An interpretable AI model for recurrence prediction after surgery in gastrointestinal stromal tumour: an observational cohort study

Dimitris Bertsimas,[a,n] Georgios Antonios Margonis,[b,n] Seehanah Tang,[a] Angelos Koulouras,[a] Cristina R. Antonescu,[c] Murray F. Brennan,[b] Javier Martin-Broto,[d,l,m] Piotr Rutkowski,[e] Georgios Stasinos,[f] Jane Wang,[g] Emmanouil Pikoulis,[h] Elzbieta Bylina,[e] Pawel Sobczuk,[e] Antonio Gutierrez,[d,l,m] Bhumika Jadeja,[b] William D. Tap,[i] Ping Chi,[i,j,k] and Samuel Singer[b,*]

[a]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
[b]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[c]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[d]Medical Oncology Department, Fundación Jimenez Diaz University Hospital, Madrid, Spain
[e]Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland
[f]Technical Chamber of Greece, Athens, Greece
[g]Department of Surgery, University of California San Francisco, San Francisco, CA, USA
[h]Third Department of Surgery, Attikon University Hospital, Athens, Greece
[i]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[j]Human Oncology and Pathogenesis Program (HOPP), Memorial Sloan Kettering Cancer Center, New York, NY, USA
[k]Department of Medicine, Weill Cornell Medical College, New York, NY, USA
[l]Hospital General de Villalba, Madrid, Spain
[m]Instituto de Investigacion Sanitaria Fundacion Jimenez Diaz (IIS/FJD; UAM), Madrid, Spain

eClinicalMedicine
2023;64: 102200

Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.102200

Summary

Background There are several models that predict the risk of recurrence following resection of localised, primary gastrointestinal stromal tumour (GIST). However, assessment of calibration is not always feasible and when performed, calibration of current GIST models appears to be suboptimal. We aimed to develop a prognostic model to predict the recurrence of GIST after surgery with both good discrimination and calibration by uncovering and harnessing the non-linear relationships among variables that predict recurrence.

Methods In this observational cohort study, the data of 395 adult patients who underwent complete resection (R0 or R1) of a localised, primary GIST in the pre-imatinib era at Memorial Sloan Kettering Cancer Center (NY, USA) (recruited 1982–2001) and a European consortium (Spanish Group for Research in Sarcomas, 80 sites) (recruited 1987–2011) were used to train an interpretable Artificial Intelligence (AI)-based model called Optimal Classification Trees (OCT). The OCT predicted the probability of recurrence after surgery by capturing non-linear relationships among predictors of recurrence. The data of an additional 596 patients from another European consortium (Polish Clinical GIST Registry, 7 sites) (recruited 1981–2013) who were also treated in the pre-imatinib era were used to externally validate the OCT predictions with regard to discrimination (Harrell's C-index and Brier score) and calibration (calibration curve, Brier score, and Hosmer-Lemeshow test). The calibration of the Memorial Sloan Kettering (MSK) GIST nomogram was used as a comparative gold standard. We also evaluated the clinical utility of the OCT and the MSK nomogram by performing a Decision Curve Analysis (DCA).

Findings The internal cohort included 395 patients (median [IQR] age, 63 [54–71] years; 214 men [54.2%]) and the external cohort included 556 patients (median [IQR] age, 60 [52–68] years; 308 men [55.4%]). The Harrell's C-index of the OCT in the external validation cohort was greater than that of the MSK nomogram (0.805 (95% CI: 0.803–0.808) vs 0.788 (95% CI: 0.786–0.791), respectively). In the external validation cohort, the slope and intercept of the calibration curve of the main OCT were 1.041 and 0.038, respectively. In comparison, the slope and intercept of the calibration curve for the MSK nomogram was 0.681 and 0.032, respectively. The MSK nomogram overestimated the recurrence risk throughout the entire calibration curve. Of note, the Brier score was lower for the OCT compared to the MSK nomogram (0.147 vs 0.564, respectively), and the Hosmer-Lemeshow test was insignificant ($P = 0.087$) for the OCT model but significant ($P < 0.001$) for the MSK nomogram. Both results confirmed the superior

*Corresponding author. Howard 1205, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY, 10065, USA.
E-mail address: singers@mskcc.org (S. Singer).

[n]Authors had equal contribution and share first-authorship.

discrimination and calibration of the OCT over the MSK nomogram. A decision curve analysis showed that the AI-based OCT model allowed for superior decision making compared to the MSK nomogram for both patients with 25–50% recurrence risk as well as those with >50% risk of recurrence.

Interpretation **We present the first prognostic models of recurrence risk in GIST that demonstrate excellent discrimination, calibration, and clinical utility on external validation. Additional studies for further validation are warranted. With further validation, these tools could potentially improve patient counseling and selection for adjuvant therapy.**

Funding **The NCI SPORE in Soft Tissue Sarcoma and NCI Cancer Center Support Grants.**

Keywords: **GIST; Prognosis; Recurrence; Artificial intelligence**

Research in context

Evidence before this study

The accurate estimation of risk of recurrence following resection of primary gastrointestinal stromal tumour (GIST) not only informs patient prognosis but can also guide patient selection for imatinib. Specifically, it can spare those who have already achieved cure from surgical resection from unnecessary cost and drug toxicity and reserve adjuvant imatinib for those at high risk for relapse. A PubMed search between January 1, 2009, and April 6, 2023, was performed using the terms "gastrointestinal stromal tumours" or "GIST" and "predictive" or "prognostic model" to identify studies on the development or validation of statistical models that aimed to predict recurrence following complete resection of primary GIST. Several of these prognostic models were published in high impact clinical oncology journals, such as *The Lancet Oncology* (ie, the Memorial Sloan Kettering (MSK) nomogram by Gold and colleagues and the Joensuu heat maps). Unfortunately, even the best models suffer from many deficiencies. For example, the original National Institutes of Health (NIH) criteria were formulated based on expert consensus and literature review and were not statistically validated. Similarly, neither the Armed Forces Institute of Pathology (AFIP) criteria by Miettinen et al. nor the modified NIH criteria by Joensuu et al. underwent formal statistical validation. The MSK nomogram by Gold and colleagues appears to consistently overestimate the actual risk of recurrence across all patient groups. Finally, the Joensuu heat maps underwent a very limited external validation of its discriminatory ability and no evaluation of calibration. In addition, because the online tool that was published by the authors in 2016 is no longer available, the scientific community cannot evaluate its concordance index or calibration.

Added value of this study

To remedy these deficiencies and improve upon existing models, one could either increase the statistical power by increasing the sample size of the cohort and/or adopt a different methodological approach. The former is not possible as only patients who underwent resection of a primary GIST in the pre-imatinib era (before 2001 in the USA) and thus retained their natural history are eligible; including patients from the imatinib era who did not receive treatment would introduce serious selection and confounding bias to an analysis. Thus, we collaborated with a group from the Massachusetts Institute of Technology (MIT) that developed their own interpretable artificial intelligence (AI)-based methodology (ie, optimal classification trees [OCT]), which has been successfully employed in the medical field. To the best of our knowledge, the GIST OCT has one of the highest concordance indices ever reported on an external validation of a GIST prognostic model and is the only GIST prognostic model that has excellent calibration on external validation.

Implications of all the available evidence

Currently, the decision to offer adjuvant imatinib relies on the clinician's expertise and available prognostic models, which have unknown or poor calibration. The GIST OCT demonstrated excellent predictive performance and superior clinical utility in an external cohort consisting of registry data from geographically distinct hospitals with high heterogeneity. This attests to the generalisability of its predictions in various practice settings. Importantly, its superior performance did not come at the cost of limited transparency, which occurs in many AI-driven models such as gradient boosting and deep learning; these models are considered "black box" models, which means it is nearly impossible for a human to understand exactly how the input was used to construct the model predictions. This advantage of OCTs is particularly important as lack of interpretability may serve as a significant barrier to clinical implementation of AI technology. Although we externally validated the GIST OCT prognostic model, additional studies for further validation are warranted.

Introduction

The standard treatment for localised, primary gastrointestinal stromal tumour (GIST) is surgical resection, but many patients recur post-operatively.[1,2] The cost and potential cumulative toxicity of imatinib mandate adjuvant application in only those at high-risk for relapse. Thus, accurate risk-stratification following surgery is needed. Current risk-stratification tools are based on models that combine well-established predictors of recurrence. Models include the original National Institutes of Health (NIH) criteria, the modified NIH criteria, the Armed Forces Institute of Pathology (AFIP) Miettinen criteria, the Joensuu contour maps, and the Memorial Sloan Kettering (MSK) nomogram.[3–7] Their popularity stems from the high area under the curve (AUC) or concordance indices reported in both the original publications and external validations.[5,6,8] A high AUC or concordance index indicates that these models can accurately rank patients according to risk of recurrence.[7] However, a perfect ranking does not necessarily equate to correct prediction of probability of recurrence. For example, three patients with recurrence risks of 20, 40, and 60% can be appropriately ranked both by a model that correctly assigns them probabilities of 20, 40 and 60% and by a model that incorrectly assigns them probabilities of 30, 50 and 70%. The measurement of how well a model matches the probability of the event (e.g., recurrence) is a distinct property called calibration.[9]

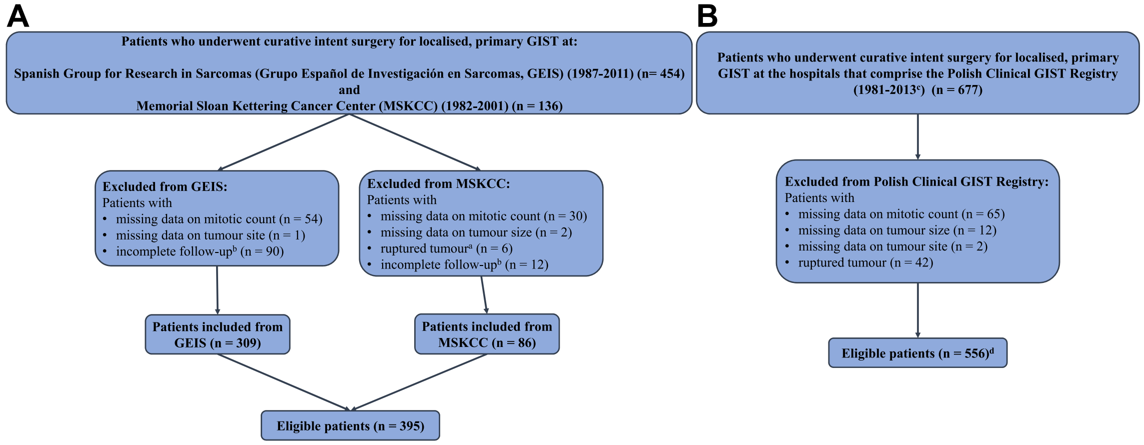
The importance of good calibration may be even more pronounced in GIST, as the model can be utilised to not only inform patients about their prognosis, but also decide who receives adjuvant imatinib. The calibration of the original NIH criteria, the modified NIH criteria, and the AFIP-Miettinen criteria cannot be assessed because they only provide a qualitative estimate of risk-stratification (e.g., low vs moderate vs high-risk).[3,4,7] The calibration of the Joensuu contour maps has also not been validated because they demonstrate ranges of probabilities of recurrence and not a specific probability for a given patient.[5] Across the most commonly used risk-stratification tools, only the MSK nomogram predicts quantifiable risk of recurrence for a given patient.[6]

Using a novel methodology could allow one to devise a model with both a high concordance index and excellent calibration. Joensuu et al. have shown a non-linear effect of tumour size and mitotic count on recurrence.[5] Thus, we hypothesised that important non-linear relationships also exist among variables that predict recurrence, and an interpretable Artificial Intelligence (AI) framework that could uncover and harness them may be an optimal option. We collaborated with the Massachusetts Institute of Technology (MIT) to employ interpretable AI-based techniques to devise a model with excellent calibration and compared it to the MSK nomogram. We also investigated whether the addition of *KIT* mutational status can further improve the interpretable AI-based model.

Methods

Internal cohort

All adult patients who underwent complete resection (R0 or R1) of a localised, primary GIST in the pre-imatinib era (and thus retained their natural history) at the Memorial Sloan Kettering Cancer Center (MSKCC) (recruited 1982–2001) and at the 80 hospitals of the Spanish Group for Research in Sarcomas (Grupo Español de Investigación en Sarcomas, GEIS) (recruited 1987–2011) were considered for inclusion (Fig. 1).


A
Patients who underwent curative intent surgery for localised, primary GIST at:
Spanish Group for Research in Sarcomas (Grupo Español de Investigación en Sarcomas, GEIS) (1987-2011) (n= 454)
and
Memorial Sloan Kettering Cancer Center (MSKCC) (1982-2001) (n = 136)
Excluded from GEIS:
Patients with
• missing data on mitotic count (n = 54)
• missing data on tumour site (n = 1)
• incomplete follow-up^b (n = 90)
Excluded from MSKCC:
Patients with
• missing data on mitotic count (n = 30)
• missing data on tumour size (n = 2)
• ruptured tumour^a (n = 6)
• incomplete follow-up^b (n = 12)
Patients included from GEIS (n = 309)
Patients included from MSKCC (n = 86)
Eligible patients (n = 395)
B
Patients who underwent curative intent surgery for localised, primary GIST at the hospitals that comprise the Polish Clinical GIST Registry (1981-2013^c) (n = 677)
Excluded from Polish Clinical GIST Registry:
Patients with
• missing data on mitotic count (n = 65)
• missing data on tumour size (n = 12)
• missing data on tumour site (n = 2)
• ruptured tumour (n = 42)
Eligible patients (n = 556)^d

[a] Given that tumour rupture was very infrequent in the MSKCC dataset (n=6) and these data were not available in the Spanish GEIS registry cohort, we excluded these patients. Since patients with tumour rupture were excluded in the training cohort, they were similarly excluded in the Polish registry dataset that was used to externally validate the OCT model.
[b] Patients who did not have recurrence but were lost to follow-up within 5 years from surgery.
[c] Because adjuvant imatinib became fully available outside clinical trials in Poland in 2013, all patients in the Polish cohort belong to the pre-imatinib era.
[d] We did not exclude any patients with incomplete follow-up from the Polish cohort because we wanted to evaluate how the OCT performed with real world data (RWD), which commonly have limited follow-up.

Fig. 1: Flowchart of study cohort selection: (A) Internal, (B) External.

External cohort

The data of patients from another European consortium (Polish Clinical GIST Registry) similarly treated in the pre-imatinib era (recruited 1981–2013) were used to externally validate the OCT predictions (Fig. 1).

Ethical standards

The study was conducted in accordance with the ethical standards of the participating institutions and was approved by their respective institutional review boards (IRBs). The IRB study protocol designation for each institution is as follows: IRB protocol 16–1583 (Memorial Sloan Kettering Cancer Center), IRB protocol 2016/195 (Grupo Español de Investigación en Sarcomas, GEIS), and IRB protocols KB/9/2011 and 119/2002 (Polish Clinical GIST Registry and Maria Sklodowska-Curie Institute—Oncology Center, respectively). The IRB approved the waiver of authorisation for MSKCC and the Polish Clinical GIST Registry. Informed consent was obtained from all participants undergoing follow-up for GEIS.

Rationale for selecting the OCT methodology

Optimal classification trees (OCTs) are state-of-the-art decision tree methods and have several technical advantages over other AI-based methods.[10,11] Importantly, OCTs not only do not share the limitations which have impeded the wide adoption of AI methodology in healthcare, but have a unique combination of interpretability and excellent performance.[12] Because of these properties, OCTs have been repeatedly used to predict outcomes in acute care surgery, cardiac surgery, and surgical oncology.[13–22] A detailed discussion of the main advantages of OCTs over the other decision tree-based methods is available in the relevant passage of the eMethods section. Given the several theoretical advantages of the OCT over other decision tree models, we were also interested in comparing its performance to “black box” machine learning (ML)-based models such as Random Forest (RF) and XGBoost as well as the popular Classification and Regression Trees (CART) (eMethods).

Feature selection and OCT training

We trained the OCTs (please see the protocol that is provided within the appendix for details) with three well-known, independent predictors of recurrence (tumour size, mitotic count, and tumour site) that are routinely available and used by all five current risk-stratification tools (the original NIH criteria, modified NIH criteria, AFIP-Miettinen criteria, Joensuu contour maps, and the MSK nomogram).[3–7]

Five feature selection analyses (optimal feature selection (OFS), XGBoost training and Shapley Additive exPlanations (SHAP) analysis, least absolute shrinkage and selection operator (LASSO), minimal redundancy maximal relevance (MRMR), and recursive feature elimination (RFE)) were performed in the same dataset to test our decision to use tumour size, mitotic count, and tumour site to train the OCTs.[23–29] The candidate features included tumour size, mitotic count, tumour site, demographic features (age and sex), histology subtype (spindle vs epithelioid vs mixed), and resection margin status (R0 vs R1).

OFS is a state-of-the-art methodology that has been shown to outperform similar approaches and results in a model that is simpler, more interpretable, and more accurate.[27–29] The SHAP method determines the average contribution of each factor to the model’s output by calculating the respective SHAP value.[23,27] The magnitude of a given SHAP value reflects the importance of the predictor in question. We opted to use XGBoost to predict recurrence because as a “black box” model, it should have superior performance over a transparent, decision tree model such as the OCT. This is supported by the well-described trade-off that exists between performance and interpretability.[30] In addition to the OFS and SHAP analyses, three other methodologies were used to select the features that predict recurrence–the LASSO estimator, the MRMR feature selection framework, and the RFE.

We used 5-fold cross-validation to tune the OCT hyperparameters. The hyperparameters used in the grid search were minbuckets (minimum number of patients included in each leaf of the tree) of 5%, 10%, and 15%, maximum depths of the tree of 3–5, and scoring criteria between gini and entropy. We trained models for every possible combination of these hyperparameters, which were determined from prior experience and familiarity with OCTs.

Evaluation of the OCT performance

There are two important properties of any prognostic model. First, the AUC or concordance index, which reflects discriminatory ability, has values that range from 0.5 (performance is equivalent to random prediction) to 1.0 (perfect discrimination). An AUC of 0.8 means that 80% of the time, the model will correctly assign a higher risk to a randomly selected patient with an event than to a randomly selected patient without an event. An AUC of 0.8–0.9 indicates excellent performance.[31,32]

Second, calibration refers to the degree of agreement between predicted probabilities and the rate of the actual outcome. A calibration plot, which is typically drawn by splitting the external validation cohort into at least five and ideally ten groups (deciles), should be close to the 45-degree line if the predictions are well calibrated. Alternatively, we can first discretise our model predictions into four interval bins (recurrence risk of 0–25%, 25–50%, 50–75%, and >75%) and calculate the average predicted probability and rate of the actual outcome of each bin. It has been previously suggested that the sample size for calibration assessment should be 400 cases with at least 100 and ideally 200 events.[33–35] Our external validation cohort included 556 cases with

173 events (recurrences). We also calculated and reported the calibration slope and intercept. Ideally, the calibration slope is 1 and the intercept is 0. An intercept of up to 0.4 is considered excellent.[31]

Finally, we evaluated the clinical utility of the OCT and the MSK nomogram by performing a Decision Curve Analysis (DCA).[36,37] Unlike traditional accuracy measures, DCA is a statistical method that determines the net benefit of a model in comparison to a competing model (e.g., MSK nomogram) and the two default strategies of treat all patients and treat no patients. DCA is most useful when there is no consensus on a single risk threshold, as it “allows one to examine risk model performance across a range of plausible risk thresholds”.[38] Thus, DCA may be ideal for GIST models that predict recurrence since there is no strict threshold of recurrence risk that mandates adjuvant imatinib but rather a range of plausible thresholds. Specifically, the ESMO-EURACAN Clinical Practice Guidelines and a recent expert review published in Nature suggest that a recurrence risk of less than 25% should dissuade a provider from offering adjuvant imatinib to a patient, whereas a risk of 50% and higher would mandate offering adjuvant imatinib.[39,40] However, the appropriate recommendation for patients with a recurrence risk between 25 and 50% is less clear.

Selection of the optimal OCT

The OCT with the highest AUC in both the training and internal validation (5-fold cross validation) cohort and that minimised overfitting was selected. The calibration of the OCT was subsequently validated in the entire external validation cohort, as well as in the subset of the external validation cohort with predicted intermediate/high recurrence risk according to the AFIP-Miettinen criteria.[7] Specifically, based on recommendations proposed by Miettinen et al., AFIP groups 3b–6b and jejunal and ileal 3a were included in the intermediate and high-risk categories.[7] Additionally, per the same recommendations, less common lower intestinal sites were grouped similarly to small intestinal GISTs.[7]

Another sub-analysis was performed in the subset of patients with tumours of known *KIT* and platelet-derived growth factor receptor alpha (*PDGFRA*) mutational status (eMethods).

Computing and statistical analysis

This study followed the STROBE reporting guidelines (Supplementary appendix). Continuous variables were presented as medians with interquartile ranges (IQR) and categorical variables as counts and percentages. Categorical variables were compared with the chi-square test, whereas continuous variables were compared with the Mann–Whitney U test. The OCT models were trained by using the Julia and Python programming languages. The Kaplan–Meier method was used for univariable survival analysis, and median follow-up was calculated using the reverse Kaplan–Meier method. The evaluation of the discrimination and calibration of the OCT and the MSK nomogram is described above. We also calculated the Harrell’s C-index, defined as the proportion of observations that the model can correctly order in terms of survival times. When censoring is present, the C-index has the added benefit of only including those patient pairs for which valid comparisons can be made. Comparisons between the OCT and the MSK nomogram model were performed with the rcorrp.cens function in the Hmisc package in R. We further evaluated the calibration of the two models by performing the Hosmer-Lemeshow test. Finally, we calculated the Brier score for the OCT and the MSK nomogram. The Brier score is an overall measure of the quality of predictions and is simultaneously influenced by both discrimination and calibration, with smaller values indicating superior model performance. All statistical analyses were conducted using R 5.3.0 (cran.r-project.org) and computing was performed on the MIT Sloan School of Management’s remote Engaging cluster (https://github.com/cran/iai/blob/master/R/optimaltrees.R).

Role of the funding source

The funding organisations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Results

Internal training cohort

The demographic and clinicopathologic features of the 395 patients included in the MSK- Spanish Group internal cohort are presented in Table 1. The median age was 63 years and 54.2% of patients were male (n = 214). Fifty-five percent of tumours were located in the stomach (n = 218). The median diameter of the tumours was 7 cm, and the median mitotic count was four mitoses per 5 mm^2. Genetic testing was performed in 81.3% of all tumours. Of those, 42.4% harbored a *KIT* exon 11 deletion (n = 136), with the second most common mutation being other isoforms of the *KIT* exon 11 mutation. The number of patients who had AFIP low-risk, intermediate-risk, and high-risk disease was 153, 88, and 154, respectively. There were 147 events (37.2%) and the 5-year recurrence-free survival (RFS) rate was 65% (60–70%) for the Spanish cohort and 56% (46–67%) for the MSK cohort. The median follow-up time was 106 (82–135) months.

External validation cohort

The demographic and clinicopathologic features of the 556 patients included in the external validation cohort are presented and compared to those of the patients included in the internal training cohort in Table 1. A

Numerical features (median [IQR])	Memorial Sloan Kettering (MSK)-Spain	Poland	P-value
	(n = 395)	(n = 556)	
Age, years	63.0 (54–71)	60.0 (52–68)	<0.001
Tumour size, cm[a]	7.0 (5–11)	6.0 (4–9.5)	<0.001
Mitotic count[b]	4.0 (1–11)	3.0 (1–9)	0.126
Categorical features			
Sex			0.741
Male	214 (54.2%)	308 (55.4%)	
Female	180 (45.6%)	248 (44.6%)	
Tumour site[c]			0.371
Gastric	218 (55.2%)	319 (57.4%)	
Non-gastric	177 (44.8%)	237 (42.6%)	
Mutation category			
KIT exon 9	17 (4.3%)	16 (2.9%)	
KIT exon 11 deletion	112 (28.4%)	87 (15.6%)	
KIT exon 11 other	102 (25.8%)	61 (11%)	
KIT exon 13	5 (1.3%)	2 (0.4%)	
KIT exon 17	0 (0.0%)	1 (0.2%)	
KIT multiple exons	24 (6.1%)	0 (0.0%)	
PDGFRA D842V or D842I	17 (4.3%)	19 (3.4%)	
PDGFRA other	14 (3.5%)	10 (1.8%)	
Wild type (WT)	21 (5.3%)	31 (5.5%)	
Not tested	74 (18.7%)	329 (59.3%)	
NF1 only	0 (0.0%)	0 (0.0%)	
SDH	0 (0.0%)	0 (0.0%)	
BRAF	9 (2.3%)	0 (0.0%)	
Armed Forces Institute of Pathology (AFIP) category			0.077
Low-risk	153 (38.7%)	253 (45.5%)	
Intermediate-risk	88 (22.3%)	101 (18.2%)	
High-risk	154 (39%)	202 (36.3%)	
TNM AJCC staging system[d] (8th edition)			0.070
I	152 (38.5%)	250 (45.7%)	
II	89 (22.5%)	102 (18.6%)	
IIIA	44 (11.2%)	69 (12.6%)	
IIIB	110 (27.8%)	126 (23%)	

The Spanish cohort had 309 patients (147 females, 161 males, and 1 unknown), while the MSK cohort had 86 patients (33 females and 53 males). No significant difference in gender was identified on chi square test (P = 0.265). Similarly, no significant difference in age distribution was identified on Mann-Whitney U test between the Spanish (median age = 63 (IQR: 54–71)) and MSK cohorts (median age = 64 (IQR: 52–71)) (P = 0.715). [a]Tumour size was ascertained in cm using the greatest diameter. [b]Mitotic count was determined to reflect per 50 HPF. [c]The primary tumour site was categorised as gastric vs non-gastric per the modified National Institutes of Health (NIH) and AFIP-Miettinen criteria; the exact site of origin was also ascertained. [d]TNM staging could not be calculated for all patients in the Polish cohort.

Table 1: **Demographic information and clinicopathologic variables of the internal training and external validation cohorts.**

significant difference in median age and tumour size was noted between the two cohorts. The median age was 60 years, and 55.4% of patients were male (n = 308). Fifty-seven percent of tumours were located in the stomach (n = 319). The median diameter of the tumours was 6 cm, and the median mitotic count was four mitoses per 5 mm^2. Genetic testing was performed in 40.8% of all tumours. Of those, more than one-third harbored a *KIT* exon 11 deletion (n = 87, 38.3%), with the second most common mutation being other isoforms of the *KIT* exon 11 mutation. The number of patients who had AFIP low-risk, intermediate-risk, and high-risk disease was 253, 101, and 202, respectively. There were 173 events (31.1%) and the 5-year RFS rate was 62% (58–67%). The median follow-up time was 61 (29–91) months.

Feature selection

Importantly, mitotic count and tumour size were selected as the two most important features by all five feature selection analyses, which supports our decision to use them to train the OCTs (eTable 1). The decision to use tumour site as the third variable was corroborated by the fact that not only was there no agreement among the five feature selection analyses with regard to the third most important feature, but tumour site was also included in most of the five current risk score systems (the original NIH criteria, the modified NIH criteria, the AFIP-Miettinen criteria, the Joensuu contour maps, and the MSK nomogram).[3–7] In turn, this would allow for a fair comparison between the OCT and the other prognostic models since the latter also include tumour size and mitotic count.

OCT structure

The OCT with the highest AUC in both the training and internal validation (5-fold cross validation) cohorts and with the least overfitting was an OCT with an AUC of 0.84 (0.80–0.87) in the training cohort and 0.777 in the 5-fold cross-validation. Fig. 2 illustrates the structure of this OCT, which assigned patients to eight distinct subgroups, each with a unique probability of recurrence within 5 years after GIST resection. The number of patients whose data were used to derive the eight predictions are reported in the respective OCT leaf. A calculator to enable rapid determination of recurrence risk based on the OCT model depicted in Fig. 2 is available online at: https://alexandriahealth.gitlab.io/apps/gastrointestinal-stromal-tumor/recurrence.html.

The ESMO-EURACAN Clinical Practice Guidelines and a recent expert review published in Nature suggest that a recurrence risk of less than 25% should dissuade a provider from offering adjuvant imatinib to a patient, whereas a risk of 50% and higher should mandate offering adjuvant imatinib.[39,40] We merged the eight subgroups into three groups based on the aforementioned cut-offs. The KM curves of the three groups separate nicely as shown in eFigure 1, and the pair wise log rank tests were all statistically significant (P < 0.05). However, given that the appropriate recommendation for patients with a recurrence risk between 25 and 50% is less clear, we recommend that readers use the original OCT, which can provide a more accurate estimate of recurrence risk and better facilitate shared physician-patient decision making. For example, a patient and/or provider may decide that 57.1% (Group B of the OCT) is a

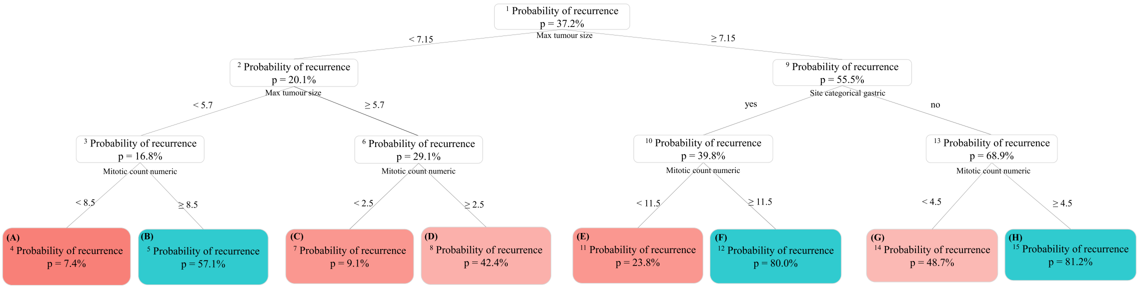

1 Probability of recurrence
p = 37.2%
Max tumour size
< 7.15
≥ 7.15
2 Probability of recurrence
p = 20.1%
Max tumour size
9 Probability of recurrence
p = 55.5%
Site categorical gastric
< 5.7
≥ 5.7
yes
no
3 Probability of recurrence
p = 16.8%
Mitotic count numeric
6 Probability of recurrence
p = 29.1%
Mitotic count numeric
10 Probability of recurrence
p = 39.8%
Mitotic count numeric
13 Probability of recurrence
p = 68.9%
Mitotic count numeric
< 8.5
≥ 8.5
< 2.5
≥ 2.5
< 11.5
≥ 11.5
< 4.5
≥ 4.5
(A)
4 Probability of recurrence
p = 7.4%
(B)
5 Probability of recurrence
p = 57.1%
(C)
7 Probability of recurrence
p = 9.1%
(D)
8 Probability of recurrence
p = 42.4%
(E)
11 Probability of recurrence
p = 23.8%
(F)
12 Probability of recurrence
p = 80.0%
(G)
14 Probability of recurrence
p = 48.7%
(H)
15 Probability of recurrence
p = 81.2%

Fig. 2: Optimal classification tree to predict recurrence. Footnote: For tumours less than 7.15 cm, tumour site was not accounted for by the algorithm; thus, for these tumours, we should clarify that the probability of recurrence is reflective of mixed gastric and non-gastric site tumours. This is in contrast to tumours equal to or larger than 7.15 cm, for which the probability of recurrence was influenced by tumour site.

risk high enough to warrant receiving imatinib while 23.8% (Group E of the OCT) is not.

The OCT that was trained in the subset of patients with known *KIT* and *PDGFRA* status is illustrated in eFigure 2 and presented in the eResults section.

OCT performance

The concordance index of the main OCT in the external validation cohort was greater than that of the MSK nomogram (0.805 (95% CI: 0.803–0.808) vs 0.788 (95% CI: 0.786–0.791), respectively). In the external validation cohort, the slope and intercept of the calibration curve of the main OCT were excellent at 1.041 and 0.038, respectively. The curve was close to the 45-degree line, indicating excellent calibration (Fig. 3A). In comparison, the slope and intercept of the calibration curve of the MSK nomogram were 0.681 and 0.032, respectively, and an overestimation of the risk of recurrence throughout the entire calibration curve was visually apparent (Fig. 3B). The calibration plots that discretise OCT and nomogram predictions into four interval bins (recurrence risk of 0–25%, 25–50%, 50–75%, and >75%) were consistent with the main plots (Fig. 4A and B).

The calibration of the main OCT was also tested in the subset of patients in the external validation cohort with intermediate/high risk of recurrence according to the AFIP-Miettinen criteria and compared to that of the MSK nomogram. The slope and intercept of the calibration curve of the main OCT were 0.921 and 0.069, respectively. The curve was relatively close to the 45-degree line, indicating good calibration (eFigure 3A). In comparison, the slope and intercept of the calibration curve of the MSK nomogram were 0.633 and 0.023, respectively, and an overestimation of the risk of recurrence throughout the entire calibration curve was visually apparent (eFigure 3B). Of note, the Hosmer-

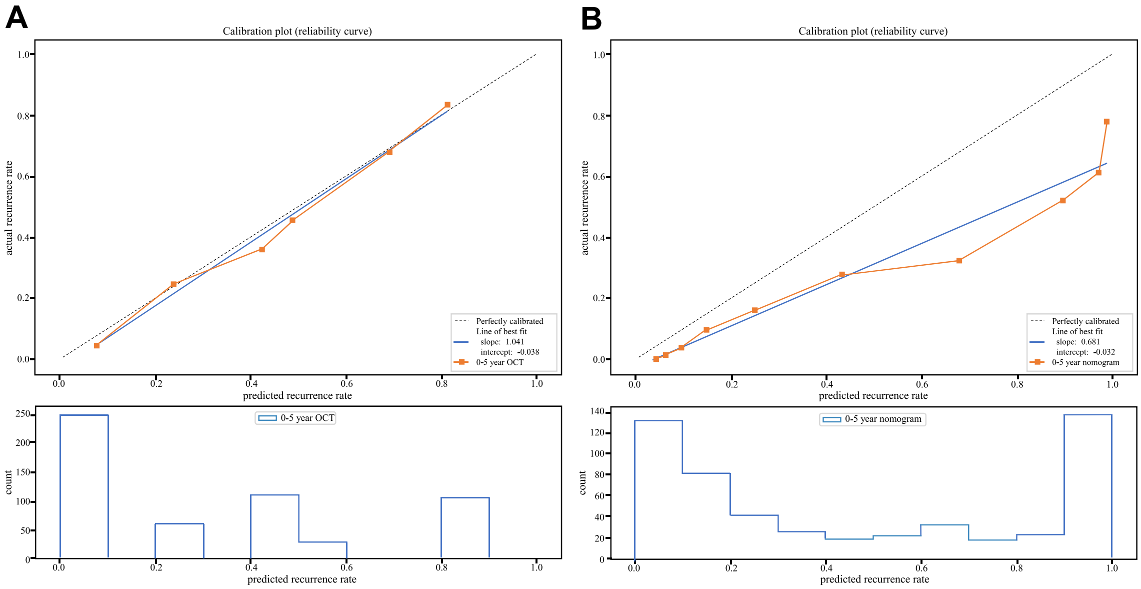


Fig. 3: Assessment of calibration of (A) the OCT (B) the Memorial Sloan Kettering (MSK) nomogram in the external validation cohort.

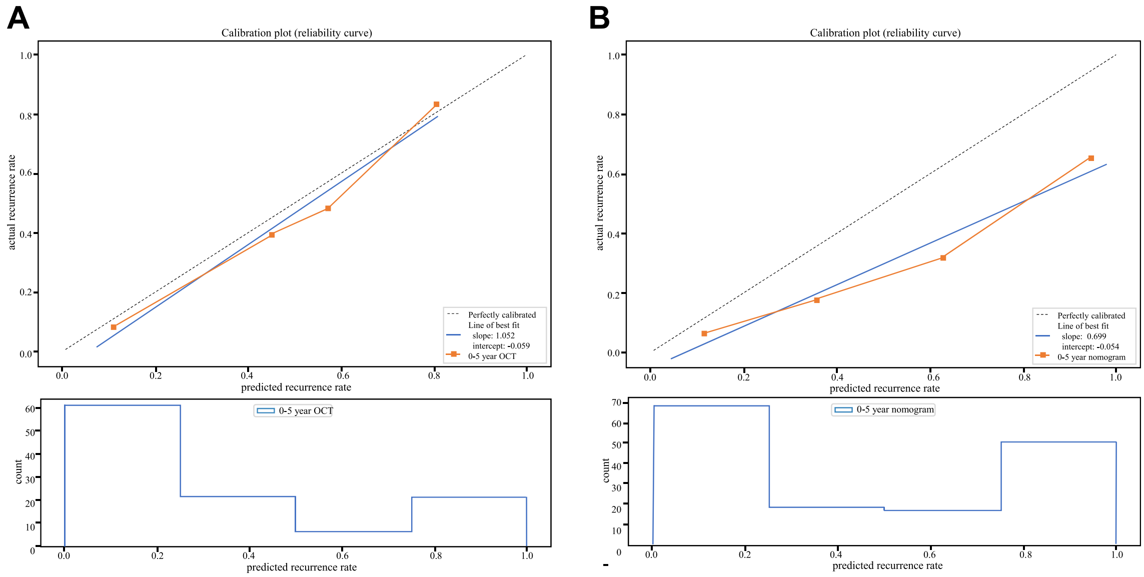


Fig. 4: Assessment of calibration of (A) the OCT (B) the MSK nomogram in the external validation cohort using the discretised OCT and nomogram predictions.

Lemeshow test was insignificant (P = 0.087) for the OCT model but significant (P < 0.001) for the MSK nomogram. We also compared the OCT predictions to the actual events in the external validation cohort. As shown in eFigure 4, the rates are remarkably consistent. We also calculated the Brier score for the OCT and the MSK nomogram (0.147 vs 0.564, respectively) which confirmed the superior discrimination and calibration of the OCT model compared to the MSK nomogram.

Finally, as shown in eFigure 5, the decision curve for the OCT model dominated the curve for the MSK nomogram and thus allowed for superior decision making compared to the MSK nomogram for both patients at 25–50% recurrence risk as well as those with >50% recurrence risk.

The performance of the OCT that was trained in the subset of patients with known *KIT* and *PDGFRA* status is reported in the eResults section and in eFigure 6. Finally, the main OCT outperformed all three ML algorithms (eResults and eFigure 7).

Discussion

In this study, we report on the first interpretable AI-based algorithm to predict recurrence risk following GIST resection. The algorithm utilised the same predictors of recurrence used by most existing models to allow for a fair comparison across different methodologies and has three main advantages. First, it had excellent calibration in an external cohort. The fact that the external cohort consisted of registry data from geographically distinct hospitals with high heterogeneity attests to the generalisability of the algorithm's predictions. In comparison, the original NIH criteria, the AFIP criteria by Miettinen et al., and the modified NIH criteria by Joensuu et al. can only make qualitative estimates. Thus, we cannot assess whether these predictions are well calibrated. The Joensuu contour or heat maps demonstrate ranges of probabilities of recurrence and not a specific probability for a given patient, and thus their calibration cannot be assessed.[5] Of note, one can calculate an estimated probability of recurrence for a given patient using contour and heat maps as long as an online tool is provided, but their associated online calculator was taken down in 2020 for unknown reasons.

In contrast, the MSK nomogram has undergone validation of its calibration, but a study from Japan by Tanimine et al. reported poor calibration.[41] Specifically, the authors found that the nomogram overestimated the recurrence risks along the entire length of the calibration curve. A group from Singapore similarly reported on poor calibration of the MSK nomogram.[42] A group from Canada that validated the MSK nomogram also reached similar conclusions; in their cohort, the observed recurrence rates were significantly lower than those predicted by the MSK nomogram.[43] Finally, an Italian group corroborated the aforementioned studies and also reported an overestimation of the risk of recurrence.[44] Ultimately, these results are consistent with our study findings, as the MSK nomogram overestimated the actual risk of recurrence across all patient groups.

Second, the OCT's high Harrell's C-index that was retained on external validation (0.80) places it among the best models that predict recurrence following GIST resection. Joensuu et al. reported that the AUCs in their validation series were 0.80 for their contour maps, 0.76

for the NIH consensus criteria, 0.76 for the modified NIH criteria, and 0.77 for the AFIP-Miettinen criteria.[5] Gold et al. reported that the AUC of their MSK nomogram ranged between 0.76 and 0.80 in the two validation cohorts, while the AUC of the original and modified NIH criteria ranged between 0.66 and 0.78 in the two validation cohorts.[6] An external validation by Chok et al. concluded that the MSK nomogram and AFIP criteria had superior predictive accuracy for tumour recurrence compared to the original and modified NIH criteria.[42]

Third, the algorithm may address a long-standing problem in GIST prognostication, which is the high heterogeneity of the intermediate/high-risk groups for the most commonly used GIST prognostic models. For example, an external validation by Gold et al. showed that the intermediate/high-risk groups not only had very heterogeneous outcomes, but they even included patients with excellent prognosis (predicted 5-year RFS of 90–100%).[6] This deficiency has been acknowledged by others. Specifically, in a systematic review of GIST prognostication systems, Khoo et al. concluded that the Joensuu modified NIH criteria have “marked prognostic heterogeneity” in its high-risk group and the AFIP-Miettinen criteria also have “extreme prognostic heterogeneity” in its high-risk group.[45] Our algorithm appears to remedy this deficiency as it can accurately predict the probability of recurrence even among patients with intermediate/high-risk according to the AFIP-Miettinen criteria.

The difference in accurate prognostication between the MSK nomogram and the OCT can be further highlighted by examining individual patient cases. For example, an 8 cm gastric GIST with a mitotic count of five mitoses per 5 mm^2 would be assigned an 88% probability of recurrence at five years by the MSK nomogram, while the OCT model predicts only a 24% probability of recurrence (eFigure 8A). On external validation, the actual rate of recurrence for this group in the Polish cohort was 24.1%. If prediction of recurrence risk was based on the MSK nomogram, this patient would be deemed very high risk and offered adjuvant imatinib; in reality, the risk of recurrence was much lower at 24.1%, which was very close to the OCT prediction of 24% and also lower than the 30–50% risk of recurrence threshold for offering imatinib.[39,40,46,47] Another example is a 4 cm GIST with a mitotic count of five mitoses per 5 mm^2. The MSK nomogram predicts a recurrence rate of 66% and 98% for a gastric and a small bowel GIST, respectively, at five years. In contrast, the OCT model predicts a 7% risk of recurrence (eFigure 8B). On external validation, the actual rate of recurrence for this group in the Polish cohort was 3.5%. Ultimately, an overestimation of recurrence risk is clinically important as it can lead to overtreatment.

We also evaluated the clinical utility of the OCT and the MSK nomogram by performing a Decision Curve Analysis. DCA is a statistical method that unlike traditional accuracy measures evaluates the net benefit of a model in comparison to a competing model (e.g., MSK nomogram). It is most useful when there is no consensus on a single risk threshold, as it allows one to examine model performance across a range of plausible risk thresholds. Thus, DCA may be ideal for GIST models that predict recurrence since there is no strict threshold of recurrence risk that mandates adjuvant imatinib but rather a range of plausible thresholds (25–50%). Notably, DCA provided a clear answer to the question about which of OCT and MSK nomogram would lead to better clinical outcomes on average among patients with resected GIST. Specifically, the OCT had a higher net benefit than the MSK nomogram. Interestingly, according to a recent systematic review on clinical prediction modeling, only 1% of all studies that reported on ML-based models evaluated their clinical utility using DCA.[48] Thus, this study is one of the very few to report on the use of DCA in ML-based models for clinical prediction modeling.

Future studies with larger cohorts may succeed in developing more granular OCTs than the ones we presented, as more data can allow for deeper OCTs without the risk of overfitting. In turn, this can smoothen the abrupt changes currently present between some OCT groups. However, it should be noted that such abrupt changes are common in all current prognostication models. For example, the MSK nomogram predicts a 5-year RFS of 87% for a typical patient with a 7 cm gastric GIST with four mitoses per 5 mm^2, but a 5-year RFS of only 16% for a similar case with five mitoses per 5 mm^2.[49] Similarly, the Joensuu contour maps predict a 10-year RFS of 60–80% for a typical patient with a 5 cm gastric GIST with nine mitoses per 5 mm^2, but a 10-year RFS of only 40–60% for a similar case with ten mitoses per 5 mm^2.[5] Nonetheless, this limitation does not appear to restrict the external use of the OCT calculator, since a calibration analysis in an independent, external cohort demonstrated that the OCT predictions are accurate. Finally, a potential limitation of the study is that while tumour size was ascertained by dedicated, highly experienced pathologists and not by imaging, we did not assess for interobserver variation in determining tumour size.

The OCT has one of the highest concordance indices ever reported on an external validation of a GIST prognostic model and is the only GIST prognostic model that has excellent calibration on external validation. The algorithm that incorporated *KIT* (eDiscussion) also had excellent calibration and yielded intuitive stratifications of *KIT* variants. It is important that the statistical properties of the GIST OCT were successfully validated in a cohort with great heterogeneity with regard to local treatment protocols, level of experience, case volume, and other pertinent factors, as this attests to the algorithm’s generalisability.[50] We attribute the excellent performance of the models to the novel OCT methodology as we used the same predictors of recurrence as previously published prognostic models. Specifically, unlike the current risk score systems which assume that variables interact in a linear and additive fashion, in the OCT methodology, some variables gain or lose significance due to the absence or presence of other variables. Finally, in contrast to most AI-driven models such as deep learning, OCTs have a tree structure that allows the reader to understand which prognostic factors and which specific cut-offs were used to calculate the probability of recurrence for a given patient. This is particularly important in GIST because these predictions can be used to guide patient selection for adjuvant imatinib, and clinicians tend to distrust predictions that derive from "black box" methods.[51]

Contributors

Conception and design: DB, GAM and SS.

Acquisition, analysis or interpretation of data: DB, GAM, ST, AK, CRA, MFB, JMB, PR, GS, JW, EP, EB, PS, AG, BJ, WDT, PC, SS.

Drafting of the manuscript: GAM, DB and SS.

Critical revision of the manuscript for important intellectual content: DB, GAM, ST, AK, CRA, MFB, JMB, PR, GS, JW, EP, EB, PS, AG, BJ, WDT, PC, SS.

Statistical analysis: DB, GAM, ST.

Administrative, technical, or material support: BJ.

Supervision: SS.

Final approval of the version to be published: DB, GAM, ST, AK, CRA, MFB, JMB, PR, GS, JW, EP, EB, PS, AG, BJ, WDT, PC, SS.

GAM, DB and SS accessed and verified the underlying data and were responsible for the decision to submit the manuscript.

Data sharing statement

The datasets used in the current study are not publicly available but may be available from the corresponding author on reasonable request depending on the policy and procedures of the institutions that participate in the consortiums.

Declaration of interests

JMB reports personal medical consulting fees from PharmaMar, GSK, Novartis, Amgen, Bayer, Roche, Lilly, Tecnofarma, Asofarma, Boehringer Ingelheim, support for attending meetings from Pfizer, PharmaMar, grants to his institution from GSK, PharmaMar, Novartis, EISAI, Lilly, Bayer, Lixte Biotechnology, Karyopharm Therapeutics, Deciphera, Blueprint Medicines, Nektar, Forma therapeutics, Amgen, Daiichi Sankyo, Immix BioPharma, BMS, Pfizer, Celgene, Arog, Adaptimmune, Rain Therapeutics, InnibRx, Ayala Pharmaceuticals, Philogen, Cebiotex, PTC Therapeutics, Springworks Therapeutics, and is on the Boards for TRACON PHARMA, PHARMAMAR, BOERHINGER, outside the submitted work.

PC reports grants to her institution from Pfizer/Array, Deciphera, Ningbo NewBay, consulting fees from Deciphera, Ningbo NewBay, and is on the Advisory board and Steering Committee for Ningbo NewBay, and on the Steering Committee for Deciphera (unpaid), outside the submitted work.

PR reports consulting fees from Bristol Myers Squibb, Merck Sharp & Dohme, Novartis, Pierre Fabre, Sanofi, Merck, Philogen and Blueprint Medicine, payment or honoraria for lectures, presentations, speakers' bureaus, manuscript writing or educational events from Bristol Myers Squibb, Merck Sharp & Dohme, Novartis, Pierre Fabre, Sanofi, Merck, Astra Zeneca, Philogen and Blueprint Medicine, outside the submitted work.

PS reports payment or honoraria for lectures, presentations, speakers' bureaus, manuscript writing or educational events from BMS, Gillead, support for attending meetings and/or travel from Novartis, BMS, MSD, is on the Advisory Board for Sandoz, is a Committee Member of the European Society of Medical Oncology and a Board Member of the Polish Society of Clinical Oncology, owns Celon Pharma stocks, and received drugs for noncomercial clinical trial from Immutep, outside the submitted work.

WDT reports personal fess from Eli Lilly, EMD Serono, Mundipharma, C4 Therapeutics, Daiichi Sankyo, Deciphera, Adcendo, Ayala Pharmaceuticals, Kowa, Servier, Bayer Pharmaceuticals, Epizyme, Cogent, Medpacto, Foghorn Therapeutics, Amgen, AmMax Bio, Boehringer Ingelheim, BioAtla, Inhibrx. In addition, WDT has a patent Companion Diagnostic for CDK4 inhibitors—14/854,329 pending to MSKCC/SKI, and a patent Enigma and CDH18 as companion Diagnostics for CDK4 inhibition—SKI2016-021-03 pending to MSKCC/SKI, outside the submitted work. WDT is on the Scientific Advisory Boards for Certis Oncology Solutions and Innova Therapeutics and owns Certis Oncology Solutions and Atropos Therapeutics stocks.

JW reports grants to her institution from the UCSF Noyce Initiative for Digital Transformation in Computational Biology & Health, Computational Innovator Postdoctoral Fellowship Award.

All other authors declare no competing interests.

Acknowledgements

The authors would like to thank Dr. Nefeli Bampatsikou for the excellent technical support.

This research was supported by the NCI SPORE in Soft Tissue Sarcoma P50 CA140146 (SS, WT, PC, CA) and NCI Cancer Center Support Grant P30 CA008748 (institutional).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102200.

References

1 DeMatteo RP, Lewis JJ, Leung D, Mudan SS, Woodruff JM, Brennan MF. Two hundred gastrointestinal stromal tumors: recurrence patterns and prognostic factors for survival. *Ann Surg.* 2000;231:51–58.

2 Eisenberg BL, Judson I. Surgery and imatinib in the management of GIST: emerging approaches to adjuvant and neoadjuvant therapy. *Ann Surg Oncol.* 2004;11:465–475.

3 Fletcher CD, Berman JJ, Corless C, et al. Diagnosis of gastrointestinal stromal tumors: a consensus approach. *Hum Pathol.* 2002;33:459–465.

4 Joensuu H. Risk stratification of patients diagnosed with gastrointestinal stromal tumor. *Hum Pathol.* 2008;39:1411–1419.

5 Joensuu H, Vehtari A, Riihimaki J, et al. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *Lancet Oncol.* 2012;13:265–274.

6 Gold JS, Gonen M, Gutierrez A, et al. Development and validation of a prognostic nomogram for recurrence-free survival after complete surgical resection of localised primary gastrointestinal stromal tumour: a retrospective analysis. *Lancet Oncol.* 2009;10:1045–1052.

7 Miettinen M, Lasota J. Gastrointestinal stromal tumors: pathology and prognosis at different sites. *Semin Diagn Pathol.* 2006;23:70–83.

8 Goh BK, Chow PK, Yap WM, et al. Which is the optimal risk stratification system for surgically treated localized primary GIST? Comparison of three contemporary prognostic criteria in 171 tumors and a proposal for a modified Armed Forces Institute of Pathology risk criteria. *Ann Surg Oncol.* 2008;15:2153–2163.

9 Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA.* 2017;318:1377–1384.

10 Bertsimas D, Dunn J. Optimal classification trees. *Mach Learn.* 2017;106:1039–1082.

11 Breiman L, Freidman JH, Olshen RA, Stone CJ, eds. *CART: classification and regression trees.* 1984.

12 Dunn J, Mingardi L, Zhuo YD. Comparing interpretability and explainability for feature selection. https://arxiv.org/abs/2105.05328.

13 Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive OpTimal trees in emergency surgery risk (POTTER) calculator. *Ann Surg.* 2018;268:574–583.

14 Bertsimas D, Margonis GA, Huang Y, et al. Toward an optimized staging system for pancreatic ductal adenocarcinoma: a clinically interpretable, artificial intelligence-based model. *JCO Clin Cancer Inform*. 2021;5:1220–1231.

15 Gebran A, Vapsi A, Maurer LR, et al. POTTER-ICU: an artificial intelligence smartphone-accessible tool to predict the need for intensive care after emergency surgery. *Surgery*. 2022;172:470–475.

16 El Hechi M, Gebran A, Bouardi HT, et al. Validation of the artificial intelligence-based trauma outcomes predictor (TOP) in patients 65 years and older. *Surgery*. 2022;171:1687–1694.

17 Bertsimas D, Margonis GA, Sujichantararat S, et al. Using artificial intelligence to find the optimal margin width in hepatectomy for colorectal cancer liver metastases. *JAMA Surg*. 2022;157(8):e221819.

18 Bertsimas D, Zhuo D, Levine J, et al. Benchmarking in congenital heart surgery using machine learning-derived optimal classification trees. *World J Pediatr Congenit Heart Surg*. 2022;13:23–35.

19 Bertsimas D, Wiberg H. Machine learning in oncology: methods, applications, and challenges. *JCO Clin Cancer Inform*. 2020;4: 885–894.

20 El Hechi MW, Maurer LR, Levine J, et al. Validation of the artificial intelligence-based predictive optimal trees in emergency surgery risk (POTTER) calculator in emergency general surgery and emergency laparotomy patients. *J Am Coll Surg*. 2021;232: 912–919.e1.

21 Maurer LR, Chetlur P, Zhuo D, et al. Validation of the AI-based predictive OpTimal trees in emergency surgery risk (POTTER) calculator in patients 65 Years and older. *Ann Surg*. 2020;277(1): e8–e15.

22 Bertsimas D, Dunn J, Steele DW, Trikalinos TA, Wang Y. Comparison of machine learning optimal classification trees with the pediatric emergency care applied research network head trauma decision rules. *JAMA Pediatr*. 2019;173:648–656.

23 Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2:749–760.

24 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58:267–288.

25 Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinf Comput Biol*. 2005;3:185–205.

26 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389–422.

27 Bertsimas D, Pauphilet J, Van Parys B. *Sparse regression: scalable algorithms and empirical performance*. 2020.

28 Bertsimas D, Pauphilet J, Van Parys BJML. Sparse classification: a scalable discrete optimization perspective. *Mach Learn*. 2021;110:3177–3209.

29 Bertsimas D, Van Parys B. *Sparse high-dimensional regression: exact scalable algorithms and phase transitions*. 2020.

30 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17:195.

31 Staartjes VE, Kernbach JM. Foundations of machine learning-based clinical prediction modeling: Part III-model evaluation and other points of significance. *Acta Neurochir Suppl*. 2022;134: 23–31.

32 Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5:1315–1316.

33 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35:214–226.

34 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–176.

35 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58:475–483.

36 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565–574.

37 Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *Am Stat*. 2008;62:314–320.

38 Kerr KF, Brown MD, Zhu K. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol*. 2016;34:2534–2540.

39 Blay JY, Kang YK, Nishida T, von Mehren M. Gastrointestinal stromal tumours. *Nat Rev Dis Primers*. 2021;7:22.

40 Casali PG, Abecassis N, Aro HT, et al. Gastrointestinal stromal tumours: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2018;29:iv267.

41 Tanimine N, Tanabe K, Suzuki T, Tokumoto N, Ohdan H. Prognostic criteria in patients with gastrointestinal stromal tumors: a single center experience retrospective analysis. *World J Surg Oncol*. 2012;10:43.

42 Chok AY, Goh BK, Koh YX, et al. Validation of the MSKCC gastrointestinal stromal tumor nomogram and comparison with other prognostication systems: single-institution experience with 289 patients. *Ann Surg Oncol*. 2015;22:3597–3605.

43 Racz JM, Brar SS, Cleghorn MC, et al. The accuracy of three predictive models in the evaluation of recurrence rates for gastrointestinal stromal tumors. *J Surg Oncol*. 2015;111:371–376.

44 Belfiori G, Sartelli M, Cardinali L, et al. Risk stratification systems for surgically treated localized primary Gastrointestinal Stromal Tumors (GIST). Review of literature and comparison of the three prognostic criteria: MSKCC Nomogramm, NIH-Fletcher and AFIP-Miettinen. *Ann Ital Chir*. 2015;86:219–227.

45 Khoo CY, Chai X, Quek R, Teo MCC, Goh BKP. Systematic review of current prognostication systems for primary gastrointestinal stromal tumors. *Eur J Surg Oncol*. 2018;44:388–394.

46 Casali PG, Le Cesne A, Velasco AP, et al. Final analysis of the randomized trial on imatinib as an adjuvant in localized gastrointestinal stromal tumors (GIST) from the EORTC Soft Tissue and bone sarcoma group (STBSG), the australasian gastro-intestinal trials group (AGITG), UNICANCER, French sarcoma group (FSG), Italian sarcoma group (ISG), and Spanish group for research on Sarcomas (GEIS). *Ann Oncol*. 2021;32:533–541.

47 Dematteo RP, Ballman KV, Antonescu CR, et al. Adjuvant imatinib mesylate after resection of localised, primary gastrointestinal stromal tumour: a randomised, double-blind, placebo-controlled trial. *Lancet*. 2009;373:1097–1104.

48 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.

49 Joensuu H. Predicting recurrence-free survival after surgery for GIST. *Lancet Oncol*. 2009;10:1025.

50 Staartjes VE, Kernbach JM. Significance of external validation in clinical machine learning: let loose too early? *Spine J*. 2020;20:1159–1160.

51 Towards trustable machine learning. *Nat Biomed Eng*. 2018;2:709–710.