes, can allow regularization of features to ensure non-informative features are weighted appropriately and able to handle both continuous and categorical variables which makes it well suited for this study. The MLR model was trained using the "nnet" package extension with L2 regularisation.

Decision Trees are again quick to grow, inherently interpretable and easy to understand. They again can handle a mixture of categorical and continuous variables and can provide useful visualization of the decision-making logic. Decision Trees were trained using the “rpart” package.

Random forests while not inherently interpretable, can be explained with explainable AI techniques readily, are high-performing (by leveraging multiple weak learners), and act to minimise variance through random selection of predictor variable for each component decision tree within the ensemble. The RF model was trained using the “randomForest” package.

Extreme Gradient Boost is a highly respected ensemble algorithm, able to handle class imbalances well, incorporates regularization and learns iteratively between successive rounds of modelling to minimise miss-classification. The XGB model was trained using the “xgboost” package.

Random Survival Forests (“randomForestsSRC package”) by Ishwaran et al., is computationally rapid, using parallel processing and generates predicted survival probabilities for a patient at every unique death time point within the training cohort.

I chose the included algorithms to provide a degree of diversity of ML techniques (regression-based, tree-based and ensemble) and to offer a computationally inexpensive approach to the classification task. Within the current digital infrastructure of the National Health Service, computationally expensive deep learning architectures would prove logistically problematic and long-term implementation of these models needs to factor in such limitations.

E.5 Classification Performance Metrics

Numerous metrics are available for assessing the predictive performance of machine learning models. Each has its own application and role within model evaluation. The following are examples of metrics by which classification models may be assessed (304).

E.5.1 Accuracy

This is one of the simplest and most readily understandable metrics. It summarizes classification performance by dividing the total correct classifications by the total number of classifications made:

$$\text{Accuracy} = (TP+TN)/TP + FN + FP + TN)$$

Where TP = true positive, TN = true negative, FP = false positive and FN = false negative. This metric is not advisable however in scenarios where the outcome classes are significantly imbalanced. For example, taking an extreme case: suppose we wish to predict the presence of a rare condition only present in 1% of patients. The correct identification of the condition is of clinical importance yet were a model to classify all cases as "negative", accuracy would be 99% despite the mis-classification rate being 100%! Consequently, accuracy is best used then groups are balanced, and all classes are equally important to the problem at hand.

E.5.2 Balanced Accuracy

The issue of class imbalance for simple accuracy is circumvented to a degree by using balanced accuracy provided there is not a high skew within the data or if some classes are of more importance than others. It is calculated as the mean of sensitivity and specificity:

$$\text{Balanced accuracy} = (\text{sensitivity} + \text{specificity})/2$$

Sensitivity (also referred to as True Positive Rate or Recall) is:

$$\text{Sensitivity} = TP/(TP+FN)$$

While Specificity (also referred to as True Negative Rate) is:

$$\text{Specificity} = TN/(TN+FP)$$

E.5.3 Recall

Recall as described above is also known as sensitivity or true positive rate. It differs slightly from accuracy: as it summarizes the total correct positive predictions made, out of all possible positive predictions:

Recall = TP/(TP+FN)

E.5.4 Precision

Precision is defined as the number of correct positive predictions made from ALL positive predictions made:

$$\text{Precision} = TP/(TP+FP)$$

It is effectively a form of accuracy for the true-positive cases and is also referred to as the Positive Predictive Value or PPV. In other words, if a case was classed as positive, what is the probability that it truly represents a positive case.

E.5.5 Area Under the Curve (AUC)

Receiver Operator Characteristics (ROC) were introduced during the second world war to test how well radar operators could discriminate noise from true signals in radar detection (305). In their original form, ROC curves graphically convey how well a test or model discriminates between two groups when the goal is accuracy of classification. A series of probability thresholds are plotted on the graph of Sensitivity (y axis) versus 1-Specificity (x axis) which can then be used to determine an acceptable sensitivity and specificity for the model to operate at. The area under the curve is calculated to reduce the plot to a single, easy to understand value. On a 1x1 square plot, if the AUC is 1.0 then it represents perfect discrimination (the curve essentially forms a square). If, however, the curve follows the diagonal line (AUC = 0.5) then the model is no better than random chance at discriminating between the two classes. The ideal AUC is thus as close to 1.0 as possible although in real-world scenarios this is rarely achieved. The AUC may be thought of as equivalent to the C-Statistic (which gives the likelihood that a randomly selected case with the desired outcome will have a higher predicted probability than a randomly selected case without the outcome).

The AUC has become widespread within the medical literature for evaluating logistic regression models however it is subject to some limitations. It is typically overly optimistic when data is skewed (306) or the same data is used to train and test the model – requiring instead some form of internal and external validation. Additionally, high AUC values suggest strong discrimination but do not automatically infer that the predicted probabilities are accurate. AUCs/C-statistics

give equal weight to sensitivity and specificity however in clinical contexts it may be desirable to err towards accepting false positives over false negatives or vice-versa in specific clinical scenarios. The C-statistic may be influenced by predictor value distribution and can be insensitive to the addition of further variables (50).

E.5.6 Precision-Recall AUC

As discussed above, precision is an assessment of how many "positive" predictions were truly positive whereas recall measures how many of the positive samples within the dataset were correctly identified. The two offer a trade-off as increasing one typically costs the other. A Precision-Recall curve illustrates both on one plot where, the closer to the upper right-hand corner our curve lies, the better the overall model performance. PR Curves are felt to be a better alternative to ROC curves in cases of highly skewed datasets (306). It allows some judgement in finding a balance between precision and recall – as the curve lets us determine at which recall value our precision starts to fall away when setting a probability threshold for the model to operate at. Where ROC curves are beneficial in situations where all classes are equally important, PR curves (and their AUC) is more useful if the positive class is more important.

E.5.7 F1 Score

F1 scores represents another method by which we may combine precision and recall to evaluate model accuracy using the harmonic mean of the two (307). This represented by:

$$\text{Harmonic mean of x and y} = 2/(1/x) + (1/y)$$

When applied to precision and recall therefore:

$$F1 = 2 \times \text{Precision} \times \text{Recall}/\text{Precision} + \text{Recall}$$

The score is derived from the predicted class (as opposed to the predicted probability of the class as we are using recall and precision). This stems from the Fβ function:

$$F\beta = (1+\beta^2) \times ([\text{precision} \times \text{recall}] / \beta^2 \times [\text{precision} + \text{recall}])$$

Where the more we care about recall over precision for a given scenario, the higher the β value. For F1 score, therefore, we care equally about both. F2 scores weight recall twice as importance as precision. Conversely if $0 < \beta < 1$ then precision is more important. However, as the score does not factor in the true negative rate, it is less useful in imbalanced datasets where negatives are more important – in this scenario a balanced accuracy may be better employed. Additionally, F1 scores may be derived for class-wise comparison as well as a net-score across the model and so is suited to multi-class problems. The intuitiveness of the F1 score however suffers owing to its use of the harmonic mean, as well as its inherent bias towards whichever constituent property (precision or recall) is represented by the smaller numerical value (307).

E.5.8 Log Loss

While many of the metrics discussed thus far assess performance at the level of the class predicted, Log Loss assesses performance at the level of the predicted probability. Specifically, it indicates how close the predicted probability value is to the true value (in the case of binary tasks for example, 0 or 1). The Log Loss value increases based on the deviation of the two and the higher the value the poorer the performance. Consequently, it can act as a penalty function during model training.

$$\text{Log Loss}_i = -[y \ln p_i + (1-y_i) \ln(1-p_i)]$$

Where i = given observation/record, y = observed outcome, = predicted probability, and ln = natural log (i.e. using base e). Log Loss is calculated on each observation based on observed class and predicted prob and averaged to provide a single value. A model with perfect skill has an overall Log Loss value 0 with no upper numerical limit. Supplemental Table 16 outlines the common advantages and disadvantages to log loss along with a selection of the key metrics discussed in this chapter.

Supplemental Table 16 - Summary of advantages and disadvantages of key evaluation metrics for classification

	Advantages	Disadvantages
AUC	Copes better with class imbalance Good where all classes are equally important. Recognisable, visualisable Good where main interest lies in the final class label rather than the predicted probability scores. Good for ranking prediction scores too	Not yet well developed for multi-class problems but well established in binary classification Less useful if only specific classes are important.
PRAUC	Good for heavily imbalanced data Good where positive class is most important	Does not care about negative classes
Log loss	Common for multiclass Good for absolute probabilistic difference Useful as a penalty function during model training	Functionally symmetrical so poor if there is significant class imbalance
F1/F β	Good if interested in rare positive classes. Good in imbalanced sets where the positive class is most important. Provides a balance between recall and precision and seeks to optimise both	Not good in imbalanced sets where the negative class is more important Need to set a β depending on whether recall or precision is more important (or equally important)
Accuracy	Good where all classes are equally important. Simple and recognisable Intuitive	Performs poorly with imbalanced classes

	Advantages	Disadvantages
Balanced Accuracy	Good if imbalanced classes Calculated on class labels	Less effective if high skew in groups Not derived from predicted probability scores for class labels

E.6 Model validation

When models are trained, we wish to know how well they will perform on new data. Where data is scarce for a given use-case, obtaining large enough quantities to optimise statistical power and confidence in model performance may be challenging. It is useful to derive insight into the generalizability of the models and the error with which they may operate. Model validation can in broad terms be split into internal and external validation. External validation is simply achieved by passing new data from an external source into the model so that predictive performance may be assessed on data the model has never seen. By using external sources, we may also get a truer picture of how well the model behaves with data from the wider population or separate but related populations. In healthcare settings however obtaining related data from other units and centres can be difficult. Internal validation aims to derive an estimation of this generalizability error from the original dataset. Two of the most well-known methods involve sampling with replacement (bootstrapping) and without replacement (cross-validation) (308).

E.6.1 Bootstrapping

Bootstrapping is a form of inferential statistics as inferences about the larger population are made using the sample we possess. We assume that the data we possess is simply one from many possible samples which ultimately comprise the wider population and attempt to simulate it through resampling with replacement many hundreds of times. We may therefore train models from the resampled datasets and average out the performance to estimate generalizability of the final model. The central assumption is however that the sample is indeed a true representation of the population.

During the bootstrapping process, we place the original data into a “bag” from which we select an observation at random to form part of the new bootstrapped dataset. The observation is then replaced into the bag and another observation is selected again at random. Due to the replacement, each observation in the bag has the same chance of being selected during each

pick and crucially, observations may consequently be selected multiple times for inclusion within the new dataset. Within an infinitely sized sample, the proportion of original observations which appear within the new dataset will naturally tend towards 63.2%. The cases which were not picked in each bootstrap iteration are termed “out of bag” and are used as test cases. This process is used frequently in ensemble training methods such as random forest models where we can build a model for each bootstrapped dataset and aggregate the models into a majority vote (classification) or average the results (regression).

A benefit over traditional hypothesis testing is that unlike the latter we do not need equations which estimate the sampling distribution using properties of the sample data, the design of the experiment and a test statistic. As the sample sizes increase the process also converges towards the true sampling distribution in most situations (nor does bootstrapping make any assumptions of the underlying distributions). There are some limitations – as mentioned previously, we assume that our sample is representative of the wider population as only observations which were seen in the original dataset can occur in the subsequently resampled sets. It can be time and computationally expensive as hundreds or thousands of datasets are derived. It is also sensitive to outliers within the original data which may skew sample statistics in the resultant datasets.

E.6.2 Cross-Validation

Cross-Validation is a process of sampling without replacement as we effectively partition our available data into a “train” and “test” groups without overlap. The simplest form of validation here would be a single train and test group – this is known as “hold-out” validation. Models are trained on the train group, predictions made on the test group and compared to the observed outcome. The ratio of data partitioned vary depending on the nature of the data and the overall sample size. A bias-variance trade-off applies, as the larger the training set the larger the variance (but lower bias). It is best avoided unless data is very expensive to train however, as to achieve adequate statistical power using simple hold-out validation of the final model would typically require many thousands of observations which may not be feasible in most healthcare settings.

(1) k-fold Cross Validation

A progression of hold-out validation is k-fold cross-validation. Here the data is partitioned into even “folds”. A single fold is reserved for testing and the remaining folds are aggregated for

training. The resulting model performance is recorded and then a new model is trained on a different group of folds (with a new fold set aside for testing). This sequence is repeated for each of the folds and each observation will therefore be used in testing exactly once. The performance statistics from each model are then averaged at the end and provided with an error estimation. This process provides a more accurate estimate of model performance as it is averaged over many runs and each observation is used once. Commonly we select a k-value of 5-10 with or without repeats (here the k-fold validation sequence is run fully and then the data is repartitions into brand new folds as many times as is set by the user).

(2) Bootstrapping versus Cross Validation

Cross-validation techniques are less prone to bias however k-fold repeats can lead to a higher variance. Conversely, bootstrapping suffers more bias and less variance. Either method is generally acceptable, and the use-case often dictates the decision for one over the other (for instance scenarios where one wishes to trade variance for bias and vice-versa).

E.7 Explainability techniques used within this work

Machine Learning models vary inherently by how understandable their predictions are to humans. This can extend from “glass-box” algorithms where the logic behind decisions made are entirely intelligible and transparent through to “black box” models which are either too complex for human comprehension or obscured for protection of proprietary technology. The field of Machine Learning has therefore shifted in recognition of the need for human-intelligible modelling especially as it relates to high-stakes decision-making.

Unfortunately, while terms such as “interpretable” and “explainable” machine learning are used almost interchangeably there remains no consensus within the field on their respective definitions (309). Rudin however offered a distinction between the two: describing “interpretability” as an inherent property of the model to be intelligible in its natural state. An example of this would be a decision tree whereby all decision-splits may be seen and understood by the human user (310). By comparison Rudin defines “explainability” as a post-hoc process whereby an opaque black-box model’s decision logic is approximated after the fact, often with need of a second model devised solely to try and explain the original model. Explanation-based techniques while popular are not without pitfalls, as they cannot by nature offer perfect fidelity with the original model (else the original model would no longer be

necessary) nor can the user be sure they use the same features in the same way as the original model even if their explanations are accurate in most cases (310).

Despite this, techniques have been devised to attempt to provide insight into the functioning of models. For the purposes of this discussion, models which are inherently intelligible will be considered "interpretable" whereas the remainder of this section will relate to "explainability" i.e. the application of post-hoc techniques to provide insight into predictions in opaque/black box models. This may be further split into "global" versus "local" explainability techniques.

Global techniques offer insight into the model itself. This is a static picture once the final model is trained as the model will now no longer change and global insights will remain the same regardless of the predictions offered on a case-by-case basis. Two such examples are Variable Importance and Partial Dependence analyses. Local techniques however offer insight at the instance-level – this means that the explanation offered varies as the prediction itself varies. It thus offers some understanding of why a particular decision was reached which is particularly beneficial in healthcare settings. Two well-known examples of local techniques include Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) (209,210).

E.7.1 Variable importance

Variable importance (also referred to as feature importance) is a simple score assigned to each feature within the model which allows relative ranking of scores. How the scores are calculated vary greatly between algorithms (for instance importance values for a tree-based model it may be derived from Gini impurity or permutation based importance (whereby a feature is selectively perturbed in isolation to determine the effect it has on the final predictions), while a logistic regression model's features may be ranked based on the T-statistics derived from the beta-coefficient for each feature in the final equation) (311). Consequently, this makes a direct comparison of absolute values misleading and instead scaling values and considering overall rank may be more useful. Importance based techniques remain a useful tool however in proving global insight into a model as well as considering features to remove when applying feature selection.

E.7.2 Partial Dependence

Partial dependence (PD) is a simple but powerful tool in observing the individual effect of specific variables to the overall model and the final predictions (205). The process involves isolating one (or maximally two) features and resetting the values for all observations within that feature to the same values within the original dataset. The new dataset is then passed back through the model to generate new predicted probabilities. This process is then repeated iteratively with the perturbed feature(s) being incrementally changed each cycle. The new probabilities from each cycle may then be plotted visually to show the exact variation in probabilities across all possible values of that feature or features. When all cases are plotted this is an individual conditional expectation plot while a partial dependence plot averages out the cases to provide a single average curve (312). As interpretation is largely visual, attempting to apply this to more than two features is rarely beneficial where the limitation becomes the human ability to comprehend such plots in more than two dimensions.

E.7.3 Local Interpretable Model-agnostic Explanations (LIME)

LIME acts to explain the individual instance prediction being evaluated and was first introduced by Ribeiro et al in 2016 (209). The first step with this technique is to establish the decision boundary between outcome classes (these may be complex, and non-linear in nature). We then focus on a very small portion of the boundary where only a few observations are in the vicinity of the target instance. Local data points near the target observation are perturbed and weighted according to distance to the main prediction instance following which a new, simpler linear model is fitted. This new local model is then used to generate variable importance values for the features used as it relates to the original target observation.

The main benefit of LIME is that it is not tied to the parent model – hence it is termed model-agnostic. It instead relies on deriving an entirely new linear model. However, defining the "neighbourhood" of points to simulate new instances into is a significant challenge. It is also susceptible to the creation of improbable instances from the perturbed dataset using gaussian distributions, all of which affect the quality of the final explanations. The presence of non-linear interactions even within the linear model can also influence the fidelity of the explanation while features important to a particular instance may not be most important to the original model

globally. Finally, simulations have shown that even for two points in proximity, explanations may still differ significantly (313).

Appendix F Health Care Professionals' perceptions of machine learning based clinical decision support systems for oesophageal cancer management

Journal: Computers in Biology and Medicine, 2025, Impact factor 6.3, CiteScore: 13.0

Comput Biol Med. 2025, Dec; 200: 111373, doi: 10.1016/j.compbiomed.2025.111373

Catherine Webb[1]*, Navamayooran Thavanesan[2]*, Mohammed Naiseh[3], Rachel Dewar-Haggart[1], Tim Underwood[2], Ganesh Vigneswaran[2]

***These authors are joint First Author for this work.**

1 School of Primary Care, Population Sciences and Medical Education, University of Southampton.

[2] Innovation for Translation Research Group (ITRG), School of Cancer Sciences, Faculty of Medicine, University of Southampton

[3] Department of Computing and Informatics, Bournemouth University

Corresponding Author: Catherine Webb

Address: School of Primary Care, Population Sciences and Medical Education, University of Southampton, South Academic Block, University Hospitals Southampton, Tremona Road, Southampton, UK, SO16 6YD

Email: Catherine.webb@soton.ac.uk

ORCID ID:

CW – 0000-0002-8545-7936
NT – 0000-0002-7127-9606
MN – 0000-0002-4927-5086
Twitter (TJU): @TimTheSurgeon

RDH – 0000-0002-3757-1152
TJU – 0000-0001-9455-2188
GV – 0000-0002-4115-428X

Twitter (GV): @ganesh_vignes

Funding Support Acknowledgement:

NT receives a joint studentship from the Institute for Life Sciences (University of Southampton) and University Hospital Southampton. The project receives additional funding from the UKRI Trustworthy Autonomous Systems Hub (TAS Hub) Pump Priming Fund. The funding sources were not involved in study design, data collection, analysis, interpretation of data or writing of this manuscript

Author Contributions:
CW - conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, and writing– review & editing.
NT - conceptualization, data curation, formal analysis, investigation, methodology, resources, supervision, validation, writing – original draft, and writing– review & editing.
MN - data curation, formal analysis, methodology, visualization, writing– review & editing.
RDH - formal analysis, methodology, visualization, supervision, validation, writing– review & editing
TJU - conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing– review & editing.

GV - conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

Conflicts of Interests to declare: None

Manuscript category: Original Article.

F.1 Acknowledgements

The study outlined within this chapter has been accepted-in-principle for publication in the journal: Computers in Biology and Medicine, pending minor revisions. My self and my colleague Dr. Catherine Webb are co-first authors on this work, and I wish to acknowledge my co-authors for their contributions.

Contributions:

1) **Catherine Webb was involved in the conception of this work, including modifying and deploying the Qualtrics survey, primary qualitative data collection, primary qualitative data analysis, manuscript drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy and in the submission process. She also acted on co-author and journal reviewer comments to amend the final manuscript.**
2) **Navamayooran Thavanesan was involved in the conception of this work, primary quantitative data collection, primary ML data analysis, manuscript drafting, and revising for critical and important intellectual content, final approval, and agreement of accountability for accuracy**
    - **NT performed clinical data collection, collation, cleaning and coding. He performed coding for the initial ML models and model evaluation methods in R as well as the variable importance plots which were then repeated in python by Ganesh Vigneswaran for interval verification. He provided an initial draft of the Qualtrics survey which was subsequently amended, modified and deployed by Dr. Webb. He assisted in drafting the initial manuscript and made amendments to the subsequent drafts based on feedback by co-authors. He received and acted on reviewer comments and undertook agreement of accountability for accuracy**
3) Mohammad Naiseh was involved in primary qualitative data analysis, the reviewing and recommending revisions for critical and important intellectual content, final approval and accountability for accuracy.

4) Rachel Dewar Haggart was involved in study supervision, methodology guidance for the qualitative aspects of this study and the final approval
5) Timothy J Underwood was involved in overall study supervision, the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy
6) Ganesh Vigneswaran provided python coding for ML models (to verify initial models in R), overall study supervision, the reviewing and recommending revisions for critical and important intellectual content, final approval, and agreement of accountability for accuracy.

The CRediT Taxonomy is as follows:

CW - conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, and writing– review & editing.
NT - conceptualization, data curation, formal analysis, investigation, methodology, resources, supervision, validation, writing – original draft, and writing– review & editing.
MN - data curation, formal analysis, methodology, visualization, writing– review & editing.
RDH - formal analysis, methodology, visualization, supervision, validation, writing– review & editing
TJU - conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing– review & editing.
GV - conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, software, supervision, validation, visualization, and writing– review & editing.

F.2 Abstract

Oesophageal cancer (OC) causes significant morbidity and mortality. Multiple treatment regimens are available, and multidisciplinary team (MDT) decisions over which to offer are complex, multi-faceted and subject to logistical constraints and human factors. A machine learning (ML) model-based clinical decision support system (CDSS) for OC has been developed, trained on historical treatment decisions. However, clinician trust in such systems is not yet established.

This study surveyed clinicians in OC MDTs in the UK and Ireland to investigate which clinical and sociodemographic factors influence conscious decision-making in OC, comparing their relative subjective importance to those derived from the ML model (reflecting previous real-world practice). It also sought to explore clinicians' views on the potential use of artificial intelligence-based CDSSs in OC.

There was agreement between clinicians and the model in many of the most influential factors in decision making, although age and gender had greater influence on the model than their conscious importance to clinicians would support. Clinicians identified a wide range of

additional clinical and holistic factors outside the current model which factor into their decision-making, including further investigations, symptoms, nutrition and social factors.

The prospect of utilising an ML CDSS in future received generally positive feedback, although opinions varied widely. However, barriers to implementation were identified, including concerns around perceived clinician superiority over ML CDSSs, patient individuality, transparency and safeguarding, the need for evidence, and additional input requirements. As ML CDSSs are increasingly offered in practice, clinicians' reservations must be addressed and their need for transparency and evidence met.

F.3 Introduction

Oesophageal cancer (OC) remains a key public health issue (314). Treatment plans are determined by Upper Gastrointestinal (UGI) multidisciplinary teams (MDTs) which assimilate multi-domain expertise across a broad range of roles (122). MDT treatment selection is directed by disease burden, patient demographics, functional status and co-morbidities (108,122) and is unsurprisingly a critical determinant of patient outcomes. However, a growing, ageing population has increased MDT caseload volume and complexity (315). Case discussions may only last a few minutes per patient, suffer from incomplete information, and encounter challenging interpersonal dynamics (23).

The need to streamline and reform MDT processes is well recognised (26,315). One potential solution uses computerised Clinical Decision Support Systems (CDSSs), now emerging across many aspects of cancer-care including screening, diagnosis and treatment planning (40). CDSSs range from simple tools summarising clinical guidelines to complex systems integrating multiple data sources for patient-specific recommendations (40,316). The ongoing Medical Artificial Intelligence (MAI) boom within healthcare is one such vehicle, representing a global market worth $5 billion USD in 2022 and projected to rise to $70 billion USD by 2032 (317). Real-world MAI implementation remains in relative infancy (45). Carefully implemented MAI, may improve safety, efficiency, cost-effectiveness and unwarranted variation as well as geographical and sociodemographic inequalities (45,63,318,319).

However, without careful analysis of their inputs and processes, there is a risk of ineffective AI CDSSs perpetuating or worsening inequity (320). Many AI solutions are also typically 'black box' (the machine's underlying logic is unclear to the human user) (321), which can be problematic

within healthcare settings in establishing clinician and patient trust (321). Consequently, they introduce novel questions and challenges ethically, legally, and in their acceptability to clinicians and patients (322).

Machine learning (ML) is a branch of AI increasingly utilised in CDSSs across many specialties (323–325). ML uses computational power to identify patterns within large, complex datasets and make predictions. An explainable ML CDSS for OC patients has recently been under development at University Hospital Southampton (UHS), based on data from 893 OC patients discussed in MDTs between 2010-2022 (108,109). For the model to be integrated into clinical practice, it is crucial that the factors on which it bases decisions are consistent with standard of care practice and sound human clinical judgement. However, the relative importance clinicians attribute to many of the factors involved in these decisions is currently unknown. Furthermore, inequalities in treatment allocation have been noted by age, gender and ethnicity (187–189,326). Identifying which factors clinicians value as explicitly important when compared to ML models trained on historical ground-truth decisions offers insight into future Human-AI interactions where inconsistencies between explicit priorities and observed practice can then be interrogated. Factors influential in the model which are not explicitly important to clinicians, may speak to potential implicit clinician bias, which is capable of significant impact on clinical practice (190).

Within this study, we present the results of a nationwide survey of OC MDT clinicians who were asked to describe how they rationalise the importance of key clinical variables when making treatment plans for OC patients. Their responses were compared against a random forests-based CDSS trained on historical MDT decisions at UHS, a high-volume tertiary referral unit (109), to identify areas of concordance and discordance between the Human and the AI. They were also surveyed on their current perceptions, concerns and views about the use of ML-based CDSSs to identify potential barriers towards implementation of AI-derived CDSSs within the OC space.

F.4 Methods

An anonymous survey with multiple choice and free text questions was hosted on the Qualtrics™ platform (327) from October - December 2023. Invitations were emailed to clinicians via professional membership organisations (The Association of Upper GI Surgeons of Great Britain and Ireland (AUGIS), the UK and Ireland Oesophagogastric Cancer Group (UKIOG) and the British Society of Gastroenterology (BSG)) and shared on social media by the

researchers. Respondents were excluded if they did not self-report as regularly contributing to the UGI MDT in the UK or Ireland.

F.4.1 Questionnaire

Participants were asked to identify factors important in UGI MDT decision-making, selecting from 18 listed factors (Supplemental Table 17) plus free text 'other', and to rank their 10 most important factors. The list included sociodemographic factors plus variables used in the ML model, excluding 2 that were not applicable (timeframe of model training data and local geography). The ML model considers the 19 Charlson Comorbidity Index conditions separately (328), but these were combined in all but 1 question of the survey to reduce questionnaire burden.

To assess attitudes to the future use of ML CDSSs in OC, respondents were asked on a Likert scale how likely they would be to use one if available, and to describe any barriers.

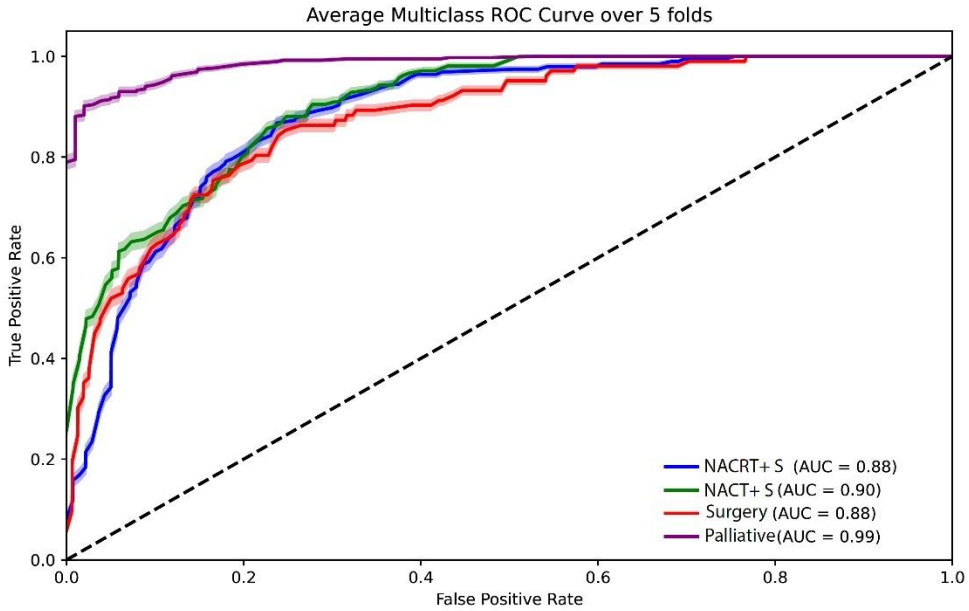
Supplemental Table 17 - Factors listed as options in the survey

Factor	In ML model?	Rationale for inclusion
Age	Y	Risk factor for frailty, comorbidities and complications
Gender	Y	Inequalities evident in literature (Dijksterhuis et al., 2021)
Performance status	Y	Fitness to withstand treatment – grading system of maximum activity tolerated
ASA grade	Y	Fitness to withstand treatment – anaesthetist's grading system
Tumour stage	Y	Size and local spread of tumour
Nodal stage	Y	Spread to lymph nodes
Metastasis stage	Y	Spread to distant organs
Tumour location	Y	Influences feasibility of different treatments
Biopsy / tumour histology	Y	Type of cancer cells
Tumour differentiation	N	Type of cancer cells – degree of abnormality
Comorbidities	Y	Fitness to withstand treatment and risk of complications
Smoking status	N	Risk factor for comorbidities/complications
Alcohol usage	N	Risk factor for comorbidities/complications
Patient geographical location	N	Access to various treatment modalities
Presence of local radiotherapy centre	N	Access to radiotherapy
Patient preference	N	Individual choice
Ethnicity	N	Inequalities evident in literature (Okereke et al., 2022)
Socioeconomic status	N	Inequalities evident in literature (Henson et al., 2018)
Other (free text)	N	Establish additional relevant factors

F.4.2 Development and validation of the ML MDT Model

The MDT ML model was developed using a random forests algorithm in Python ("Ranger" Library, sklearn v1.2.2, max depth = 6 (based on k=5 cross-validation)) using variables consistently available to the MDT prior to a final treatment decision (Supplemental Table 17). The model was trained on a cohort of 843 oesophageal cancer patients over a 12-year period managed at a tertiary referral centre. Previous publications described the model's development, established utility, and confirmed performance using area under the curve (AUC) from the multi-class Receiver Operator Characteristic (ROC) using "one" vs "others" approach

(Supplemental Figure 14) (108,109). Permutation-based feature importance was derived for each included variable.



Supplemental Figure 14 - Multiclass ROC curve for random forests treatment classifier representing a "one vs others" class-prediction performance. K = 5 Cross-validation was conducted using an 80:20 split. Mean ROC is presented +/- 1 Standard Error of the Mean Reproduced from Thavanesan et al., 2024. Computers in Biology and Medicine 180: 108978, p4.

F.4.3 Literature search string in development of conceptual framework in Supplemental Figure 18 performed on 6th March 2024

"Artificial Intelligence" OR "AI" OR "Machine Learning" OR "ML" OR "neural net*" OR "deep learning" OR "expert system"

AND

Clinician OR doctor OR physician OR nurse OR "health* professional" OR "health* staff" OR "medic* professional" OR "medic* staff"

AND

Views OR perceptions OR thoughts OR beliefs OR opinions OR qualitative OR interviews OR focus groups

AND

"Decision support" OR "decision aid" or "decision assist*"

(MESH: Decision Support Systems, Clinical/ or Decision Support Techniques/)

AND

Cancer OR tumour OR tumor OR neoplas*

Inclusion: all study designs, all applications of clinical decision support system (e,g. screening, diagnosis, treatment).

Exclusion: studies not specific to cancer (e.g. nephrology in general, not only kidney cancer).

F.4.4 Analysis

Data were managed in IBM SPSS Statistics for Macintosh v28.0.1.1 and Microsoft Excel v16.80. 'Top ten' responses were converted to numeric scores (top = ten points, second = nine etc).

Qualitative responses were analysed independently by 2 reviewers (CW and MN) through thematic analysis (284). To ensure consistency and objectivity, an inter-rater reliability test was conducted to assess the level of agreement between different coders. The resulting score of 0.73 indicated good reliability, aligning with the commonly accepted threshold of 0.7 or above. For respondent validation, themes were discussed with clinicians in semi-structured interviews conducted as part of the wider research programme.

F.4.5 Ethical Approval

This work was approved by the University of Southampton Ethics Committee (ERGO: 70375) and the Health Research Authority Research Ethics Committee (IRAS: 319540) as part of the Machine Learning in Oesophageal Cancer (M-LOC) study.

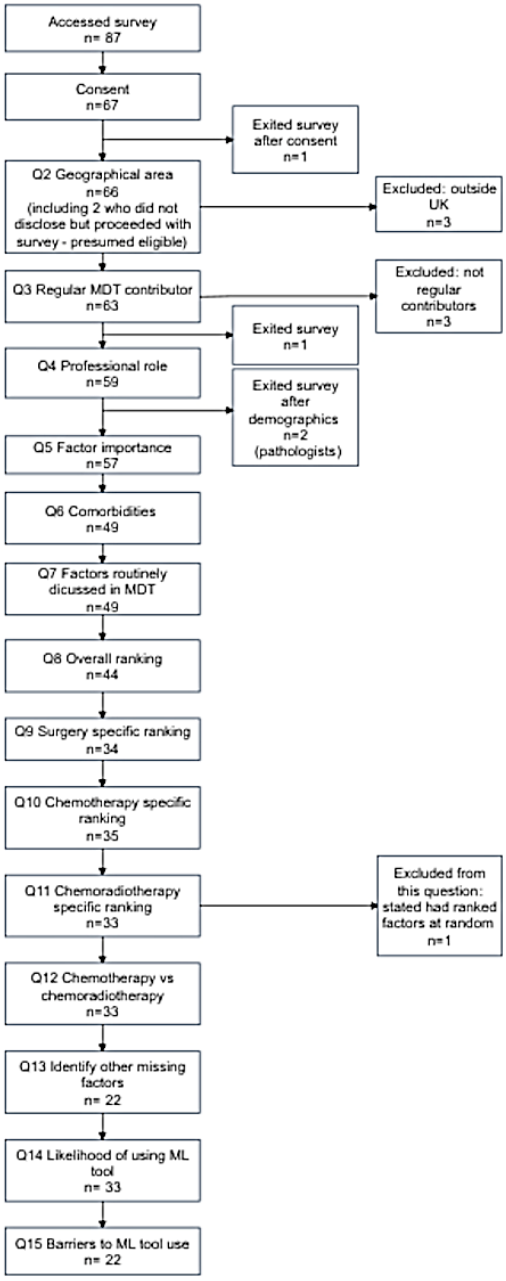
F.5 Results

F.5.1 Sample Size and Demographics

A total of 87 participants accessed the survey, of which 67 consented. Three respondents were excluded as they do not routinely attend the UGI MDT, three as they work outside the UK, and one left the survey immediately after consenting. Supplemental Table 18 shows the location and professional roles of the remaining 60 respondents. Response numbers vary by question as none were compulsory (except consent) and there was attrition throughout (Supplemental Figure 15).

Supplemental Table 18 - Work location, professional group and seniority of eligible respondents. Valid percentage excludes missing responses

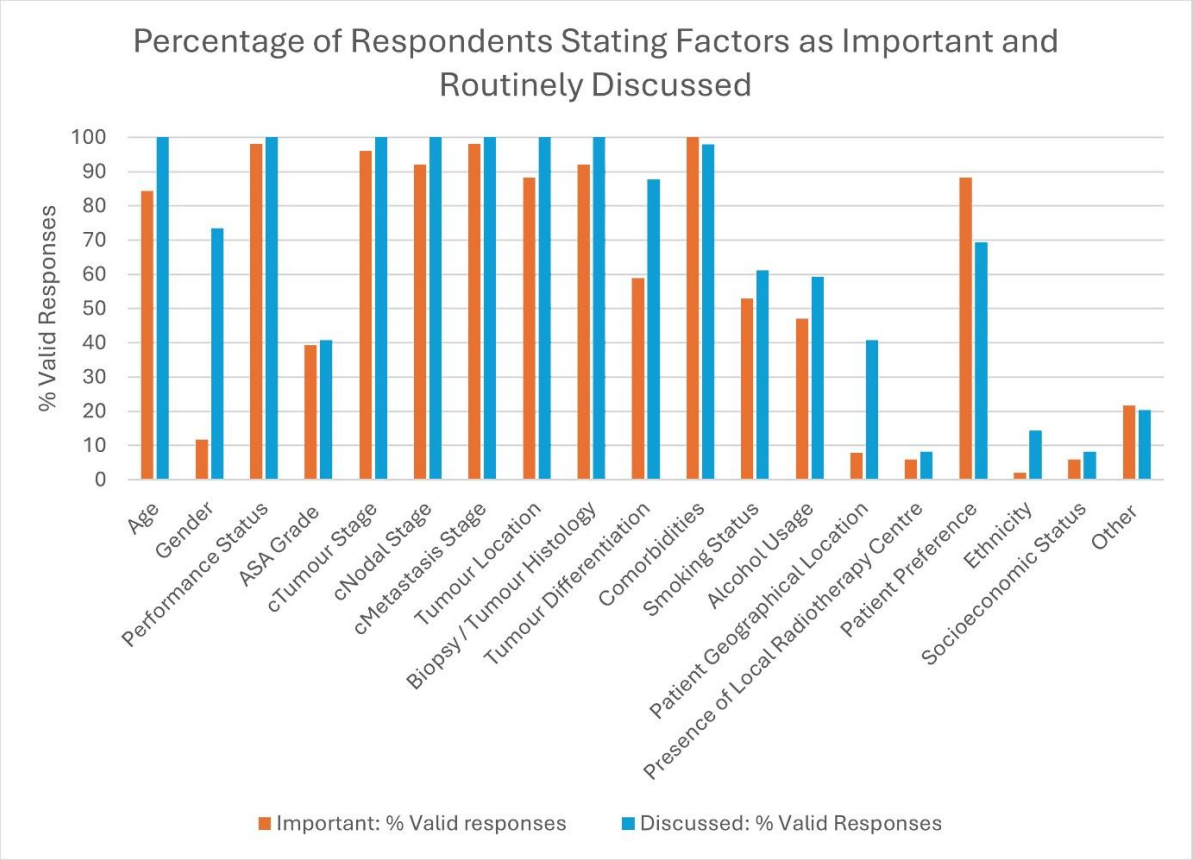
UK Geographical Area	**Number of Respondents (n= 58)**	**Valid Percentage (%)**
East Midlands (England)	3	5.2
West Midlands (England)	3	5.2
East of England	7	12.1
Kent, Surrey and Sussex	1	1.7
London	7	12.1
North East England	2	3.4
Yorkshire and The Humber	3	5.2
North West England	4	6.9
South West England	2	3.4
Thames Valley	3	5.2
Wessex	5	8.6
Scotland	7	12.1
Wales	4	6.9
Northern Ireland	5	8.6
Republic of Ireland	2	3.4
Missing	*2*	
Professional Role	**Number of Respondents (n=59)**	**Valid Percentage (%)**
Surgeon	21	35.6
Gastroenterologist	7	11.9
Medical Oncologist	5	8.5
Clinical Oncologist	17	28.8
Radiologist	1	1.7
Specialist Nurse	6	10.2
Pathologist*	2	3.4
Missing	*1*	
Grade	**Number of Respondents (n=54)**	**Valid Percentage (%)**
Consultant	43	79.6
Registrar	3	5.6
Fellow	2	3.7
Specialist Nurse	6	11.1
Valid Total	54	100.0
Missing	*6*	



Supplemental Figure 15 - Flow chart showing number of respondents per question and reasons for exclusions

F.5.2 Factor Importance and Discussion Frequency

Most factors deemed important by respondents were also identified as routinely discussed (Supplemental Figure 16). A disparity was seen for patient preference, which was more often deemed important (88%) than discussed (69%).



Supplemental Figure 16 - Clustered bar chart showing the percentages of respondents deeming each factor important to them in OC treatment decisions (n = 57) and the percentage reporting that each factor is routinely discussed in MDTs (n = 49)

F.5.3 Relative Factor Importance in Overall Treatment Decisions

Supplemental Figure 14shows the ROC curve for the ML model. Supplemental Table 19 compares the relative importance of each factor from the survey ranking to that of the ML model when determining overall treatment decisions.

Supplemental Table 19 - Comparison of overall ranking of factors from the survey with rankings from the ML model

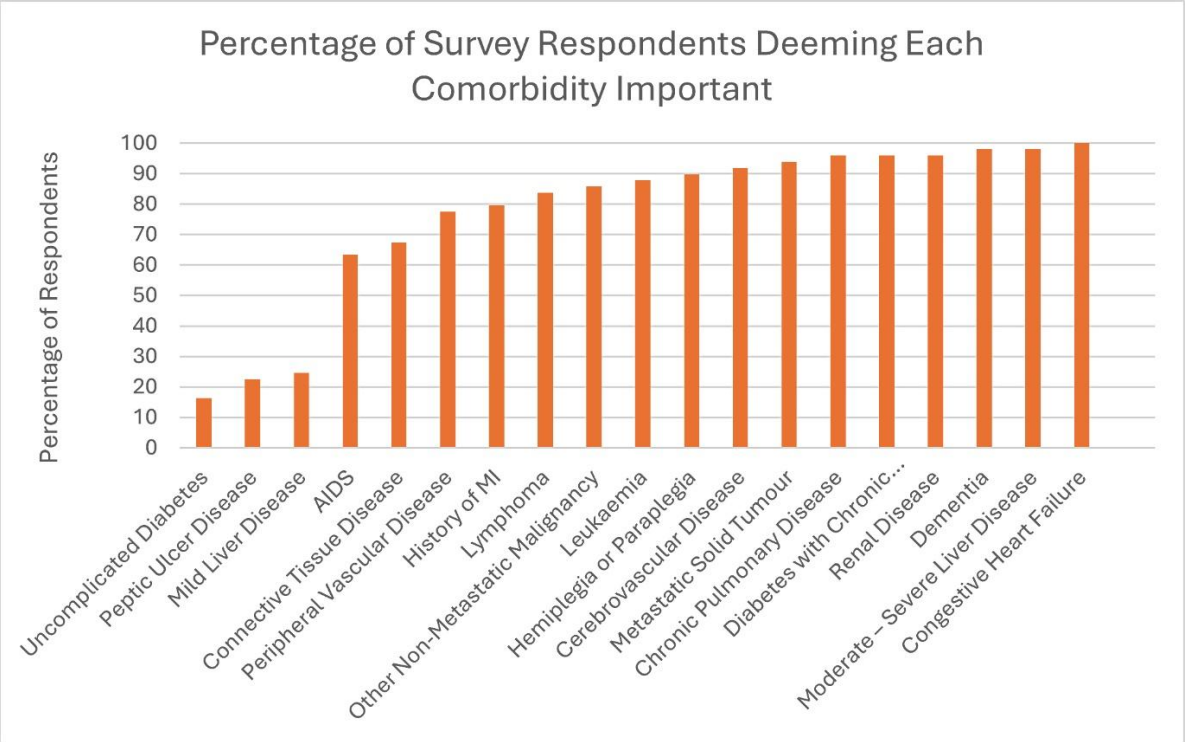
Rank	Survey		Comparison ML model
1	cMetastasis stage		cMetastasis stage
2	Performance status		Performance status
3	cTumour stage		Age
4	Biosy / tumour histology		Epoch *(not in survey)*
5	Comorbidities		cNodal stage
6	cNodal stage		cTumour stage
7	Patient preference *(not in ML model)*		Referring location *(not in survey)*
8	Tumour location		Comorbidities *(summed)*
9	Age		Tumour location
10	ASA grade *(not in ML model)*		Tumour histology

Metastasis stage and performance status were top ranking for both the respondents and ML model. However, while age ranked third in the model (with high relative importance at 16%) it ranked 9th in priority within the survey. Tumour stage and histology both ranked more highly for respondents than the model.

Gender was not ranked in the top ten by any survey respondent but has a 1.13% variable importance in the ML model, higher than any individual comorbidity. Patient preference and ASA grade featured in the top ten survey rankings but are not featured within the current ML model - this data has not been introduced, and in the case of patient preference frequently unavailable at first MDT discussion.

F.5.4 Comorbidities

Respondents were asked which co-morbidities they deemed to be important and were free to choose as many or few as they wished (Supplemental Figure 17). Congestive heart failure and dementia were considered important by more than 97% of survey respondents but ranked relatively low within the preliminary ML model at 0.2% and 0.04% respectively. Amongst comorbidities, uncomplicated diabetes was considered the least important to respondents but proved third most important in the model.



Supplemental Figure 17 - Survey Data - percentage of respondents who considered each comorbidity to be important when able to select as many as they wished (n = 49)

F.5.5 Additional factors recommended by respondents

Respondents were given the opportunity to highlight any additional factors they felt were important to their decision-making, using free text (Supplemental Table 20). They were organised into 4 categories through thematic analysis – additional investigation findings and assessments, symptoms, medical history, nutrition and social factors.

Supplemental Table 20 - Additional factors identified through the survey as important in making OC treatment decisions. Underlined factors were identified by some respondents as also regularly discussed at their MDT (n = 22).

Category	Factors
Additional investigation findings and assessments	Molecular markers (e.g. PD-L1, CPS, HER-2, MMR, MSI), DYPD genetic testing (determines how well certain chemotherapy agents are metabolised), tumour length, total length of disease (relevant to radiotherapy field), nodal distribution, stomach involvement, lung scan appearances, exercise testing, lung function test results, subjective impression of fitness ('end of bed' assessment), Anaesthetist's opinion, frailty
Symptoms and medical history	Dysphagia (swallowing difficulty), stridor (narrowed windpipe), vomiting (gastric outlet obstruction), disease-related quality of life, medication use, hearing impairment (risk of hearing loss as side-effect)
Nutrition	Weight / weight loss, body mass index, nutritional indicators
Social factors	Social, family and community support, employment, level of understanding and ability to report symptoms / side effects, likely compliance with treatment, psychological wellbeing and coping, religious beliefs (especially for Jehovah's Witnesses), profession and hobbies (risk of neuropathy (nerve damage) as side effect.

In addition, some respondents highlighted the impact of differing knowledgeability amongst MDT members, differing policies between units, the expertise of the surgical team in operating after chemoradiotherapy, and local research activity.

As neoadjuvant chemotherapy and chemoradiotherapy are currently both accepted treatments within the UK, respondents were also asked how they ordinarily chose between the two for a given patient. In addition to the general considerations previously outlined, several specific views emerged from across the range of professional roles (n=33) (Supplemental Table 21).

Supplemental Table 21 - Respondent views specific to the decision between chemotherapy and chemoradiotherapy. FLOT chemotherapy = 5-Flurouracil (5-FU), Leucovorin, Oxaloplatin, Docetaxel. CROSS Chemotherapy regimen = Carboplatin and Paclitaxel with concurrent radiotherapy.

Consideration	**Details From Some Respondents**
Histology – Squamous Cell Carcinoma or Adenocarcinoma?	Many reported using chemoradiotherapy for all squamous cell carcinomas (if size and position are appropriate), and defaulting to chemotherapy for most or all adenocarcinomas
Tumour Location	Chemotherapy for Squamous Cell Carcinoma if Siewert grade 2/3 (a grading system describing tumour position) *(Respondent 35, Clinical Oncologist)*
Tumour length / total disease length	Many raised the importance of whether a tumour is small enough to be encompassed in a radiotherapy field, <8-10cm.
Patient fitness for surgery	Some reported avoiding chemoradiotherapy if the patient is or may be fit for surgery
Is the circumferential resection margin threatened? I.e. is there a risk of incomplete removal with surgery?	May favour chemoradiotherapy for adenocarcinomas *(Respondent 48, Clinical Oncologist)*
Dysphagia (swallowing difficulties)	Chemotherapy may be more likely to improve this symptom *(Respondent 4, Clinical Oncologist)*
Dihydropyrimidine dehydrogenase (DPD) testing result (a genetic blood test determining ability to metabolise certain chemotherapy agents)	May favour chemoradiotherapy as cannot use 5FU chemotherapy *(Respondent 4, Clinical Oncologist)*
Patient frailty	Frail patients may find chemoradiotherapy more tolerable than FLOT chemotherapy *(Respondent 11, Surgeon)*
Molecular test results	If favour post-operative immunotherapy, may select CROSS chemoradiotherapy regimen *(Respondent 11, Surgeon)*
Impact of potential side effects	Pre-existing hearing impairment, hobbies or profession *(Respondent 2, Medical Oncologist)*

F.5.6 Clinician perceptions of a potential ML decision-support system for OC

When scoring their likelihood of using an ML CDSS from 0 (definitely not) to 100 (definitely yes), responses ranged from 0 to 100 (n=33) with a median of 75.0 (IQR 60.0-82.5)

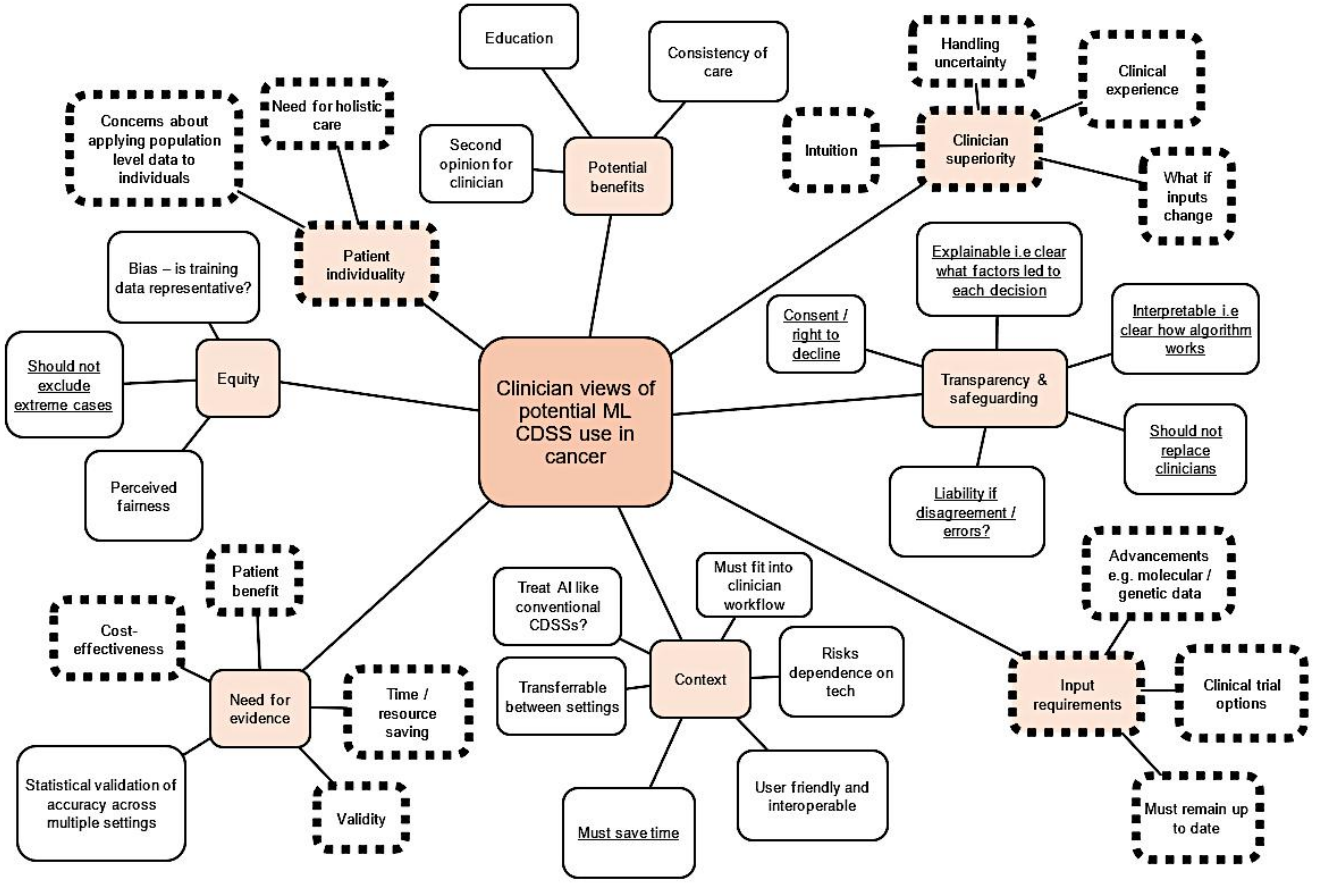
When asked about perceived barriers to ML CDSS use, five themes were identified (n=22) relating to: Clinician Superiority, Patient Individuality, Transparency and Safeguarding, a Need for Evidence, and Input Requirements (Supplemental Table 22).

Supplemental Table 22 - Thematic analysis of current barriers to adopting ML CDSSs in OC as highlighted by respondents.

Theme	Comments
Clinician Superiority	Some emphasised the importance of clinicians' intuition, 'gut feeling' and wealth of clinical experience, which they felt an algorithm could not emulate. Some considered clinicians better able to handle uncertainty when investigation results are inconclusive. *"Sometimes the true nodal stage cannot be accurately defined - are the nodes related to cancer or other lung pathology for instance. This uncertainty cannot be entered into an algorithm*." (Respondent 3, Surgeon) Several respondents emphasised that a CDSS should be limited to decision support and not over-ride clinical judgement.
Patient Individuality	Several respondents voiced concerns that ML CDSS use may hinder the individualised, holistic care that clinicians aim to provide, and questioned the application of population-level data to individual decisions. Some emphasised the multitude of factors involved, along with concerns that the model decision may not be valid if the patient's history, tumour or circumstances fell outside of the norm. *"I would be wary of using such a tool as I believe each patient is individual and unless you meet them and use a holistic approach it is hard to decide on the right treatment for them. People often differ from their 'on paper' selves. I do not believe AI can do this."* (Respondent 34, Specialist Nurse)
Transparency and Safeguarding	Respondents emphasised the importance of the model being open and explainable, so the reasoning behind its decision for each patient could be scrutinised to build trust. Some felt that patients should be informed of CDSS use with the right to decline. A number raised the need for a safeguarding process if clinicians disagree with the CDSS.

Theme	Comments
Need for evidence	Many respondents stated that they would need evidence of benefit before using an ML CDSS, including validity, cost effectiveness, patient benefit and time and resource savings. *"We discuss 40 patients in 90 minutes. It would have to be very quick and very useful!"* (Respondent 8, Surgeon) One respondent suggested that a CDSS may only save time if it were able to identify a group of patients who did not need to be discussed in the MDT.
Input requirements	Alongside the additional factors highlighted in Supplemental Table 21, some respondents felt that a CDSS should include information of available clinical trial options. Some questioned how the CDSS would stay up to date, suggesting for example that models trained on data before newer diagnostic and treatment advancements may no longer be valid. *"Because this is an area of rapid flux, I would be concerned that ML models built using data derived before the routine use of molecular testing and the availability of immunotherapy may not be able to keep pace with changes in clinical management."* (Respondent 11, Surgeon)

A conceptual framework created from a literature search of clinician perceptions of AI in cancer (described in Appendix F.4.3), is combined with the findings of this study (Supplemental Figure 18).



Supplemental Figure 18 - Conceptual framework of clinician views of potential CDSS use in cancer, updated to include the findings of this study. Dotted borders indicate content added from this work. Underlined indicates content from the literature review which was reinforced by the results of this study

F.6 Discussion

F.6.1 Summary

This study identified clinical and sociodemographic factors important in OC treatment decision-making, comparing their relative importance as identified by human experts against an ML model trained on historical MDT decisions from a leading tertiary referral unit. The study also explored clinicians' perceptions of using an ML-driven CDSSs within OC.

Specific factors of importance to the human experts included additional investigation findings and assessments (e.g. molecular markers and frailty), symptoms and medical history (e.g. swallowing difficulties and medication use), nutrition (e.g. body mass index (BMI)) and social factors (e.g. support and coping). Discordance between human and machine were apparent however in other areas such as demographic factors, where age and gender were valued much higher within the model than they were consciously to clinicians. Additionally certain co-morbidities ranking relatively low in importance in the ML model were weighted highly by participants (congestive heart failure and dementia), and vice versa (uncomplicated diabetes).

Clinicians' perceptions of ML CDSSs were favourable overall. However, several barriers to implementation were identified, grouped into five themes: perceptions of clinician superiority over ML CDSSs (e.g. human intuition), patient individuality (e.g. the need for person-centred care), transparency and safeguarding (e.g. explainability), the need for evidence (e.g. of effectiveness), and input requirements (e.g. molecular data).

F.6.2 Interpretation

This study has revealed that while ML can model OC decision-making with high performance, there remains some discordance between what humans consciously value within this process versus what the machine identifies when analysing historic decisions. Despite this, the overall sentiment towards the role of ML within MDT processes appears to be positive provided certain barriers can be overcome, such as the sense of clinician superiority, transparency, and granularity of the input variables.

F.6.3 Implications

Comorbidities such as congestive heart failure and dementia undoubtedly increase perioperative risk in surgical candidates (329,330). Their importance to human experts is expected, and their comparative insignificance to the model may be explained by the low overall incidence of these conditions within the cohort (2.4% for CHF and 1.1% for dementia). Caution is therefore needed when considering CDSS recommendations based on infrequent features within the training data.

The disparate ranking of age in this study was also a notable finding. UK guidelines for OC are not directed by age, as it is a poor predictor at an individual level in cancer treatment decision-making (331). However, the possibility that clinicians treat it as a surrogate marker of overall fitness for major therapy has been raised previously (109). It appears that age's disproportionate influence in practice may be sub-conscious, and independent of PS in planning cancer treatment - a finding echoed elsewhere in the literature (187). Gender, ethnicity, socioeconomic status and geographical location were only deemed important by a minority of clinicians. However previous studies have found inequitable treatments for women, non-caucasians and those of lower socioeconomic backgrounds (187–189,326). ML CDSSs learning from historical decisions risk perpetuating inequalities, something the European Union Artificial Intelligence Act looks to mitigate against (332,333). A key benefit of ML within cancer care decision-making is not just the modelling and automation of treatment planning but also reversing the technology to examine and interrogate team-based decisions. This offers the capability to audit decisions over time and elicit potential sources of bias. Within this cohort, our model weighed Age as a significant factor and we have shown previously that age for example significantly influences OC treatment options (108,109). There is some nuance to this phenomenon however as we must balance the drive to minimise bias from ML models with the recognition that not all biases are indeed problematic and may even represent domain expertise borne form experiential learning. In these situations, it is recognised that some biases may be useful provided they do not propagate inequity (227). In our use-case we know that advanced age carries additional clinical risk in aggressive medical interventions and so must be considered when treatment planning. Once such biases are recognised this may then allow future training datasets to be modified to up-scale representation in under-represented groups or be fed-back to clinicians directly to modify human decision-making prior to training subsequent model generations.

Desirable factors outside of the current model identified by respondents included exercise testing, BMI, presenting symptoms, genetic testing and novel molecular markers. Patient preference was identified as a key factor but could not be included within the current ML model as it is trained on MDT discussions made early in the patient pathway. An estimated 11% of treatment decisions deviate from clinicians' recommendations due to patient preference (325), and some have argued that preferences should be more formally investigated and recorded for MDTs (325,334). Shared decision-making between patient and clinician may benefit from future CDSSs during consultations to support this process (325).

Participant perceptions of ML CDSSs in OC conveyed previously identified preferences for explainability, interpretability (335–339), a safeguarding process in the event of clinician disagreement with the CDSS (340,341), the need for speed (325), and the importance of reliability for extreme clinical cases (340). Building trust between clinicians and AI tools is paramount if adoption of such tools is to be achieved and requires ethical, transparent innovation (55). In addition to quality control, data governance and bias mitigation, explainability (the ability to extract insight and understanding of machine logic when provided a given prediction or AI output) has become key for Medical AI in recent years (59). To avoid biasing respondents in their survey responses the feature importance values were not provided within the survey. However, in the context of a validated ML model being used in clinical scenarios, the authors advocate for the insights from explainability tools to be made readily available for clinicians using medical CDSSs. This allows clinicians the opportunity to evaluate how well such rationales align with their own human decision making to develop trust in recommendations or to exercise clinician agency in disagreeing with the output. Within our modelling we utilised a form of global explainability technique (permutation-based feature importance) to provide insight. Global techniques provide a static over-arching overview of how the model uses and weighs inputs and this does not change from instance to instance. However, at the point of use, local-explainability tools such as Local Interpretable Model Agnostic Explanations (LIME (209) and Shappley Values (SHAP (210)) are also necessary in providing insight at the individual patient level. In combining these techniques and presenting them to clinicians as built-in aspects of user interfaces maximises transparency. This may then be combined with features such as counterfactual explanations (explanations derived from changing maximally one or two model inputs for a given patient and comparing how the resultant output varies from the original scenario), clinician override (the ability to override an AI recommendation by the human operator and record those specific cases for future training)

and feedback systems for clinicians to highlight areas of discrepancy to incorporate into subsequent model iterations.

Areas of hesitancy for respondents surrounded factoring in patient individuality, perceived clinician superiority, and novel clinical predictors. In a survey of doctors and medical students in Korea, only 44% believed that AI was diagnostically superior to doctors (255). Research has also highlighted the limitation of AI CDSS in handling uncertainty (267), a sentiment strongly echoed by participants within this study. With regards to patient individuality, it remains controversial whether medical AI inhibits or enhances person-centred care. Some fear MAI may fail to incorporate patients' values and preferences, while proponents argue for its potential to release clinician time to build patient trust, aiding counselling for shared decision-making (342). This demonstrates that in the short term, such tools will play a primarily decision-support role rather than an automated decision-maker role.

The results of this study have highlighted that while MAI represents a growth market globally, offering potential gains across several clinical performance indicators, successful implementation requires buy-in by end-users through overcoming the barriers highlighted here. It has also demonstrated discordance between human perception and machine learned variables for OC decision making, suggesting that what the human agents within MDTs perceive they value may not necessarily match subsequent outcomes. Human oversight remains key to mitigating MAI bias (320), and research such as this study, is necessary to achieve long-term clinician buy-in, as well as safe and equitable implementation.

F.6.4 Study limitations and strengths

The study was undoubtedly subject to limitations, the main one being a low response rate (approximately 6% - 8% based on society mailing list sizes) - a common challenge with clinician surveys (343), as well as response-attrition over the course of the survey. Only the most motivated respondents continued onto the latter stages of the questionnaire exploring perceptions of ML CDSSs, potentially exacerbating response bias as the latter stages of the survey were also where perceptions towards ML were dealt with. That our results found a generally positive sentiment towards ML may thus be in part, owing to self-selection of respondents who were already supportive of ML. We feel however that the national reach of the

survey will have helped mitigate this to some degree but nevertheless recognise it is a potential confounder.

Exploring nuanced decision-making within an online survey naturally risks over-simplifying a complex process where factors may not carry equal weight across their range and may even act synergistically with others. Nevertheless, this study leveraged a mixed methodology approach to bridge knowledge between ML modelling of MDT decision-making with quantitative and qualitative data on respondent's priorities (344) allowing direct comparison between real world practice as viewed through an AI paradigm versus clinicians' own conscious decision-making. The anonymous, online nature of the survey mitigates some of the risk of social acceptability response bias, particularly compared to co-design approaches in which interviewers may be affiliated to the CDSS. This survey also achieved strong geographical coverage across the UK, spanning all relevant professional groups within the MDT and representing a full range of opinions on using ML CDSSs. The insights derived from the study will be key in guiding future CDSS design and ML modelling approaches if designers wish to maximise the adoption of these CDSSs into widespread clinical practice. The need for a "co-design" ethos within medical AI which factors in stakeholder needs and concerns is increasingly important if the barriers identified within this study are to be overcome.

F.7 Conclusion

This survey revealed insights into clinicians' priorities in OC decision-making and highlighted the range and complexity of factors influencing treatment recommendations. Sociodemographic factors appear to have greater impact in practice than would be anticipated by their conscious importance to clinicians. A wide range of potential barriers to the future use of an ML CDSS in OC were identified, many of which aligned with the findings of work on other cancers and throughout medicine. These must be addressed if the potential gains of clinical AI CDSS are to be realised.

References

1. Heartburn Cancer UK. Oesophageal Cancer [Internet]. 2020 [cited 2024 Mar 18]. Available from: https://heartburncanceruk.org/oesophageal-cancer/

2. Cromwell D, Palser T, van der Meulen J, Hardwick R, Riley S, Greenaway K, et al. An audit of the care received by people with Oesophago-Gastric Cancer in England and Wales, National Oesophago-gastric Cancer Audit - 2010. 2010. Report.

3. Chadwick G, Fellow R, Groene O, Cromwell D, Greenaway K. An audit of the care received by people with Oesophago-Gastric Cancer in England and Wales National Oesophago-Gastric Cancer Audit 2013. 2013. Report.

4. Arnold M, Ferlay J, Van Berge Henegouwen MI, Soerjomataram I. Global burden of oesophageal and gastric cancer by histology and subsite in 2018. Gut. 2020 Sep 1;69(9):1564–71. doi:10.1136/gutjnl-2020-321600 PubMed PMID: 32606208.

5. Reynolds J V., 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). Journal of Clinical Oncology. 2021 May 20;39(15_suppl):4004–4004. doi:10.1200/JCO.2021.39.15_suppl.4004

6. 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 (D. n engl j med [Internet]. 2006. Report. Available from: www.nejm.org

7. Allum WH, Stenning SP, Bancewicz J, Clark PI, Langley RE. Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in esophageal cancer. Journal of Clinical Oncology. 2009 Oct 20;27(30):5062–7. doi:10.1200/JCO.2009.22.2083 PubMed PMID: 19770374.

8. 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 Sep 1;16(9):1090–8. doi:10.1016/S1470-2045(15)00040-6 PubMed PMID: 26254683.

9. 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. The Lancet. 2019 May 11;393(10184):1948–57. doi:10.1016/S0140-6736(18)32557-1 PubMed PMID: 30982686.

10. Reynolds J V, Preston SR, O'Neill B, Lowery MA, Baeksgaard L, Crosby T, et al. Trimodality therapy versus perioperative chemotherapy in the management of locally advanced adenocarcinoma of the oesophagus and oesophagogastric junction (Neo-AEGIS): an open-label, randomised, phase 3 trial. Lancet Gastroenterol Hepatol. 2023 Sep. doi:10.1016/s2468-1253(23)00243-1

11. Hoeppner J, Brunner T, Lordick F, Schmoor C, Kulemann B, Neumann UP, et al. Prospective randomized multicenter phase III trial comparing perioperative chemotherapy (FLOT protocol) to neoadjuvant chemoradiation (CROSS protocol) in patients with adenocarcinoma of the esophagus (ESOPEC trial). Journal of Clinical Oncology. 2024 Jun 10;42(17_suppl):LBA1–LBA1. doi:10.1200/JCO.2024.42.17_suppl.LBA1

12. Noble F, Lloyd MA, Turkington R, Griffiths E, O'Donovan M, O'Neill JR, et al. Multicentre cohort study to define and validate pathological assessment of response to neoadjuvant therapy in oesophagogastric adenocarcinoma. British Journal of Surgery. 2017 Dec 1;104(13):1816–28. doi:10.1002/bjs.10627 PubMed PMID: 28944954.

13. Jiang W, de Jong JM, van Hillegersberg R, Read M. Predicting Response to Neoadjuvant Therapy in Oesophageal Adenocarcinoma. Cancers. MDPI; 2022. doi:10.3390/cancers14040996

14. 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 May 27;17(1). doi:10.1186/s12957-019-1630-8 PubMed PMID: 31133018.

15. Findlay JM, Bradley KM, Wang LM, Franklin JM, Teoh EJ, Gleeson F V., et al. Predicting pathologic response of esophageal cancer to neoadjuvant chemotherapy: The implications of metabolic nodal response for personalized therapy. Journal of Nuclear Medicine. 2017 Feb 1;58(2):266–75. doi:10.2967/jnumed.116.176313 PubMed PMID: 27635027.

16. Goense L, van Rossum PSN, Xi M, Maru DM, Carter BW, Meijer GJ, et al. Preoperative Nomogram to Risk Stratify Patients for the Benefit of Trimodality Therapy in Esophageal

Adenocarcinoma. Ann Surg Oncol. 2018 Jun 1;25(6):1598–607. doi:10.1245/s10434-018-6435-4 PubMed PMID: 29569125.

17. Bott RK, George G, McEwen R, Zylstra J, Knight WRC, Baker CR, et al. Predicting response to neoadjuvant chemotherapy in patients with oesophageal adenocarcinoma. Acta Oncol (Madr). 2021;60(12):1629–36. doi:10.1080/0284186X.2021.1986228 PubMed PMID: 34613874.

18. Al-Batran SE, Ajani JA. Impact of chemotherapy on quality of life in patients with metastatic esophagogastric cancer. Cancer. 2010. p. 2511–8. doi:10.1002/cncr.25064 PubMed PMID: 20301114.

19. 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. Annals of Thoracic Surgery. 2011 Oct;92(4):1239–43. doi:10.1016/j.athoracsur.2011.05.057 PubMed PMID: 21867990.

20. Publishing Asia B, Stephens MR, Lewis WG, Brewster AE, Lord I, J C Blackshaw GR, et al. Multidisciplinary team management is associated with improved outcomes after surgery for esophageal cancer. Diseases of the Esophagus [Internet]. 2006. Report. Available from: https://academic.oup.com/dote/article/19/3/164/2194890

21. 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 Apr;18(2):214–9. doi:10.1007/s10147-011-0362-8 PubMed PMID: 22193638.

22. NHS England. B11/S/a 2013/14 NHS STANDARD CONTRACT FOR CANCER: OESOPHAGEAL AND GASTRIC (ADULT). 2013. Report.

23. Lamb BW, Brown KF, Nagpal K, Vincent C, Green JSA, Sevdalis N. Quality of care management decisions by multidisciplinary cancer teams: A systematic review. Annals of Surgical Oncology. 2011. p. 2116–25. doi:10.1245/s10434-011-1675-6 PubMed PMID: 21442345.

24. 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(9):1091–9. doi:10.1080/0284186X.2021.1937308 PubMed PMID: 34313177.

25. Gray R, Gordon B, Meredith M. Meeting patients' needs: Improving the effectiveness of multidisciplinary team meetings in cancer services [Internet]. 2017. Report. Available from: www.cancerresearchuk.org

26. Gray R, Gordon B, Meredith M. CRUK: Meeting patients' Needs - Improving the effectiveness of multidisciplinary team meetings in cancer services [Internet]. 2017. Report. Available from: www.cancerresearchuk.org

27. Munro AJ. Multidisciplinary Team Meetings in Cancer Care: An Idea Whose Time has Gone? Clin Oncol (R Coll Radiol). 2015 Dec;27(12):728–31. doi:10.1016/j.clon.2015.08.008 PubMed PMID: 26365047.

28. NHS England. NHS England [Internet]. 2024 [cited 2025 May 4]. NHS National Cost Collection Publication: National Schedule 2023/24. Available from: https://app.powerbi.com/view?r=eyJrIjoiZGQxYjNkOGUtOTIwMC00N2VjLWEyM2EtYjAzOGMwNWU5ODQ1IiwidCI6IjM3YzM1NGIyLTg1YjAtNDdmNS1iMjIyLTA3YjQ4ZDc3NGVlMyJ9

29. Knight WRC, Baker CR, Griffin N, Wulaningsih W, Kelly M, Davies AR, et al. Does a high Mandard score really define a poor response to chemotherapy in oesophageal adenocarcinoma? Br J Cancer. 2021 May 11;124(10):1653–60. doi:10.1038/s41416-021-01290-4 PubMed PMID: 33742143.

30. Moore JL, Green M, Santaolalla A, Deere H, Evans RPT, Elshafie M, et al. Pathologic Lymph Node Regression After Neoadjuvant Chemotherapy Predicts Recurrence and Survival in Esophageal Adenocarcinoma: A Multicenter Study in the United Kingdom. Journal of Clinical Oncology. 2023 Oct 1;41(28):4522–34. doi:10.1200/JCO.23.00139

31. Blazeby JM. Minimally invasive or open oesophagectomy for localized oesophageal cancer: Results of the ROMIO phase 3 randomized controlled trial. Journal of Clinical Oncology. 2021 May 20;39(15_suppl):e16057–e16057. doi:10.1200/JCO.2021.39.15_suppl.e16057

32. National Oesophagogastric Cancer Audit. National Oesophago-Gastric Cancer Audit 2016. 2016. Report.

33. Michalowski J, Salvador A, Napper R. Commissioned by Healthcare Quality Improvement Partnership National Oesophago-Gastric Cancer Audit 2018 An audit of the care received by people with Oesophago-Gastric Cancer in England and Wales 2018 Annual Report. 2018. Report.

34. Wahedally H, Data Manager S/, Cromwell D. National Oesophago-Gastric Cancer Audit 2021 An audit of the care received by people with Oesophago-Gastric Cancer in England and Wales Commissioned by the Healthcare Quality Improvement Partnership. 2021. Report.

35. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May 23;j2099. doi:10.1136/bmj.j2099

36. Vaughan TL, Onstad L, Dai JY. Interactive decision support for esophageal adenocarcinoma screening and surveillance. BMC Gastroenterol. 2019 Jun 27;19(1). doi:10.1186/s12876-019-1022-0 PubMed PMID: 31248371.

37. Collins GS, Altman DG. Identifying patients with undetected gastro-oesophageal cancer in primary care: External validation of QCancer® (Gastro-Oesophageal). Eur J Cancer. 2013 Mar;49(5):1040–8. doi:10.1016/j.ejca.2012.10.023 PubMed PMID: 23159533.

38. Hendriks MP, Jager A, Ebben KCWJ, van Til JA, Siesling S. Clinical decision support systems for multidisciplinary team decision-making in patients with solid cancer: Composition of an implementation model based on a scoping review. Critical Reviews in Oncology/Hematology. Elsevier Ireland Ltd; 2024. doi:10.1016/j.critrevonc.2024.104267 PubMed PMID: 38311011.

39. Vu E, Steinmann N, Schröder C, Förster R, Aebersold DM, Eychmüller S, et al. Applications of Machine Learning in Palliative Care: A Systematic Review. Cancers. MDPI; 2023. doi:10.3390/cancers15051596

40. Beauchemin M, Murray MT, Sung L, Hershman DL, Weng C, Schnall R. Clinical decision support for therapeutic decision-making in cancer: A systematic review. Int J Med Inform. 2019 Oct 1;130. doi:10.1016/j.ijmedinf.2019.07.019 PubMed PMID: 31450082.

41. 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. 2022 Mar 9;126(4):562–8. doi:10.1038/s41416-021-01506-7 PubMed PMID: 34471257.

42. 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 Jun 1;10(1):619–33. doi:10.1007/s40122-021-00251-2

43. Diller GP, Kempny A, Babu-Narayan S V, 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 Apr 1;40(13):1069–77. doi:10.1093/eurheartj/ehy915 PubMed PMID: 30689812.

44. 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 Dec 1;16(1). doi:10.1186/s12885-016-2972-z PubMed PMID: 27905893.

45. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: Systematic literature review. JMIR Medical Informatics. JMIR Publications Inc.; 2020. doi:10.2196/18599

46. Hunter Emergency Laparotomy Collaborator Group, Hunter Emergency Laparotomy Collaborator Group. High-Risk Emergency Laparotomy in Australia: Comparing NELA, P-POSSUM, and ACS-NSQIP Calculators. J Surg Res. 2020 Feb;246:300–4. doi:10.1016/j.jss.2019.09.024 PubMed PMID: 31648068.

47. Mak M, Hakeem AR, Chitre V. Pre-NELA vs NELA - has anything changed, or is it just an audit exercise? Ann R Coll Surg Engl. 2016 Nov;98(8):554–9. doi:10.1308/rcsann.2016.0248 PubMed PMID: 27502336.

48. IBM. IBM Cloud Education [Internet]. 2020 [cited 2024 Mar 18]. Deep Learning. Available from: https://www.ibm.com/cloud/learn/deep-learning#toc-deep-learn-md_Q_Of3

49. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. Medicina (Lithuania). 2020 Sep 1;56(9):1–10. doi:10.3390/medicina56090455 PubMed PMID: 32911665.

50. Bradley A, van der Meer R, McKay C. Personalized Pancreatic Cancer Management: A Systematic Review of How Machine Learning Is Supporting Decision-making. Pancreas. 2019 May 1;48(5):598–604. doi:10.1097/MPA.0000000000001312 PubMed PMID: 31090660.

51. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001. Report.

52. Geeks for Geeks. geeksforgeeks.org [Internet]. 2025 [cited 2025 Mar 3]. Difference Between Machine Learning and Statistics. Available from: https://www.geeksforgeeks.org/difference-between-machine-learning-vs-statistics/

53. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. Yearb Med Inform. 2019 Aug 1;28(1):128–34. doi:10.1055/s-0039-1677903 PubMed PMID: 31022752.

54. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan 1;28(1):31–8. doi:10.1038/s41591-021-01614-0 PubMed PMID: 35058619.

55. Upadhyay U, Gradisek A, Iqbal U, Dhar E, Li YC, Syed-Abdul S. Call for the responsible artificial intelligence in the healthcare. BMJ Health Care Inform. 2023 Dec 21;30(1). doi:10.1136/bmjhci-2023-100920 PubMed PMID: 38135293.

56. European Commision. The EU Artificial Intelligence Act. Official Journal (OJ) of the European Union. 2024 Jul 12.

57. KPMG international. Decoding the EU AI Act. 2024. Report.

58. Lashbrook A. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. The Atlantic [Internet]. 2018 Aug 16 [cited 2024 Jul 22]. Available from: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

59. Zhang J, Zhang Z ming. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak. 2023 Dec 1;23(1). doi:10.1186/s12911-023-02103-9 PubMed PMID: 36639799.

60. Portillo V, Greenhalgh C, Craigon PJ, Ten Holter C. Responsible Research and Innovation (RRI) Prompts and Practice Cards: A Tool to Support Responsible Practice. In: ACM International Conference Proceeding Series. Association for Computing Machinery; 2023. doi:10.1145/3597512.3599721

61. Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. Res Policy. 2013;42(9):1568–80. doi:10.1016/j.respol.2013.05.008

62. Jirotka M, Grimpe B, Stahl B, Eden G, Hartswood M. Responsible research and innovation in the digital age. Commun ACM. 2017 May 1;60(5):62–8. doi:10.1145/3064940

63. Thavanesan N, Vigneswaran G, Bodala I, Underwood TJ. The Oesophageal Cancer Multidisciplinary Team: Can Machine Learning Assist Decision-Making? Journal of Gastrointestinal Surgery. 2023. doi:10.1007/s11605-022-05575-8

64. Calman K, Hine D. A policy framework for commissioning cancer services. 1995. Report.

65. Vermeulen BD, Bruggeman L, Bac DJ, Schrauwen RWM, Epping LSM, Scheffer RCH, et al. Impact of multidisciplinary tumor board discussion on palliation of patients with esophageal or gastro-esophageal junction cancer: a population-based study. Acta Oncol (Madr). 2020 Apr 2;59(4):410–6. doi:10.1080/0284186X.2020.1725240 PubMed PMID: 32067535.

66. 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 Sep;35(9):1970–6. doi:10.1007/s00268-011-1152-1 PubMed PMID: 21604049.

67. Haward R, Amir Z, Borrill C, Dawson J, Scully J, West M, et al. Breast cancer teams: the impact of constitution, new cancer workload, and methods of operation on their effectiveness. Br J Cancer. 2003 Jul 7;89(1):15–22. doi:10.1038/sj.bjc.6601073 PubMed PMID: 12838294.

68. Hamilton DW, Heaven B, Thomson RG, Wilson JA, Exley C. Multidisciplinary team decision-making in cancer and the absent patient : a qualitative study. 2016;1–8. doi:10.1136/bmjopen-2016-012559

69. 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(9):1091–9. doi:10.1080/0284186X.2021.1937308 PubMed PMID: 34313177.

70. The National Institute for Health and Care Excellence (NICE). Oesophago-gastric cancer: Assessment and management in adults (NG83). NICE Guideline. 2018;4(January 2018):970–6.

71. Smyth EC, Lagergren J, Fitzgerald RC, Lordick F, Shah MA, Lagergren P, et al. Oesophageal cancer. Nat Rev Dis Primers. 2017 Jul 27;3:17048. doi:10.1038/nrdp.2017.48 PubMed PMID: 28748917.

72. Lang CCJ, Lloyd M, Alyacoubi S, Rahman S, Pickering O, Underwood T, et al. The Use of miRNAs in Predicting Response to Neoadjuvant Therapy in Oesophageal Cancer. Cancers (Basel). 2022;14(5). doi:10.3390/cancers14051171

73. Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA, et al. The AUGIS Survival Predictor: Prediction of Long-Term and Conditional Survival After Esophagectomy Using Random Survival Forests. Ann Surg. 2023 Feb 1;277(2):267–74. doi:10.1097/SLA.0000000000004794 PubMed PMID: 33630434.

74. Goense L, van Rossum PSN, Xi M, Maru DM, Carter BW, Meijer GJ, et al. Preoperative Nomogram to Risk Stratify Patients for the Benefit of Trimodality Therapy in Esophageal Adenocarcinoma. Ann Surg Oncol. 2018;25(6):1598–607. doi:10.1245/s10434-018-6435-4 PubMed PMID: 29569125.

75. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: A practical guide for beginners. Diagnostic and Interventional Radiology. 2019;25(6):485–95. doi:10.5152/dir.2019.19321 PubMed PMID: 31650960.

76. Dimitriou N, Arandjelović O, Caie PD. Deep Learning for Whole Slide Image Analysis: An Overview. Front Med (Lausanne). 2019;6(November):1–7. doi:10.3389/fmed.2019.00264

77. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological Image Analysis: A Review. IEEE Rev Biomed Eng. 2009;2:147–71. doi:10.1109/RBME.2009.2034865 PubMed PMID: 20671804.

78. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J. 2018;16:34–42. doi:10.1016/j.csbj.2018.01.001

79. Rahman S, Early J, De Vries M, Lloyd M, Grace B, Ramchurn G, et al. Predicting response to neoadjuvant therapy using image capture from diagnostic biopsies of oesophageal adenocarcinoma. European Journal of Surgical Oncology. 2021 Jan;47(1):e4. doi:10.1016/j.ejso.2020.11.022

80. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides. JAMA Netw Open. 2019 Nov 11;2(11). doi:10.1001/jamanetworkopen.2019.14645 PubMed PMID: 31693124.

81. Kieffer B, Babaie M, Kalra S, Tizhoosh HR. Convolutional Neural Networks for Histopathology Image Classification : Training vs . Using Pre-Trained Networks. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). 2017. p. 1–6. doi:10.1109/IPTA.2017.8310149

82. Bogowicz M, Vuong D, Huellner MW, Pavic M, Andratschke N, Gabrys HS, et al. CT radiomics and PET radiomics: Ready for clinical implementation? Quarterly Journal of Nuclear Medicine and Molecular Imaging. 2019;63(4):355–70. doi:10.23736/S1824-4785.19.03192-3 PubMed PMID: 31527578.

83. Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Texture analysis of imaging: What radiologists need to know. American Journal of Roentgenology. American Roentgen Ray Society; 2019. p. 520–8. doi:10.2214/AJR.18.20624 PubMed PMID: 30645163.

84. Daiko H, Kato K. Updates in the 8th edition of the TNM staging system for esophagus and esophagogastric junction cancer. Jpn J Clin Oncol. 2020 Aug 4;50(8):847–51. doi:10.1093/jjco/hyaa082 PubMed PMID: 32519741.

85. Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Radiologists Need to Know. Ajr. 2019;(212):1–9.

86. Xie C yi, Pang C lap, Chan B, Wong EY yuen, Dou Q, Vardhanabhuti V. Machine Learning and Radiomics Applications in Esophageal Cancers Using Non-Invasive Imaging Methods—A Critical Review of Literature. Cancers (Basel). 2021 May 19;13(10):2469. doi:10.3390/cancers13102469

87. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging. 2020;11(1). doi:10.1186/s13244-020-00887-2

88. Brunzell H, Eriksson J. Feature reduction for classification of multidimensional data. Pattern Recognit. 2000;33(10):1741–8. doi:10.1016/S0031-3203(99)00142-9

89. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26(3):303–4. doi:10.1038/nbt0308-303 PubMed PMID: 18327243.

90. Balakrishnama S, Ganapathiraju A. Linear Discriminant Analysis - A Brief Tutorial. Vol. 18. 1998. Report.

91. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17. doi:10.1093/bioinformatics/btm344 PubMed PMID: 17720704.

92. Rhys HI. Machine Learning with R, the tidyverse, and mlr. 1st Editio. New York: Manning; 2020. 511 p.

93. Findlay JM, Bradley KM, Wang LM, Franklin JM, Teoh EJ, Gleeson F V., et al. Predicting pathologic response of esophageal cancer to neoadjuvant chemotherapy: The implications of metabolic nodal response for personalized therapy. Journal of Nuclear Medicine. 2017;58(2):266–75. doi:10.2967/jnumed.116.176313 PubMed PMID: 27635027.

94. Beukinga RJ, Hulshoff JB, Van Dijk L V., Muijs CT, Burgerhof JGM, Kats-Ugurlu G, et al. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment 18F-FDG PET/CT imaging. Journal of Nuclear Medicine. 2017;58(5):723–9. doi:10.2967/jnumed.116.180299 PubMed PMID: 27738011.

95. Ou J, Li R, Zeng R, Wu CQ, Chen Y, Chen TW, et al. CT radiomic features for predicting resectability of oesophageal squamous cell carcinoma as given by feature analysis: A case control study. Cancer Imaging. 2019;19(1):1–10. doi:10.1186/s40644-019-0254-0 PubMed PMID: 31619297.

96. Hou Z, Ren W, Li S, Liu J, Sun Y, Yan J, et al. Radiomic analysis in contrast-enhanced CT: Predict treatment response to chemoradiotherapy in esophageal carcinoma. Oncotarget. 2017;8(61):104444–54. doi:10.18632/oncotarget.22304

97. Larue RTHM, Klaassen R, Jochems A, Leijenaar RTH, Hulshof MCCM, Henegouwen MIVB, et al. Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer. Acta Oncol (Madr). 2018;57(11):1475–81. doi:10.1080/0284186X.2018.1486039

98. Tan X, Ma Z, Yan L, Ye W, Liu Z, Liang C. Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma. Eur Radiol. 2019 Jan;29(1):392–400. doi:10.1007/s00330-018-5581-1 PubMed PMID: 29922924.

99. Simoni N, Rossi G, Benetti G, Zuffante M, Micera R, Pavarana M, et al. F-FDG PET / CT Metrics Are Correlated to the Pathological Response in Esophageal Cancer Patients Treated With Induction Chemotherapy Followed by. Vol. 10. 2020;10(November):1–11. doi:10.3389/fonc.2020.599907

100. Pan LL, Gu P, Huang G, Xue HP, Wu SQ. Prognostic significance of SUV on PET/CT in patients with esophageal cancer: A systematic review and meta-analysis. Eur J Gastroenterol Hepatol. 2009;21(9):1008–15. doi:10.1097/MEG.0b013e328323d6fa PubMed PMID: 19352191.

101. Cao Q, Li Y, Li Z, An D, Li B, Lin Q. Development and validation of a radiomics signature on differentially expressed features of 18F-FDG PET to predict treatment response of concurrent chemoradiotherapy in thoracic esophagus squamous cell carcinoma. Radiotherapy and Oncology. 2020;146:9–15. doi:10.1016/j.radonc.2020.01.027 PubMed PMID: 32065875.

102. Zhang H, Tan S, Chen W, Kligerman S, Kim G, D'Souza WD, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal 18F-FDG PET features, clinical parameters, and demographics. Int J Radiat Oncol Biol Phys. 2014;88(1):195–203. doi:10.1016/j.ijrobp.2013.09.037 PubMed PMID: 24189128.

103. Qiu Q, Duan J, Deng H, Han Z, Gu J, Yue NJ, et al. Development and Validation of a Radiomics Nomogram Model for Predicting Postoperative Recurrence in Patients With Esophageal Squamous Cell Cancer Who Achieved pCR After Neoadjuvant Chemoradiotherapy Followed by Surgery. Front Oncol. 2020;10(August):1–10. doi:10.3389/fonc.2020.01398

104. Reynolds J V, 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). Journal of Clinical Oncology. 2021 May 20;39(15_suppl):4004. doi:10.1200/JCO.2021.39.15_suppl.4004

105. Yang CK, Yeh JCY, Yu WH, Chien LI, Lin KH, Huang WS, et al. Deep convolutional neural network-based positron emission tomography analysis predicts esophageal cancer outcome. J Clin Med. 2019;8(6):1–9. doi:10.3390/jcm8060844

106. Shen C, Liu Z, Wang Z, Guo J, Zhang H, Wang Y, et al. Building CT Radiomics Based Nomogram for Preoperative Esophageal Cancer Patients Lymph Node Metastasis Prediction. Transl Oncol. 2018;11(3):815–24. doi:10.1016/j.tranon.2018.04.005

107. Hoshino I, Yokota H, Ishige F, Iwatate Y, Takeshita N, Nagase H, et al. Radiogenomics predicts the expression of microRNA-1246 in the serum of esophageal cancer patients. Sci Rep. 2020;10(1):1–8. doi:10.1038/s41598-020-59500-7 PubMed PMID: 32054931.

108. Thavanesan N, Bodala I, Walters Z, Ramchurn S, Underwood TJ, Vigneswaran G. Machine learning to predict curative multidisciplinary team treatment decisions in oesophageal cancer. European Journal of Surgical Oncology. 2023 Jul;106986. doi:10.1016/j.ejso.2023.106986 PubMed PMID: 37463827.

109. Thavanesan N, Farahi A, Parfitt C, Belkhatir Z, Azim T, Vallejos EP, et al. Insights from explainable AI in oesophageal cancer team decisions. Comput Biol Med. 2024 Sep;180:108978. doi:10.1016/j.compbiomed.2024.108978

110. Cancer research UK. Cancer Mortality for common cancers [Internet]. 2022. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared#heading-Zero

111. 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 Nov;68(6):394–424. doi:10.3322/caac.21492 PubMed PMID: 30207593.

112. 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. Annals of Thoracic Surgery. 2011;92(4):1239–43. doi:10.1016/j.athoracsur.2011.05.057 PubMed PMID: 21867990.

113. 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):1–9. doi:10.1186/s12957-019-1630-8 PubMed PMID: 31133018.

114. Al-Batran SE, Ajani JA. Impact of chemotherapy on quality of life in patients with metastatic esophagogastric cancer. Cancer. 2010 Jun 1;116(11):2511–8. doi:10.1002/cncr.25064 PubMed PMID: 20301114.

115. 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. Diseases of the Esophagus. 2006;19(3):164–71. doi:10.1111/j.1442-2050.2006.00559.x PubMed PMID: 16722993.

116. 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(2):214–9. doi:10.1007/s10147-011-0362-8 PubMed PMID: 22193638.

117. 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(8):2116–25. doi:10.1245/s10434-011-1675-6 PubMed PMID: 21442345.

118. 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(9):1970–6. doi:10.1007/s00268-011-1152-1 PubMed PMID: 21604049.

119. Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA, et al. The AUGIS Survival Predictor. Ann Surg. 2021;Publish Ah(0). doi:10.1097/sla.0000000000004794

120. 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;13(Ml):6240–51. doi:10.21037/jtd-21-1107

121. 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(2):1–11. doi:10.2196/14122 PubMed PMID: 32130123.

122. National Institute for Health and Care Excellence. Oesophago-gastric cancer: assessment and management in adults NICE guideline [Internet]. 2018. Report. Available from: www.nice.org.uk/guidance/ng83

123. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. Springer; 2002.

124. Breiman L. Random Forests. Mach Learn. 2001;45:5–32. doi:10.1023/A:1010933404324

125. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–94. doi:10.1145/2939672.2939785

126. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. Routledge; 2017. doi:10.1201/9781315139470

127. 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 (D. n engl j med [Internet]. 2006. Report. Available from: www.nejm.org

128. 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 May 18;359(9319):1727–33. doi:10.1016/S0140-6736(02)08651-8 PubMed PMID: 12049861.

129. 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 Sep 1;16(9):1090–8. doi:10.1016/S1470-2045(15)00040-6 PubMed PMID: 26254683.

130. 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):1–9. doi:10.1186/s12913-021-06203-y PubMed PMID: 33676492.

131. National Cancer Equality Initiative/Pharmaceutical Oncology Initiative. The impact of patient age on clinical decision-making in oncology. 2012. Report.

132. Ahamat N. Access all ages: assessing the impact of age on access to surgical treatment. The Bulletin of the Royal College of Surgeons of England. 2012 Oct 1;94(9):300–300. doi:10.1308/147363512x13448516926748

133. 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 Nov 7;3(CSCW):1–24. doi:10.1145/3359206

134. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? 2017 Dec 28;(Ml):1–28.

135. Chen V, Li J, Kim JS, Plumb G, Talwalkar A. Interpretable Machine Learning. Queue. 2021;19(6):28–56. doi:10.1145/3511299

136. 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 Nov 19;9(11):e030907. doi:10.1136/bmjopen-2019-030907

137. 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 Aug 1;266(2):232–6. doi:10.1097/SLA.0000000000002171 PubMed PMID: 28187044.

138. 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 Apr 1;156(4):323–32. doi:10.1001/jamasurg.2020.7081 PubMed PMID: 33595631.

139. 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 May;151(5):667–73. doi:10.1016/j.surg.2011.12.006

140. 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 Cardiothorac Surg. 2011 Dec;40(6):1455–63; discussion 1463-4. doi:10.1016/j.ejcts.2011.01.086 PubMed PMID: 21514837.

141. 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. doi:10.1001/jamanetworkopen.2021.29228 PubMed PMID: 34724556.

142. Washington K, Watkins JR, Jay J, Jeyarajah DR. Oncologic resection in laparoscopic versus robotic transhiatal esophagectomy. Journal of the Society of Laparoendoscopic Surgeons. 2019 Apr 1;23(2). doi:10.4293/JSLS.2019.00017 PubMed PMID: 31148912.

143. 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):1–10. doi:10.1186/s12885-016-2972-z PubMed PMID: 27905893.

144. Diller GP, Kempny A, Babu-Narayan S V., 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(13):1069–77. doi:10.1093/eurheartj/ehy915 PubMed PMID: 30689812.

145. 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(1):619–33. doi:10.1007/s40122-021-00251-2

146. 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;(July):1–7. doi:10.1038/s41416-021-01506-7

147. Bolger JC, Donohoe CL, Lowery M, Reynolds J V. Advances in the curative management of oesophageal cancer. British Journal of Cancer. Springer Nature; 2022. p. 706–17. doi:10.1038/s41416-021-01485-9 PubMed PMID: 34675397.

148. Cancer Research UK. Cancer Research UK Oesophageal Cancer Survival [Internet]. 2023 [cited 2023 Oct 27]. Available from: https://www.cancerresearchuk.org/about-cancer/oesophageal-cancer/survival

149. NHS England, NHS Improvement. Streamlining Multi-Disciplinary Team Meetings Guidance for Cancer Alliances. 2019. Report.

150. Liao F, Yu S, Zhou Y, Feng B. A machine learning model predicting candidates for surgical treatment modality in patients with distant metastatic esophageal adenocarcinoma: A propensity score-matched analysis. Front Oncol. 2022 Jul 22;12. doi:10.3389/fonc.2022.862536

151. Ma X, Pierce E, Anand H, Aviles N, Kunk P, Alemazkoor N. Early prediction of response to palliative chemotherapy in patients with stage-IV gastric and esophageal cancer. BMC Cancer. 2023 Dec 1;23(1):910. doi:10.1186/s12885-023-11422-z PubMed PMID: 37759332.

152. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. Routledge; 2017. doi:10.1201/9781315139470

153. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Annals of Applied Statistics. 2008 Sep;2(3):841–60. doi:10.1214/08-AOAS169

154. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982 May 14;247(18):2543–6. PubMed PMID: 7069920.

155. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Weather Rev. 1950 Jan;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

156. Tasci E, Zhuge Y, Camphausen K, Krauze A V. Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. Cancers. MDPI; 2022. doi:10.3390/cancers14122897

157. Seraphin M, Silberstein PT, Cichon G. Trends in palliative care interventions for stage IV esophageal cancer: An analysis of the NCDB. Journal of Clinical Oncology. 2022 Jun 1;40(16_suppl):e24104–e24104. doi:10.1200/JCO.2022.40.16_suppl.e24104

158. Bleiberg H, Conroy T, Paillot B, Lacave AJ, Blijham G, Jacob JH, et al. Randomised phase II study of cisplatin and 5-fluorouracil (5-FU) versus cisplatin alone in advanced squamous cell oesophageal cancer. Eur J Cancer. 1997 Jul;33(8):1216–20. doi:10.1016/s0959-8049(97)00088-9 PubMed PMID: 9301445.

159. Ilson DH, Ajani J, Bhalla K, Forastiere A, Huang Y, Patel P, et al. Phase II trial of paclitaxel, fluorouracil, and cisplatin in patients with advanced carcinoma of the esophagus. J Clin Oncol. 1998 May;16(5):1826–34. doi:10.1200/JCO.1998.16.5.1826 PubMed PMID: 9586897.

160. Kojima T, Shah MA, Muro K, Francois E, Adenis A, Hsu CH, et al. Randomized Phase III KEYNOTE-181 Study of Pembrolizumab Versus Chemotherapy in Advanced Esophageal Cancer. J Clin Oncol. 2020 Dec 10;38(35):4138–48. doi:10.1200/JCO.20.01888 PubMed PMID: 33026938.

161. Jiang Y, Xie J, Han Z, Liu W, Xi S, Huang L, et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. Clinical Cancer Research. 2018 Nov 15;24(22):5574–84. doi:10.1158/1078-0432.CCR-18-0848 PubMed PMID: 30042208.

162. Que SJ, Chen QY, Zhong Q, Liu ZY, Wang J Bin, Lin JX, et al. Application of preoperative artificial neural network based on blood biomarkers and clinicopathological parameters for predicting long-term survival of patients with gastric cancer. World J Gastroenterol. 2019 Nov 21;25(43):6451–64. doi:10.3748/wjg.v25.i43.6451 PubMed PMID: 31798281.

163. Janjigian YY, Shitara K, Moehler M, Garrido M, Salman P, Shen L, et al. First-line nivolumab plus chemotherapy versus chemotherapy alone for advanced gastric, gastro-oesophageal junction, and oesophageal adenocarcinoma (CheckMate 649): a randomised, open-label,

phase 3 trial. The Lancet. 2021 Jul 3;398(10294):27–40. doi:10.1016/S0140-6736(21)00797-2 PubMed PMID: 34102137.

164. Janjigian YY, Kawazoe A, Bai Y, Xu J, Lonardi S, Metges JP, et al. Pembrolizumab plus trastuzumab and chemotherapy for HER2-positive gastric or gastro-oesophageal junction adenocarcinoma: interim analyses from the phase 3 KEYNOTE-811 randomised placebo-controlled trial. The Lancet. 2023. doi:10.1016/S0140-6736(23)02033-0 PubMed PMID: 37871604.

165. Shitara K, Lordick F, Bang YJ, Enzinger P, Ilson D, Shah MA, et al. Zolbetuximab plus mFOLFOX6 in patients with CLDN18.2-positive, HER2-negative, untreated, locally advanced unresectable or metastatic gastric or gastro-oesophageal junction adenocarcinoma (SPOTLIGHT): a multicentre, randomised, double-blind, phase 3 trial. The Lancet. 2023 May 20;401(10389):1655–68. doi:10.1016/S0140-6736(23)00620-7 PubMed PMID: 37068504.

166. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. New England Journal of Medicine. 2015 Jun 25;372(26):2509–20. doi:10.1056/nejmoa1500596 PubMed PMID: 26028255.

167. 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. Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery; 2019. doi:10.1145/3359206

168. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? [Internet]. 2017 Dec 28. Available from: http://arxiv.org/abs/1712.09923

169. Chen V, Li J, Kim JS, Plumb G, Talwalkar A. Interpretable Machine Learning. Queue. 2021 Dec 31;19(6):28–56. doi:10.1145/3511299

170. Lander S, Lander E, Gibson MK. Esophageal Cancer: Overview, Risk Factors, and Reasons for the Rise. Current Gastroenterology Reports. Springer; 2023. p. 275–9. doi:10.1007/s11894-023-00899-0 PubMed PMID: 37812328.

171. Soukup T, Lamb BW, Morbi A, Shah NJ, Bali A, Asher V, et al. Cancer multidisciplinary team meetings: Impact of logistical challenges on communication and decision-making. BJS Open. 2022 Aug 1;6(4). doi:10.1093/bjsopen/zrac093 PubMed PMID: 36029030.

172. Soukup T, Morbi A, Lamb BW, Gandamihardja TAK, Hogben K, Noyes K, et al. A measure of case complexity for streamlining workflow in multidisciplinary tumor boards: Mixed methods

development and early validation of the MeDiC tool. Cancer Med. 2020 Jul 1;9(14):5143–54. doi:10.1002/cam4.3026 PubMed PMID: 32476281.

173. Wihl J, Falini V, Borg S, Stahl O, Jiborn T, Ohlsson B, et al. Implementation of the measure of case discussion complexity to guide selection of prostate cancer patients for multidisciplinary team meetings. Cancer Med. 2023 Jul 1;12(14):15149–58. doi:10.1002/cam4.6189 PubMed PMID: 37255390.

174. Lamb B, Green JSA, Vincent C, Sevdalis N. Decision making in surgical oncology. Surgical Oncology. 2011. p. 163–8. doi:10.1016/j.suronc.2010.07.007 PubMed PMID: 20719499.

175. Ebben KCWJ, Sieswerda MS, Luiten EJT, Heijns JB, Van Der Pol CC, Bessems M, et al. Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting. JCO Clin Cancer Inform [Internet]. 2020. Report. Available from: https://doi.org/10.

176. Walker RC, Underwood TJ. Oesophageal cancer. Surgery (United Kingdom). Elsevier Ltd; 2017. p. 627–34. doi:10.1016/j.mpsur.2017.09.010

177. Lin HS, Watts JN, Peel NM, Hubbard RE. Frailty and post-operative outcomes in older surgical patients: A systematic review. BMC Geriatrics. BioMed Central Ltd.; 2016. doi:10.1186/s12877-016-0329-8 PubMed PMID: 27580947.

178. Cancer Research UK. Oesophageal cancer incidence statistics [Internet]. 2021 Oct [cited 2023 Apr 19]. Report. Available from: Oesophageal cancer incidence statistics

179. Cheng Z, Johar A, Gottlieb-Vedi E, Nilsson M, Lagergren J, Lagergren P. Impact of co-morbidity on reoperation or death within 90 days of surgery for oesophageal cancer. BJS Open. 2021 Jan 1;5(1). doi:10.1093/bjsopen/zraa035 PubMed PMID: 33609378.

180. Racz J, Dubois L, Katchky A, Wall W. Elective and emergency abdominal surgery in patients 90 years of age or older. Can J Surg. 2012 Oct;55(5):322–8. doi:10.1503/cjs.007611 PubMed PMID: 22992421.

181. McGillicuddy EA, Schuster KM, Davis KA, Longo WE. Factors predicting morbidity and mortality in emergency colorectal procedures in elderly patients. Arch Surg. 2009 Dec;144(12):1157–62. doi:10.1001/archsurg.2009.203 PubMed PMID: 20026835.

182. Belinda DS, Fausto C. Emergency Surgery for Colorectal Cancer in Patients Aged Over 90 Years: Review of the Recent Literature. Journal of Tumour. 2016;4(1).

183. Guthrie B, Barnett K, Mercer SW, Norbury M, Watt G, Wyke S. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012;380:37–43. doi:10.1016/S0140

184. Ahamat N. Access all ages: assessing the impact of age on access to surgical treatment. The Bulletin of the Royal College of Surgeons of England. 2012 Oct 1;94(9):300–300. doi:10.1308/147363512x13448516926748

185. Thavanesan N, Bodala I, Walters Z, Ramchurn S, Underwood TJ, Vigneswaran G. Machine learning to predict curative multidisciplinary team treatment decisions in oesophageal cancer. European Journal of Surgical Oncology. 2023 Jul;106986. doi:10.1016/j.ejso.2023.106986

186. McLoughlin JM, Lewis JM, Meredith KL. The impact of age on morbidity and mortality following esophagectomy for esophageal cancer. Cancer Control. 2013 Apr;20(2):144–50. doi:10.1177/107327481302000208 PubMed PMID: 23571705.

187. Henson KE, Fry A, Lyratzopoulos G, Peake M, Roberts KJ, McPhail S. Sociodemographic variation in the use of chemotherapy and radiotherapy in patients with stage IV lung, oesophageal, stomach and pancreatic cancer: evidence from population-based data in England during 2013-2014. Br J Cancer. 2018 May;118(10):1382–90. doi:10.1038/s41416-018-0028-7 PubMed PMID: 29743552.

188. Okereke IC, Westra J, Tyler D, Klimberg S, Jupiter D, Venkatesan R, et al. Disparities in esophageal cancer care based on race: a National Cancer Database analysis. Dis Esophagus. 2022 Jun 15;35(6). doi:10.1093/dote/doab083 PubMed PMID: 34918057.

189. Kalff MC, Dijksterhuis WPM, Wagner AD, Oertelt-Prigione S, Verhoeven RHA, Lemmens VEPP, et al. Sex differences in treatment allocation and survival of potentially curable gastroesophageal cancer: A population-based study. Eur J Cancer. 2023 Jul;187:114–23. doi:10.1016/j.ejca.2023.04.002 PubMed PMID: 37146505.

190. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J. 2021 Mar;8(1):40–8. doi:10.7861/fhj.2020-0233 PubMed PMID: 33791459.

191. Walker RC, Underwood TJ. Oesophageal cancer. Surgery (United Kingdom). Elsevier Ltd; 2017. p. 627–34. doi:10.1016/j.mpsur.2017.09.010

192. Favareto SL, Sousa CF, Pinto PJ, Ramos H, Chen MJ, Castro DG, et al. Clinical Prognostic Factors for Patients With Esophageal Cancer Treated With Definitive Chemoradiotherapy. Cureus. 2021 Oct 19. doi:10.7759/cureus.18894

193. Lamb B, Green JSA, Vincent C, Sevdalis N. Decision making in surgical oncology. Surgical Oncology. 2011. p. 163–8. doi:10.1016/j.suronc.2010.07.007 PubMed PMID: 20719499.

194. Farzaneh N, Ansari S, Lee E, Ward KR, Sjoding MW. Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. NPJ Digit Med. 2023 Dec 1;6(1). doi:10.1038/s41746-023-00797-9

195. Naiseh M, Webb C, Underwood T, Ramchurn G, Walters Z, Thavanesan N, et al. XAI for Group-AI Interaction: Towards Collaborative and Inclusive Explanations [Internet]. 2024. Available from: https://orcid.org/0000-0002-4927-5086

196. Di Ieva A. AI-augmented multidisciplinary teams: hype or hope? [Internet]. 2019. doi:10.1016/S01406736(19)326261

197. Choo JM, Ryu HS, Kim JS, Cheong JY, Baek SJ, Kwak JM, et al. Conversational artificial intelligence (chatGPT™) in the management of complex colorectal cancer patients: early experience. ANZ J Surg. 2024 Mar 1;94(3):356–61. doi:10.1111/ans.18749 PubMed PMID: 37905713.

198. Tjhin Y, Kewlani B, Singh HKSI, Pawa N. Artificial intelligence in colorectal multidisciplinary team meetings. What are the medicolegal implications? Colorectal Disease. 2024. doi:10.1111/codi.17091

199. Lee K, Lee SH. Artificial intelligence-driven oncology clinical decision support system for multidisciplinary teams. Sensors (Switzerland). 2020 Sep 1;20(17):1–12. doi:10.3390/s20174693 PubMed PMID: 32825296.

200. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak. 2019 Jul 29;19(1). doi:10.1186/s12911-019-0874-0 PubMed PMID: 31357998.

201. Boshier PR, Swaray A, Vadhwana B, O'sullivan A, Low DE, Hanna GB, et al. Systematic review and validation of clinical models predicting survival after oesophagectomy for adenocarcinoma. British Journal of Surgery. Oxford University Press; 2022. p. 418–25. doi:10.1093/bjs/znac044 PubMed PMID: 35233634.

202. Zhang Y, Zhang Z, Wei L, Wei S. Construction and validation of nomograms combined with novel machine learning algorithms to predict early death of patients with metastatic colorectal cancer. Front Public Health. 2022;10:1008137. doi:10.3389/fpubh.2022.1008137 PubMed PMID: 36605237.

203. Ren Y, Zhang Y, Zhan J, Sun J, Luo J, Liao W, et al. Machine learning for prediction of delirium in patients with extensive burns after surgery. CNS Neurosci Ther. 2023 Oct 1;29(10):2986–97. doi:10.1111/cns.14237 PubMed PMID: 37122154.

204. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Vol. 29. 2001. Report.

205. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics. 2015 Jan 2;24(1):44–65. doi:10.1080/10618600.2014.907095

206. Park J, Lee WH, Kim KT, Park CY, Lee S, Heo TY. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of the Total Environment. 2022 Aug 1;832. doi:10.1016/j.scitotenv.2022.155070 PubMed PMID: 35398119.

207. Feng Y, Wang X, Zhang J. A Heterogeneous Ensemble Learning Method for Neuroblastoma Survival Prediction. IEEE J Biomed Health Inform. 2022 Apr 1;26(4):1472–83. doi:10.1109/JBHI.2021.3073056 PubMed PMID: 33848254.

208. Yu C, Li Y, Yin M, Gao J, Xi L, Lin J, et al. Automated Machine Learning in Predicting 30-Day Mortality in Patients with Non-Cholestatic Cirrhosis. J Pers Med. 2022 Nov 1;12(11). doi:10.3390/jpm12111930

209. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016 Feb 16.

210. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017 May 22.

211. Ma S, Tourani R, Thuc E:, Le D, Liu L, Zhang K, et al. Predictive and Causal Implications of using Shapley Value for Model Interpretation. Journal of Machine Learning Research [Internet]. 2020. Report. Available from: http://jmlr.org/papers/v/.html.

212. Lagergren J, Bottai M, Santoni G. Patient Age and Survival After Surgery for Esophageal Cancer. Ann Surg Oncol. 2021 Jan 1;28(1):159–66. doi:10.1245/s10434-020-08653-w PubMed PMID: 32468352.

213. Tougeron D, Hamidou H, Scotté M, Fiore F Di, Antonietti M, Paillot B, et al. Esophageal cancer in the elderly: an analysis of the factors associated with treatment decisions and outcomes [Internet]. 2010. Report. Available from: http://www.biomedcentral.com/1471-2407/10/510 doi:10.1186/1471-2407-10-510

214. Vining P, Birdas TJ. Management of clinical T2N0 esophageal cancer: A review. Journal of Thoracic Disease. AME Publishing Company; 2019. p. S1629–32. doi:10.21037/jtd.2019.07.85

215. Dolan JP, Kaur T, Diggs BS, Luna RA, Sheppard BC, Schipper PH, et al. Significant understaging is seen in clinically staged T2N0 esophageal cancer patients undergoing esophagectomy. Diseases of the Esophagus. 2016 May 1;29(4):320–5. doi:10.1111/dote.12334 PubMed PMID: 25707341.

216. Markar SR, Gronnier C, Pasquer A, Duhamel A, Beal H, Thréreaux J, et al. Role of neoadjuvant treatment in clinical T2N0M0 oesophageal cancer: Results from a retrospective multi-center European study. Eur J Cancer. 2016 Mar 1;56:59–68. doi:10.1016/j.ejca.2015.11.024 PubMed PMID: 26808298.

217. van Rossum PSN, van Laarhoven HWM. CROSS versus modified MAGIC or FLOT in oesophageal and gastro-oesophageal junction adenocarcinoma. Lancet Gastroenterol Hepatol. 2023 Sep. doi:10.1016/S2468-1253(23)00278-9

218. Zhao X, Ren Y, Hu Y, Cui N, Wang X, Cui Y. Neoadjuvant chemotherapy versus neoadjuvant chemoradiotherapy for cancer of the esophagus or the gastroesophageal junction: A meta-analysis based on clinical trials. PLoS One. 2018 Aug 1;13(8). doi:10.1371/journal.pone.0202185 PubMed PMID: 30138325.

219. Abrams JA, Buono DL, Strauss J, McBride RB, Hershman DL, Neugut AI. Esophagectomy compared with chemoradiation for early stage esophageal cancer in the elderly. Cancer. 2009 Nov 1;115(21):4924–33. doi:10.1002/cncr.24536 PubMed PMID: 19637343.

220. Sugimura K, Miyata H, Tanaka K, Makino T, Takeno A, Shiraishi O, et al. Multicenter Randomized Phase 2 Trial Comparing Chemoradiotherapy and Docetaxel plus 5-Fluorouracil and Cisplatin Chemotherapy as Initial Induction Therapy for Subsequent Conversion Surgery in Patients with Clinical T4b Esophageal Cancer: Short-term Results. Ann Surg. 2021 Dec 1;274(6):E465–72. doi:10.1097/SLA.0000000000004564 PubMed PMID: 33065643.

221. Donlon NE, Moran B, Kamilli A, Davern M, Sheppard A, King S, et al. CROSS Versus FLOT Regimens in Esophageal and Esophagogastric Junction Adenocarcinoma: A Propensity-Matched Comparison. Ann Surg. 2022 Nov 1;276(5):792–8. doi:10.1097/SLA.0000000000005617 PubMed PMID: 35876385.

222. Zhang G, Zhang C, Sun N, Xue L, Yang Z, Fang L, et al. Neoadjuvant chemoradiotherapy versus neoadjuvant chemotherapy for the treatment of esophageal squamous cell carcinoma: a propensity score-matched study from the National Cancer Center in China. J Cancer Res Clin

Oncol. 2022 Apr 1;148(4):943–54. doi:10.1007/s00432-021-03659-7 PubMed PMID: 34013382.

223. Hoeppner J, Lordick F, Brunner T, Glatz T, Bronsert P, Röthling N, et al. ESOPEC: prospective randomized controlled multicenter phase III trial comparing perioperative chemotherapy (FLOT protocol) to neoadjuvant chemoradiation (CROSS protocol) in patients with adenocarcinoma of the esophagus (NCT02509286). BMC Cancer. 2016 Jul 19;16:503. doi:10.1186/s12885-016-2564-y PubMed PMID: 27435280.

224. Bouvier AM, Launoy G, Lepage C, Faivre J. Trends in the management and survival of digestive tract cancers among patients aged over 80 years. Aliment Pharmacol Ther. 2005 Aug 1;22(3):233–41. doi:10.1111/j.1365-2036.2005.02559.x PubMed PMID: 16091061.

225. McGillicuddy EA, Schuster KM, Davis KA, Longo WE. Factors predicting morbidity and mortality in emergency colorectal procedures in elderly patients. Arch Surg. 2009 Dec 21;144(12):1157–62. doi:10.1001/archsurg.2009.203 PubMed PMID: 20026835.

226. Racz J, Dubois L, Katchky A, Wall W. Elective and emergency abdominal surgery in patients 90 years of age or older. Canadian Journal of Surgery. 2012;55(5):322–8. doi:10.1503/cjs.007611 PubMed PMID: 22992421.

227. Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. Insights into Imaging. Springer Science and Business Media Deutschland GmbH; 2021. doi:10.1186/s13244-020-00955-7 PubMed PMID: 33564955.

228. Shea C, Khawaja AR, Sofi K, Nabi G. Association of metabolic equivalent of task (MET) score in length of stay in hospital following radical cystectomy with urinary diversion: a multi-institutional study. Int Urol Nephrol. 2021 Jul;53(7):1305–10. doi:10.1007/s11255-021-02813-x PubMed PMID: 33675471.

229. Brzezicki MA, Bridger NE, Kobetić MD, Ostrowski M, Grabowski W, Gill SS, et al. Artificial intelligence outperforms human students in conducting neurosurgical audits. Clin Neurol Neurosurg. 2020 May 1;192. doi:10.1016/j.clineuro.2020.105732 PubMed PMID: 32058200.

230. Ólafsdóttir HS, Dalqvist E, Onjukka E, Klevebro F, Nilsson M, Gagliardi G, et al. Postoperative complications after esophagectomy for cancer, neoadjuvant chemoradiotherapy compared to neoadjuvant chemotherapy: A single institutional cohort study. Clin Transl Radiat Oncol. 2023 May;40:100610. doi:10.1016/j.ctro.2023.100610

231. Bosset JF, Gignoux M, Triboulet JP, Tiret E, Mantion G, Elias D, et al. Chemoradiotherapy followed by surgery compared with surgery alone in squamous-cell cancer of the esophagus.

N Engl J Med. 1997 Jul 17;337(3):161–7. doi:10.1056/NEJM199707173370304 PubMed PMID: 9219702.

232. Walsh TN, Noonan N, Hollywood D, Kelly A, Keeling N, Hennessy TP. A comparison of multimodal therapy and surgery for esophageal adenocarcinoma. N Engl J Med. 1996 Aug 15;335(7):462–7. doi:10.1056/NEJM199608153350702 PubMed PMID: 8672151.

233. Fiorica F, Di Bona D, Schepis F, Licata A, Shahied L, Venturi A, et al. Preoperative chemoradiotherapy for oesophageal cancer: a systematic review and meta-analysis. Gut. 2004 Jul;53(7):925–30. doi:10.1136/gut.2003.025080 PubMed PMID: 15194636.

234. Mukkamalla SKR, Recio-Boiles A, Babiker HM. Esophageal Cancer. StatPearls Publishing; 2023. PubMed PMID: 29083661.

235. Reynolds J V., 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). Journal of Clinical Oncology. 2021 May 20;39(15_suppl):4004–4004. doi:10.1200/JCO.2021.39.15_suppl.4004

236. Deng HY, Wang WP, Wang YC, Hu WP, Ni PZ, Lin YD, et al. Neoadjuvant chemoradiotherapy or chemotherapy? A comprehensive systematic review and meta-analysis of the options for neoadjuvant therapy for treating oesophageal cancer. European Journal of Cardio-thoracic Surgery. European Association for Cardio-Thoracic Surgery; 2017. p. 421–31. doi:10.1093/ejcts/ezw315 PubMed PMID: 27694253.

237. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. Health Services Research and Managerial Epidemiology. SAGE Publications Inc.; 2024. doi:10.1177/23333928241234863

238. Haward R, Amir Z, Borrill C, Dawson J, Scully J, West M, et al. Breast cancer teams: The impact of constitution, new cancer workload, and methods of operation on their effectiveness. Br J Cancer. 2003 Jul 7;89(1):15–22. doi:10.1038/sj.bjc.6601073 PubMed PMID: 12838294.

239. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology. Springer; 2024. p. 3–15. doi:10.1007/s11604-023-01474-3 PubMed PMID: 37540463.

240. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med. 2023 Dec 1;6(1). doi:10.1038/s41746-023-00955-z

241. Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human Computer Studies. 2023 Jan 1;169. doi:10.1016/j.ijhcs.2022.102941

242. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, et al. Artificial Intelligence Education Programs for Health Care Professionals: Scoping Review. JMIR Medical Education. JMIR Publications Inc.; 2021. doi:10.2196/31043

243. Jirotka M, Grimpe B, Stahl B, Eden G, Hartswood M. Responsible research and innovation in the digital age. Commun ACM. 2017 Apr 24;60(5):62–8. doi:10.1145/3064940

244. Cancer Research UK. Cancer Research UK Cancers with unmet needs [Internet]. [cited 2024 Jun 9]. Available from: https://www.cancerresearchuk.org/funding-for-researchers/research-opportunities-in-harder-to-treat-cancers#portfolio2

245. Liu CQ, Ma YL, Qin Q, Wang PH, Luo Y, Xu PF, et al. Epidemiology of esophageal cancer in 2020 and projections to 2030 and 2040. Thorac Cancer. 2023 Jan;14(1):3–11. doi:10.1111/1759-7714.14745 PubMed PMID: 36482832.

246. Taylor C, Munro AJ, Glynne-Jones R, Griffith C, Trevatt P, Richards M, et al. Multidisciplinary team working in cancer: what is the evidence? BMJ. 2010 Mar 23;340(mar23 2):c951–c951. doi:10.1136/bmj.c951

247. NHS England. NHS England [Internet]. 2024 [cited 2025 May 4]. NHS National Cost Collection Publication: National Schedule 2023/24. Available from: https://app.powerbi.com/view?r=eyJrIjoiZGQxYjNkOGUtOTIwMC00N2VjLWEyM2EtYjAzOGMwNWU5ODQ1IiwidCI6IjM3YzM1NGIyLTg1YjAtNDdmNS1iMjIyLTA3YjQ4ZDc3NGVlMyJ9

248. Kapoor V, Mittal A, Garg S, Diwakar M, Mishra AK, Singh P. Lung Cancer Detection Using VGG16 and CNN. doi:10.1109/AIC.2023.125

249. Kumar S, Singh J, Ravi V, Singh P, Al Mazroa A, Diwakar M, et al. Utilizing Multi-layer Perceptron for Esophageal Cancer Classification Through Machine Learning Methods. Open Public Health J. 2024 Oct 9;17(1). doi:10.2174/0118749445335423240808062700

250. Yang KY, Mukundan A, Tsao YM, Shi XH, Huang CW, Wang HC. Assessment of hyperspectral imaging and CycleGAN-simulated narrowband techniques to detect early esophageal cancer. Sci Rep. 2023 Dec 1;13(1). doi:10.1038/s41598-023-47833-y PubMed PMID: 37993660.

251. Pucher PH, Rahman SA, Bhandari P, Blencowe N, Chidambaram S, Crosby T, et al. Prevalence and Risk Factors for Malignant Nodal Involvement in Early esophago-gastric Adenocarcinoma. Ann Surg. 2024 Sep 2. doi:10.1097/SLA.0000000000006496

252. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? [Internet]. 2022 Jul 18. Available from: http://arxiv.org/abs/2207.08815

253. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol. 2017 Apr;70(2):144–56. doi:10.4097/kjae.2017.70.2.144 PubMed PMID: 28367284.

254. Higgins D, Madai VI. From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. Advanced Intelligent Systems. 2020 Oct;2(10). doi:10.1002/aisy.202000052

255. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: An online mobile survey. J Med Internet Res. 2019 Mar 1;21(3). doi:10.2196/12422 PubMed PMID: 30907742.

256. Reddy S, Lebrun A, Chee A, Kalogeropoulos D. The Role of Explainable AI and Evaluation Frameworks for Safe and Effective Integration of Large Language Models in Healthcare. Telehealth and Medicine Today. 2024 Apr 30;9(2). doi:10.30953/thmt.v9.485

257. European Commision. The EU Artificial Intelligence Act. Official Journal (OJ) of the European Union. 2024 Jul 12.

258. Collingridge D. The Social Control of Technology. American Political Science Review. 1980.

259. Rahman SA, Walker RC, Lloyd MA, Grace BL, van Boxel GI, Kingma BF, et al. Machine learning to predict early recurrence after oesophageal cancer surgery. British Journal of Surgery. 2020 Jul 1;107(8):1042–52. doi:10.1002/bjs.11461 PubMed PMID: 31997313.

260. Kalff MC, Dijksterhuis WPM, Wagner AD, Oertelt-Prigione S, Verhoeven RHA, Lemmens VEPP, et al. Sex differences in treatment allocation and survival of potentially curable gastroesophageal cancer: A population-based study. Eur J Cancer. 2023 Jul;187:114–23. doi:10.1016/j.ejca.2023.04.002 PubMed PMID: 37146505.

261. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. Nature Research; 2019. p. 206–15. doi:10.1038/s42256-019-0048-x

262. Janjigian YY, Van Cutsem E, Muro K, Wainberg Z, Al-Batran SE, Hyung WJ, et al. MATTERHORN: phase III study of durvalumab plus FLOT chemotherapy in resectable gastric/gastroesophageal junction cancer. Future Oncol. 2022 Jun;18(20):2465–73. doi:10.2217/fon-2022-0093 PubMed PMID: 35535555.

263. Jones CM, Hawkins M, Mukherjee S, Radhakrishna G, Crosby T. Considerations for the Treatment of Oesophageal Cancer With Radiotherapy During the COVID-19 Pandemic. Clin Oncol (R Coll Radiol). 2020 Jun;32(6):354–7. doi:10.1016/j.clon.2020.04.001 PubMed PMID: 32299723.

264. Lin HS, Watts JN, Peel NM, Hubbard RE. Frailty and post-operative outcomes in older surgical patients: A systematic review. BMC Geriatrics. BioMed Central Ltd.; 2016. doi:10.1186/s12877-016-0329-8 PubMed PMID: 27580947.

265. Datta SS, Ghosal N, Daruvala R, Chakraborty S, Shrimali RK, van Zanten C, et al. How do clinicians rate patient's performance status using the ECOG performance scale? A mixed-methods exploration of variability in decision-making in oncology. Ecancermedicalscience. 2019;13:913. doi:10.3332/ecancer.2019.913 PubMed PMID: 31123496.

266. Simcock R, Wright J. Beyond Performance Status. Clin Oncol (R Coll Radiol). 2020 Sep;32(9):553–61. doi:10.1016/j.clon.2020.06.016 PubMed PMID: 32684503.

267. Pfisterer KJ, Lohani R, Janes E, Ng D, Wang D, Bryant-Lukosius D, et al. An Actionable Expert-System Algorithm to Support Nurse-Led Cancer Survivorship Care: Algorithm Development Study. JMIR Cancer. 2023 Oct 4;9:e44332. doi:10.2196/44332

268. Pucher PH, Rahman SA, Bhandari P, Blencowe N, Chidambaram S, Crosby T, et al. Prevalence and Risk Factors for Malignant Nodal Involvement in Early esophago-gastric Adenocarcinoma: Results from the Multicenter Retrospective Congress Study (endosCopic resectiON, esophaGectomy or Gastrectomy For Early Esophagogastric Cancers). Ann Surg. 2024 Sep 2. doi:10.1097/SLA.0000000000006496 PubMed PMID: 39219545.

269. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ Qual Saf. 2013 Oct;22 Suppl 2(Suppl 2):ii58–64. doi:10.1136/bmjqs-2012-001712 PubMed PMID: 23882089.

270. Stanovich K. Rationality and the Reflective Mind. Oxford University Press; 2010. doi:10.1093/acprof:oso/9780195341140.001.0001

271. Arkes HR. Costs and benefits of judgment errors: Implications for debiasing. Psychol Bull. 1991;110(3):486–98. doi:10.1037/0033-2909.110.3.486

272. Kalff MC, Van Berge Henegouwen MI, Gisbertz SS. Textbook outcome for esophageal cancer surgery: An international consensus-based update of a quality measure. Diseases of the Esophagus. 2021 Jul 1;34(7). doi:10.1093/dote/doab011 PubMed PMID: 33744921.

273. Buchholz V, Hazard R, Lee DK, Liu DS, Zhang W, Chen S, et al. Textbook outcomes after oesophagectomy: a single-centre observational study. BMC Surg. 2023 Dec 1;23(1). doi:10.1186/s12893-023-02253-7 PubMed PMID: 38066440.

274. NHS innovation services. Innovation Service Your guide to innovation in the NHS Regulation stage. Report.

275. Overgaard SM, Graham MG, Brereton T, Pencina MJ, Halamka JD, Vidal DE, et al. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. npj Digital Medicine. Nature Research; 2023. doi:10.1038/s41746-023-00968-8

276. US FDA, Health Canada, UK MHRA. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. Report.

277. Care Quality Commision. Southampton General Hospital Quality Report. 2015. Report.

278. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12–8. doi:10.11613/BM.2014.003 PubMed PMID: 24627710.

279. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. The Lancet Digital Health. Elsevier Ltd; 2021. p. e337–8. doi:10.1016/S2589-7500(21)00076-5 PubMed PMID: 33933404.

280. Observatory for Responsible Research and Innovation in ICT. ORBIT AREA 4 Ps [Internet]. 2024 [cited 2024 Aug 5]. Available from: https://orbit-rri.org/

281. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020 Jun 9;117(23):12592–4. doi:10.1073/pnas.1919012117 PubMed PMID: 32457147.

282. KPMG international. Decoding the EU AI Act. 2024. Report.

283. Greenhalgh C, Craigon P, Portillo V, Dowthwaite L, Perez Vallejos E, Webb H, et al. University of Nottingham Research Data Management Service [Internet]. 2023 [cited 2024 Jul 22]. Responsible Innovation (RI) Prompts and Practice Cards (version 3.1.1, November 2023). Available from: https://rdmc.nottingham.ac.uk/handle/internal/10930 doi:10.17639/nott.7353

284. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006 Jan;3(2):77–101. doi:10.1191/1478088706qp063oa

285. Baştanlar Y, Ozuysal M. Introduction to machine learning. Methods Mol Biol. 2014;1107:105–28. doi:10.1007/978-1-62703-748-8_7 PubMed PMID: 24272434.

286. Hefin I. Rhys. Machine Learning with R, the Tidyverse and mlr. 1st ed. Manning; 2020.

287. Moubayed A, Injadat M, Nassif AB, Lutfiyya H, Shami A. E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics. IEEE Access. 2018 Jul 20;6:39117–38. doi:10.1109/ACCESS.2018.2851790

288. Levett DZH, Jack S, Swart M, Carlisle J, Wilson J, Snowden C, et al. Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation. Br J Anaesth. 2018 Mar;120(3):484–500. doi:10.1016/j.bja.2017.10.020 PubMed PMID: 29452805.

289. Zannoni J, Guazzi M, Milani V, Bandera F, Alfonzetti E, Arena R. Prognostic value of cardiopulmonary exercise testing in a European cohort with cardiovascular risk factors absent of a cardiovascular disease diagnosis. Int J Cardiol. 2023 Jan 1;370:402–4. doi:10.1016/j.ijcard.2022.10.016 PubMed PMID: 36228767.

290. Andonian BJ, Hardy N, Bendelac A, Polys N, Kraus WE. Making Cardiopulmonary Exercise Testing Interpretable for Clinicians. Curr Sports Med Rep. 2021 Oct 1;20(10):545–52. doi:10.1249/JSR.0000000000000895 PubMed PMID: 34622820.

291. Stubbs DJ, Grimes LA, Ercole A. Performance of cardiopulmonary exercise testing for the prediction of post-operative complications in non cardiopulmonary surgery: A systematic review. PLoS One. 2020 Feb 3;15(2):e0226480. doi:10.1371/journal.pone.0226480

292. Stack exchange Contributors. Stack Exchange [Internet]. 2020 [cited 2026 Mar 14]. Relevanc e of Power Calculations to Machine Learning models. Available from: https://stats.stackexchange.com/questions/471822/are-there-any-power-calculation-formulas-for-ml-methods-beyond-logistic-regressi

293. Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. The Lancet Digital Health. Elsevier Ltd; 2025. doi:10.1016/j.landig.2025.01.013 PubMed PMID: 40461350.

294. Juan C Olamendy. Medium.com [Internet]. 2024 [cited 2025 Mar 3]. Real world ML — Determine the Optimal Sample Size. Available from: https://medium.com/@juanc.olamendy/real-world-ml-determine-the-optimal-sample-size-

210c2e01651e#:~:text=Have%20you%20ever%20wondered%20how,your%20project's%20time%20and%20costs.

295. Doroudi S, Rastegar SA. The Bias–Variance Tradeoff in Cognitive Science. Cognitive Science. John Wiley and Sons Inc; 2023. doi:10.1111/cogs.13241 PubMed PMID: 36655991.

296. Sainani KL. Logistic regression. PM and R. 2014 Dec 1;6(12):1157–62. doi:10.1016/j.pmrj.2014.10.006 PubMed PMID: 25463689.

297. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspect Clin Res. 2017 Jul 1;8(3):148–51. doi:10.4103/picr.PICR_87_17 PubMed PMID: 28828311.

298. Kwak C, Clayton-Matthews A. Multinomial logistic regression. Nurs Res. 2002;51(6):404–10. doi:10.1097/00006199-200211000-00009 PubMed PMID: 12464761.

299. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015 Apr 1;27(2):130–5. doi:10.11919/j.issn.1002-0829.215044 PubMed PMID: 26120265.

300. Breiman L, Friedman J, Stone C, Olshen R. Classification and regression trees. Taylor & Francis 1984; 1984. 1–368 p.

301. Kingsford C, Salzberg SL. What are decision trees? Nature Biotechnology. 2008. p. 1011–2. doi:10.1038/nbt0908-1011 PubMed PMID: 18779814.

302. Raileanu LE, Stoffel K. Theoretical comparison between the Gini Index and Information Gain criteria *. Annals of Mathematics and Artificial Intelligence. Kluwer Academic Publishers; 2004. Report.

303. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7(DEC). doi:10.3389/fnbot.2013.00021

304. Owusu-Adjei M, Ben Hayfron-Acquah J, Frimpong T, Abdul-Salaam G. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. PLOS Digital Health. 2023 Nov 30;2(11):e0000290. doi:10.1371/journal.pdig.0000290

305. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007 Feb;115(5):654–7. doi:10.1161/CIRCULATIONAHA.105.594929 PubMed PMID: 17283280.

306. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning - ICML '06. New York, New York, USA: ACM Press; 2006. p. 233–40. doi:10.1145/1143844.1143874

307. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. Mach Learn. 2021 Mar 1;110(3):451–6. doi:10.1007/s10994-021-05964-1

308. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J Anal Test. 2018 Jul 1;2(3):249–62. doi:10.1007/s41664-018-0068-2

309. Marcinkevičs R, Vogt JE. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. Wiley Interdiscip Rev Data Min Knowl Discov. 2023. doi:10.1002/widm.1493

310. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. Nature Research; 2019. p. 206–15. doi:10.1038/s42256-019-0048-x

311. Wei P, Lu Z, Song J. Variable importance analysis: A comprehensive review. Reliability Engineering and System Safety. Elsevier Ltd; 2015. p. 399–432. doi:10.1016/j.ress.2015.05.018

312. Curia F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. Health Technol (Berl). 2021 Jul 1;11(4):875–85. doi:10.1007/s12553-021-00554-6

313. Christopher Molnar. Interpretable Machine Learning [Internet]. Second Edition. Independently Published; 2022 [cited 2024 May 9]. Available from: https://christophm.github.io/interpretable-ml-book/

314. Huang J, Koulaouzidis A, Marlicz W, Lok V, Chu C, Ngai CH, et al. Global Burden, Risk Factors, and Trends of Esophageal Cancer: An Analysis of Cancer Registries from 48 Countries. Cancers (Basel). 2021 Jan 5;13(1). doi:10.3390/cancers13010141 PubMed PMID: 33466239.

315. Independent Cancer Taskforce. Achieving world-class cancer outcomes: a strategy for England 2015-2020 [Internet]. 2015 [cited 2024 Apr 19]. Available from: http://bit.ly/1Idwf5W

316. NHS England. Supporting clinical decisions with health information technology: an implementation guide for clinical decision support systems [Internet]. 2023 [cited 2024 Apr 23]. Available from: https://www.england.nhs.uk/long-read/supporting-clinical-decisions-with-health-information-technology/

317. Global Market Insights. Artificial intelligence in healthcare market - by application (medical imaging & diagnosis, drug discovery, therapy planning, hospital workflow, wearables, virtual assistants), by region & global forecast, 2023 - 2032 [Internet]. 2023 [cited 2024 Apr 2]. Available from: https://www.gminsights.com/industry-analysis/healthcare-artificial-intelligence-market

318. Khanna NN, Maindarkar MA, Viswanathan V, Fernandes JFE, Paul S, Bhagawati M, et al. Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. Healthcare (Basel). 2022 Dec 9;10(12). doi:10.3390/healthcare10122493 PubMed PMID: 36554017.

319. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking Bias: The Role of Artificial Intelligence in Improving Clinical Decision-Making. Cureus. 2023 Mar;15(3):e36415. doi:10.7759/cureus.36415 PubMed PMID: 37090406.

320. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med. 2023 Jun 14;6(1):113. doi:10.1038/s41746-023-00858-z PubMed PMID: 37311802.

321. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 2018 Dec 4;320(21):2199–200. doi:10.1001/jama.2018.17163 PubMed PMID: 30398550.

322. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020 Nov 30;20(1):310. doi:10.1186/s12911-020-01332-6 PubMed PMID: 33256715.

323. Sanchez-Martinez S, Camara O, Piella G, Cikes M, González-Ballester MÁ, Miron M, et al. Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging. Front Cardiovasc Med. 2021;8:765693. doi:10.3389/fcvm.2021.765693 PubMed PMID: 35059445.

324. Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. Med (N Y). 2021 Jun 11;2(6):642–65. doi:10.1016/j.medj.2021.04.006 PubMed PMID: 35590138.

325. Ankolekar A, van der Heijden B, Dekker A, Roumen C, De Ruysscher D, Reymen B, et al. Clinician perspectives on clinical decision support systems in lung cancer: Implications for shared decision-making. Health Expect. 2022 Aug;25(4):1342–51. doi:10.1111/hex.13457 PubMed PMID: 35535474.

326. Dijksterhuis WPM, Kalff MC, Wagner AD, Verhoeven RHA, Lemmens VEPP, van Oijen MGH, et al. Gender Differences in Treatment Allocation and Survival of Advanced Gastroesophageal

Cancer: A Population-Based Study. J Natl Cancer Inst. 2021 Nov 2;113(11):1551–60. doi:10.1093/jnci/djab075 PubMed PMID: 33837791.

327. Qualtrics. Qualtrics Homepage.

328. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373–83. doi:10.1016/0021-9681(87)90171-8 PubMed PMID: 3558716.

329. Alcorn S, Foo I. Perioperative management of patients with dementia. BJA Educ. 2017 Mar;17(3):94–8. doi:10.1093/bjaed/mkw038

330. Kotzé A, Howell SJ. Heart failure: pathophysiology, risk assessment, community management and anaesthesia. Continuing Education in Anaesthesia Critical Care & Pain. 2008 Sep;8(5):161–6. doi:10.1093/bjaceaccp/mkn028

331. Soto-Perez-de-Celis E, Li D, Yuan Y, Lau YM, Hurria A. Functional versus chronological age: geriatric assessments to guide decision making in older patients with cancer. Lancet Oncol. 2018 Jun;19(6):e305–16. doi:10.1016/S1470-2045(18)30348-6 PubMed PMID: 29893262.

332. van Kolfschooten H. The AI cycle of health inequity and digital ageism: mitigating biases through the EU regulatory framework on medical devices. J Law Biosci. 2023;10(2):lsad031. doi:10.1093/jlb/lsad031 PubMed PMID: 38075950.

333. European Union (EU). 2024/1689 laying down harmonised rules on artificial intelligence [Internet]. 2024 [cited 2024 Aug 20]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

334. Soukup T, Sevdalis N, Green JSA, Lamb BW, Chapman C, Skolarus TA. Making Tumor Boards More Patient-Centered: Let's Start With the Name. JCO Oncol Pract. 2021 Oct;17(10):591–3. doi:10.1200/OP.20.00588 PubMed PMID: 33734827.

335. Anjara SG, Janik A, Dunford-Stenger A, Mc Kenzie K, Collazo-Lorduy A, Torrente M, et al. Examining explainable clinical decision support systems with think aloud protocols. PLoS One. 2023;18(9):e0291443. doi:10.1371/journal.pone.0291443 PubMed PMID: 37708135.

336. Brigden T, Mitchell C, Redrup Hill E, Hall A. Ethical and legal implications of implementing risk algorithms for early detection and screening for oesophageal cancer, now and in the future. PLoS One. 2023;18(10):e0293576. doi:10.1371/journal.pone.0293576 PubMed PMID: 37903120.

337. Emani S, Rui A, Rocha HAL, Rizvi RF, Juaçaba SF, Jackson GP, et al. Physicians' Perceptions of and Satisfaction With Artificial Intelligence in Cancer Treatment: A Clinical Decision Support System Experience and Implications for Low-Middle-Income Countries. JMIR Cancer. 2022 Apr 7;8(2):e31461. doi:10.2196/31461 PubMed PMID: 35389353.

338. Pumplun L, Peters F, Gawlitza JF, Buxmann P. Bringing Machine Learning Systems into Clinical Practice: A Design Science Approach to Explainable Machine Learning-Based Clinical Decision Support Systems. J Assoc Inf Syst. 2023;24(4):953–79. doi:10.17705/1jais.00820

339. Staes CJ, Beck AC, Chalkidis G, Scheese CH, Taft T, Guo JW, et al. Design of an interface to communicate artificial intelligence-based prognosis for patients with advanced solid tumors: a user-centered approach. J Am Med Inform Assoc. 2023 Dec 22;31(1):174–87. doi:10.1093/jamia/ocad201 PubMed PMID: 37847666.

340. Kočo L, Siebers CCN, Schlooz M, Meeuwis C, Oldenburg HSA, Prokop M, et al. The Facilitators and Barriers of the Implementation of a Clinical Decision Support System for Breast Cancer Multidisciplinary Team Meetings-An Interview Study. Cancers (Basel). 2024 Jan 17;16(2). doi:10.3390/cancers16020401 PubMed PMID: 38254891.

341. Helenason J, Ekström C, Falk M, Papachristou P. Exploring the feasibility of an artificial intelligence based clinical decision support system for cutaneous melanoma detection in primary care - a mixed method study. Scand J Prim Health Care. 2024 Mar;42(1):51–60. doi:10.1080/02813432.2023.2283190 PubMed PMID: 37982736.

342. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak. 2023 Apr 20;23(1):73. doi:10.1186/s12911-023-02162-y PubMed PMID: 37081503.

343. Shiyab W, Ferguson C, Rolls K, Halcomb E. Solutions to address low response rates in online surveys. European journal of cardiovascular nursing. 2023 May 25;22(4):441–4. doi:10.1093/eurjcn/zvad030 PubMed PMID: 36827086.

344. Bazeley P. Analysing Mixed Methods Data. In: Mixed Methods Research for Nursing and the Health Sciences. Wiley; 2009. p. 84–118. doi:10.1002/9781444316490.ch6