Methods 220 (2023) 55–60

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Machine learning models based on routinely sampled blood tests can predict the presence of malignancy amongst patients with suspected musculoskeletal malignancy

Kieran Bentick [a], Joel Runevic [b], Sriram Akula [b], Theocharis Kyriacou [b], Paul Cool [a,b,*], Peter Andras [b]

[a] Robert Jones and Agnes Hunt Orthopaedic Hospital, Oswestry SY10 7AG, United Kingdom
[b] Keele University, Keele ST5 5BG, United Kingdom

ARTICLE INFO

Keywords:
Blood tests
Diagnosis
Cancer
Muscluloskeletal
Machine Learning

ABSTRACT

Aims: This study explores the possibility of using routinely taken blood tests in the diagnosis and triage of patients with suspected musculoskeletal malignancy.
Methods: A retrospective study was performed on results of patients who had presented for assessment to a regional musculoskeletal tumour unit. Blood results of patients with a histologically confirmed diagnosis between 2010 and 2020 were retrieved. 33 distinct blood tests were available for model forming. Results were standardised by calculating z-scores. Data were split into a training set (70%) and a test set (30%). The training set was balanced by resampling underrepresented classes. The random forest algorithm performed best and was selected for model forming. Receiver operating characteristic curves were used to find the optimum threshold. Models were calibrated and performance metrics evaluated with confusion tables.
Results: 2371 patients formed the study population. 1080 had a malignant diagnosis in one of three categories: sarcoma, metastasis, or haematological malignancy. 1291 had a benign condition. Metastasis could be predicted with an accuracy of 79% (AUC 87%, sensitivity 79%, specificity 80% NPV 91%). Haematological malignancy accuracy 79% (AUC 81%, sensitivity 77%, specificity 79%, NPV 97%). Sarcoma accuracy 64% (AUC 73%, sensitivity 76%, specificity 61%, NPV 88%) and all malignancy accuracy 74% (AUC 80%, sensitivity 72%, specificity 75%, NPV 76%).
Conclusion: Routinely performed blood tests can be useful in triage of musculoskeletal tumours and can be used to predict presence of musculoskeletal malignancy.

1. Introduction

Whether consciously or not a large part of medicine is stratification of risk. Much of our estimation of risk is based on clearly defined parameters, symptoms and clinical signs known to be associated with conditions of interest. This forms the basis of diagnostic medicine.

There are clinical and biochemical parameters used in initial evaluation of or investigation for disease that have less direct correlation with the disease, parameters such as a patient's age and blood tests including full blood count, inflammatory markers, clotting, renal, liver and bone profiles. These are considered in assessment of a patient's fitness in general, suitability for intervention, treatment planning and prognostication. These same parameters are often overlooked as being of diagnostic value particularly when they fall within what is considered the normal range for that result. These parameters can have more complex links with pathology in general and malignancy in particular when evaluated either in isolation or assessed in combination.

Previous studies have assessed this phenomenon in specific cancer types and useful decision trees or algorithms have been developed in reference to lung and ovarian cancer [1,2]. Similar techniques have demonstrated use of blood tests in prognostication in prostate [3] and breast [4] cancer and are employed to guide orthopaedic management of these conditions. The modified Glasgow prognostic score employs C-reactive protein and albumin to stratify prognosis in cancer irrespective of site and can be useful in guiding treatment in in patients with soft tissue sarcoma [5,6] and osteosarcoma [7,8]. Many of these models are

* Corresponding author.
E-mail address: paul.cool@nhs.net (P. Cool).

https://doi.org/10.1016/j.ymeth.2023.10.012
Received 31 July 2023; Received in revised form 9 October 2023; Accepted 26 October 2023
Available online 10 November 2023

tumour site specific and have limited roles in diagnosis or triage.

Pertaining to prediction of the presence disease and its magnitude there have been studies demonstrating relationship between the presence of cancer and elevated platelet count [9,10] and similar models have been developed and proved useful in prediction of outcome in non-cancerous conditions. This was shown in spinal cord injury where routinely measured blood parameters can usefully predict impairment following injury [11]. This has not however been tested in triaging or risk stratification of patients referred with potential musculoskeletal malignancy.

The aim of this study was to evaluate whether results from the blood tests that are ordinarily undertaken in the course of a patient's assessment, i.e. neither tumour specific nor specialised blood tests, could be used to predict the presence of cancer in patients referred with suspected musculoskeletal malignancy. Within this set of patients referred for assessment could there be demonstrable differences used to stratify cases into high risk and low risk groups which could be a consideration when choosing the referral pathway used for ultimate assessment.

2. Methods

This study forms part of a larger study evaluating the use of machine learning in the diagnosis of orthopaedic conditions approved by local ethics committee. Individual consent to use data was not required. All results were depersonalised and the data used for model forming only included blood test results, Diagnostic group (sarcoma / metastatic cancer / haematological cancer and benign) sex and age.

All patients referred our orthopaedic oncology unit with a suspected musculoskeletal tumour between 2010 and 2020 who had a confirmed histological result were included in this study. Patients without a histological diagnosis were excluded. All blood tests that were taken at the point of biopsy or resection of the tumour were retrieved from the hospital server. These were all routine blood tests undertaken as part of usual care patient care.

Uncommonly performed blood tests which were done on less than 10 % patients were excluded from analysis leaving 33 blood tests as demonstrated in Table 1. in addition to age and sex as parameters for evaluation. Results from plasma electrophoresis and urine tests were not included.

Processing of the data was done in Python using pandas [12] and NumPy [13] packages. Graphical representation employed Matplotlib [14]. Several open-source machine learning methods were evaluated using the scikit-learn package in Python [15]. Following in initial exploratory analysis, the random forest algorithm [16] proved best performing and was selected for model-forming.

Data was prepared for machine learning. Blood test results were transformed to z-scores allowing better comparison of features and avoidance of statistical leverage that use of raw values would have. When appropriate, sex-specific reference ranges were used. Only transformed z-scores were used for model building.

Missing data was managed with imputation. 20 % of the data contained missing values. There were no blood tests without missing values. There were 209 complete cases. Using only complete cases or removing values would have reduced the data set to a point of potentially introducing bias. Missing data was therefore imputed with the mean z-score (zero) to have no effect on model formation.

Each output group was selected in turn and least significant features from runs with the full feature set were dropped for evaluation of that group. The data was divided into a training set comprising 70 % and a test set comprising 30 % of the data. The partitions were validated to ensure appropriately stratified and that the test set was a good representation of the training set. Numerical attributes were scaled and the balance of data was checked. Imbalances within the dataset were balanced with resampling.

Imbalances within the dataset wer balanced with resampling. X and y parameters were converted to NumPy arrays. Hyper-parameter were tuned [17] and the model trained with the training-set using K-fold cross validation with k set as 5 and configured to optimise the area under the curve (AUC) of the receiver operator characteristic (ROC).

Precision-Recall curves and best thresholds were calculated based on sensitivity as well as ROC and best threshold based on Youden's J statistic [18]. Youden's J statistic was selected to base best threshold upon.

The trained model was applied to the test set and assessed on the unseen test set using best threshold value to compute confusion tables. Accuracy, precision (PPV), recall (sensitivity), specificity, F1 (harmonic mean of precision and recall), F2 (more weight on recall/sensitivity) and negative predictive values (NPV) were calculated.

Calibration curves were evaluated to avoid overfitting and to ensure appropriate models were developed. Feature values and importance were computed using the shapely additive explanations (ShAP) [19] package. Feature value plots were drawn to explain models.

Four different models were evaluated in diagnosis of malignant disease. These were malignancy (all patients with malignant disease), metastasis, haematological malignancy, and sarcoma.

Table 1
Blood tests used in model forming.

Adjusted Calcium (mmol/L)	Prothrombin Time (s)
Alanine Transaminase (international units/L)	Partial Thromboplastin Time (s)
Albumin (g/L)	Red blood count (10*12/L)
Alkaline Phosphatase (international units/L)	Sodium (mmol/L)
Basophils (10*9/L)	Total Protein (g/L)
Bilirubin (micromol/L)	Urea (mmol/L)
Calcium (mmol/L)	White blood count (10*9/L)
Creatinine (micromol/L)	C-Reactive Protein (mg/L)
Eosinophils (10*9/L)	Glomerular Filtration Rate (mL/min/1.7)
ESR (mm/h)	Red cell distribution width (%)
Gamma GT (international units/L)	
Haematocrit (L/L)	
Haemoglobin (g/L)	
International Normalised Ratio (INR)	
Lymphocytes (10*9/L)	
Magnesium (mmol/L)	
Mean Cell Haemoglobin (pg)	
Mean Cell Volume (fL)	
Monocytes (10*9/L)	
Neutrophils (10*9/L)	
Phosphate (mmol/L)	
Platelets (10*9/L)	
Potassium (mmol/L)	

3. Results

2371 patients formed the population of this study. 1080 patients (46 %) had a malignant tumour. 1291 patients (54 %) had a benign condition. Amongst the patients with malignancy 506 (47 %) had metastatic cancer, 400 (37 %) sarcoma (213 malignant primary bone tumours and 187 soft tissue sarcomas, Table 3 in appendix) and 174 (16 %) haematological malignancy. Benign conditions seen included benign neoplasia, inflammatory conditions, infection, and non-specific changes (Table 4 in appendix).

Results of the four models are demonstrated in Table 2.

These results demonstrate the performance of the four models generated when formed to optimise area under the curve.

Calculated negative predicted values (NPV) from this data demonstrate the NPV for haematological malignancy, metastasis, sarcoma and all malignancy are 97 %, 91 %, 88 % and 76 % respectively for these models.

The calibration curves are displayed in Fig. 1 and demonstrate good performance with linear curves closely following the ideal slope demonstrating agreement between predicted and observed values. The gradient of the slopes being 1.13, 1.05, 1.02 and 1.23 respectively for

Table 2
Performance metrics for diagnosing disease.

Class	Model	Accuracy %	Sensitivity %	Specificity %	F1 %	F2 %	AUC %	Train set n
Benign vs Malignant	Random Forest	74	72	75	71	72	80	1659
Benign vs Metastatic	Random Forest	79	79	80	69	75	87	1257
Benign vs Haematological	Random Forest	79	77	79	44	59	81	1025
Benign vs Sarcoma	Random Forest	64	76	61	52	64	73	1183

AUC: Area Under Receiver Operator Characteristic curve.
F1: Harmonic mean of precision and recall.
F2: Harmonic mean of precision and recall with added weight on recall.

Table 3
Demonstrating the split of the sarcoma group.

Group	Sum
Soft tissue sarcoma	187
Chondrosarcoma	106
Osteosarcoma	77
Ewing sarcoma	29
Adamantinoma	1

each of the four models.

ShAP summary plots demonstrating the most important features in each model are shown in Fig. 2. It is interesting but not entirely unexpected that the most important features differ per disease. In the metastatic model the most important parameters being alkaline phosphatase, age, gamma GT and haemoglobin. In the case of haematological malignancy age is the most important value along with haemoglobin, estimated GFR and red blood count. In the sarcoma model the most important features were alkaline phosphatase, ESR, INR and neutrophils with age following behind this.

4. Discussion

In this study we have observed that models can be developed based on the results of routinely undertaken blood tests which can be used to predict the presence of malignant disease in patients referred with suspected orthopaedic malignancy.

The models to evaluate metastatic disease, haematological malignancy and sarcoma all have high negative predictive values with a marked difference in the NPV in the malignant model.

These are three distinct pathological entities each with differences in their effect on physiology. Even within each group there are differences in the pathological processes for example haematological malignancy was considered as a single entity with lymphoma and myeloma assessed together. Within the sarcoma model soft tissue sarcoma, and bone sarcomas were considered together (Table 3 in appendix), each known to have differences. It is noted that within the sarcoma group that alkaline phosphatase takes position at the top of features of importance. This is likely to be due to the bone sarcomas in which alkaline phosphatase will be raised as a marker of bone formation [20] there are also known differences in the age ranges of different types of sarcoma [21]. This is a limitation of this project however these groups however would result in groups so small as to introduce error.

When we look at the ShAP plot there are differences in the features of importance between all these groups as expected from their disease characteristics and natural history therefore when look at a combined group or in an overall manner as in the malignancy model patterns may not be as easily forthcoming.

Despite these differences between groups and within groups the models formed have demonstrated capacity to discriminate between benign and malignant processes with AUC acceptable for sarcoma and excellent for other models.

Table 4
Demonstrating the split of diagnoses within the benign group.

Multi-Disciplinary Team Diagnosis	Sum
lipoma	244
non-specific / no neoplasia	207
enchondroma	78
giant cell tumour of bone	68
reactive	61
infection / osteomyelitis	56
epidermoid / sebaceous cyst	43
schwannoma (incl. variants)	41
osteochondroma	34
fibrous dysplasia	31
angioleiomyoma	29
tenosynovial giant cell tumour - diffuse type	25
fracture	23
bone island / bone infarct	23
haemangioma / artero-venous malformation	22
degenerative disease	21
ganglion	19
benign spindle cell tumour	18
osteoid osteoma / osteoblastoma	16
granlomatous disease / sarcoid	16
chondroblastoma	15
aneurysmal bone cyst	15
intramuscular myxoma	15
tenosynovial giant cell tumour - localised type	15
synovial chondromatosis	15
simple bone cyst	14
eosinophilic granuloma / haematopoetic island	14
benign skin adnexal tumour	12
non ossifying fibroma / fibro-osseous lesion	10
chondromyxoid fibroma	8
desmoid tumour	8
haematoma / seroma	8
chronic recurrent multifocal osteitis	7
Paget's disease	7
gout	6
osteoporosis	6
neuroma	6
fibroma / elastofibroma	6
heterotopic ossification	5
hibernoma	5
glomus tumour (and variants)	4
Dupuytren's disease	4
congenital abnormality	3
bizarre parosteal osteochondromatous proliferation	2
osteofibrous dysplasia	2
arthropathy	2
tuberculosis	2
	1291

When assessing a population-wide screening tool sensitivity, specificity, PPV ad NPV are all factors which are valuable. When we consider a scenario where there is a preselected cohort or patients who have been referred due to the presence of symptoms and signs raising concern about the potential presence orthopaedic malignancy the situation is a

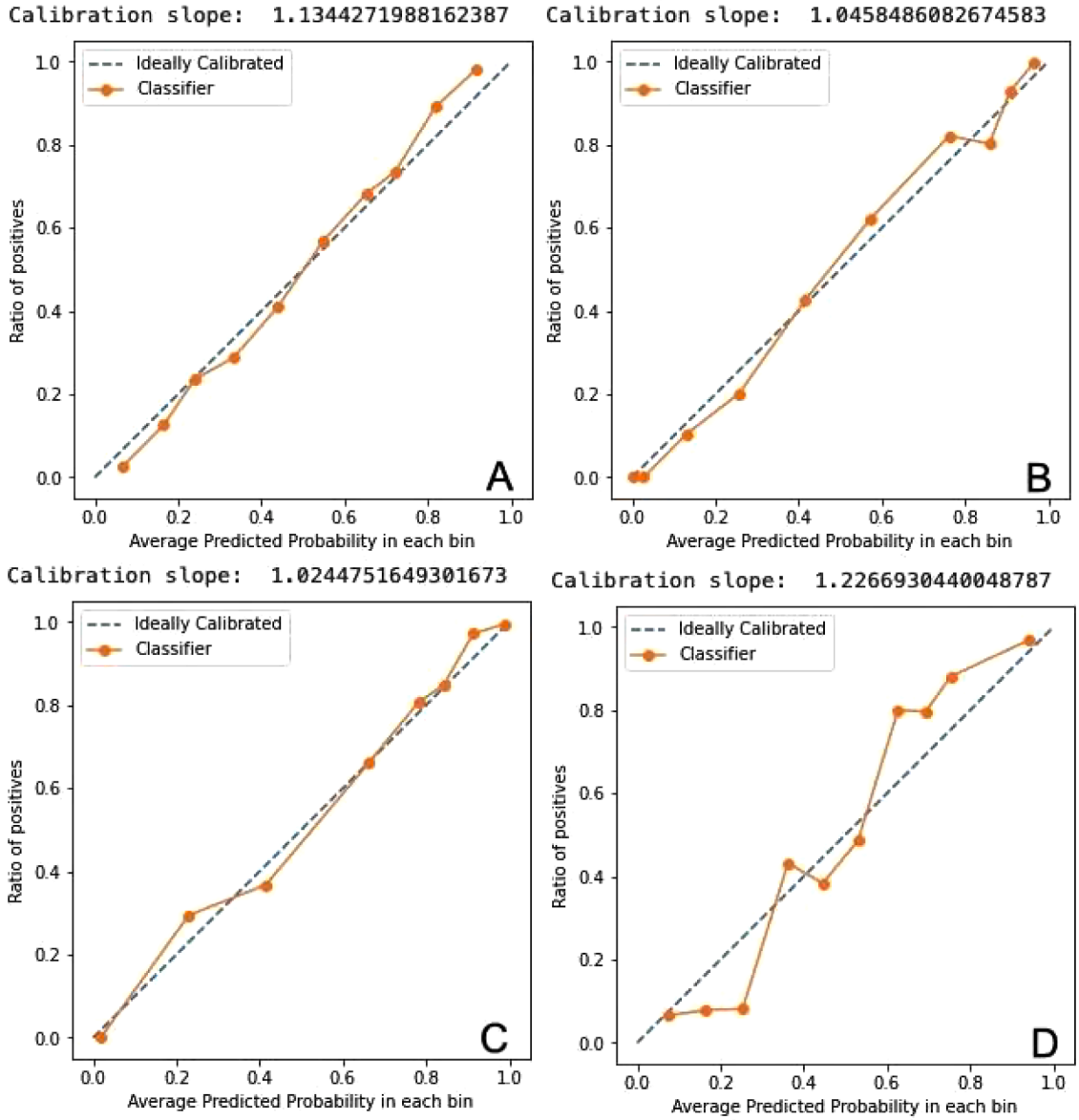

Fig. 1. Calibration curves for different models: Malignant (A), Metastatic (B), Haematological (C), Sarcoma (D).

little different. Here we are looking stratify this cohort into groups of higher or lower risk in order that those with the highest risk can be identified and their definitive assessment expedited. Negative predictive value is arguably a more useful characteristic to employ thereby minimising the frequency of cases incorrectly characterised as lower risk.

The models formed have been optimised to maximise the area under the ROC curve however the methods employed here can be adapted to maximise any outcome characteristic desired thereby optimising for example the negative predictive value at the expense of specificity which could prove useful in formation of a tool to stratify groups into higher and lower risk cohorts and minimise those incorrectly determined to have a lower risk profile or indeed sensitivity could be the focus should this be preferred.

This has potential for application in practice where all cases can be seen in a timely manner, but cases stratified into a higher risk cohort prioritised and assessed in an expedited fashion. There is also possibility for its application in circumstances where healthcare resources are less available to provide an inexpensive method of stratification of risk.

Whilst providing useful information this study has some limitations. Our dataset was a large dataset and common to retrospective studies where routinely collected data is subsequently used for a different purpose there was missing data. Whilst it would be ideal to have complete datasets, imputing with the mean z-score was selected as it would introduce least bias without affecting the model acknowledging that this may attenuate any correlation between the variable and outcome. Omitting all cases without complete datasets would have resulted in a large proportion of the data being discarded resulting in introduction of greater bias and use of regression imputation poses the opposite problem to use of the mean with potential to overestimate a relationship. A prospective validation study would provide more confidence in the models assessed. It is likely that the models will require adaptation and fine tuning in the future to enhance the predictive power. The addition of cancer specific markers could enhance the model [22] and could be considered in a future study. This is likely to provide a model with improvements in sensitivity, specificity, PPV and NPV however any improvements in the output of the model would need to be weighed up against applicability for a screening measure. Furthermore, radiological features (i.e. latent or aggressive benign, blastic or lytic metastases) could enhance future models. With the method employed retraining the model with additional data should prove straight-forward.

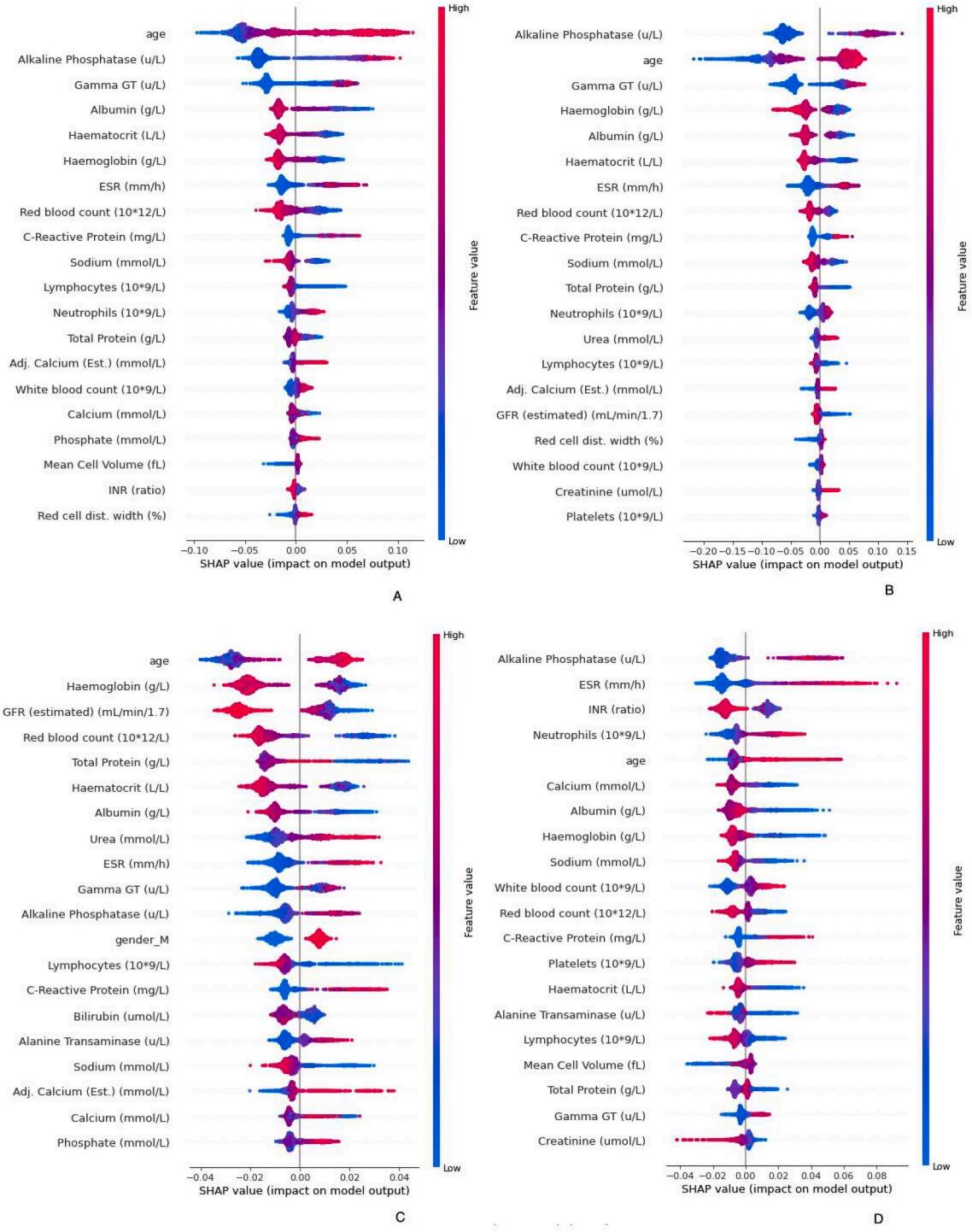

Fig. 2. ShAP (Shapley additive explanations) values for each model. Malignant (A), Metastatic disease (B), Haematological malignancy (C),Sarcoma (D). A high feature value is indicated in red and the impact on the model on the X axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

Routine blood tests can be useful in triage and risk stratification in patients with suspected orthopaedic malignancy. Further work should validate the developed models in a clinical settling.

CRediT authorship contribution statement

Kieran Bentick: Data curation, Writing – original draft. **Joel Runevic:** Methodology, Data curation, Writing – original draft. **Sriram Akula:** Methodology, Data curation, Writing – original draft. **Theocharis Kyriacou:** Methodology, Data curation, Writing – original draft. **Paul Cool:** Conceptualization, Methodology, Data curation, Writing – original draft. **Peter Andras:** Methodology, Data curation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] J. Wu, X. Zan, L. Gao, J. Zhao, J. Fan, H. Shi, et al., A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study, JMIR Med Inform. 7 (3) (2019 Aug 15) e13476.

[2] M.Y. Lu, T.Y. Chen, D.F.K. Williamson, M. Zhao, M. Shady, J. Lipkova, et al., AI-based pathology predicts origins for cancers of unknown primary, Nature 594 (7861) (2021 Jun 3) 106–110.

[3] Shepherd KL, Cool P, Cribb G. Prognostic indicators of outcome for patients with skeletal metastases from carcinoma of the prostate. Bone Jt J. 2018 Dec;100-B(12): 1647-54.

[4] Stevenson J, McMair M, Cribb GL, Cool P. Quadruple A blood parameter in the prediction of survival of patients presenting with bone metastases of breast cancer.

[5] The mGPS Study Group, S. Spence, J. Doonan, O.M. Farhan-Alanie, C.D. Chan, D. Tong, et al., Does the modified Glasgow Prognostic Score aid in the management of patients undergoing surgery for a soft-tissue sarcoma?: an international multicentre study, Bone Jt J. 104 (B(1)) (2022) 168–176.

[6] T. Nakamura, R. Grimer, C. Gaston, M. Francis, J. Charman, P. Graunt, A. Uchida, A. Sudo, L. Jeys, The value of C-reactive protein and comorbidity in predicting survival of patients with high grade soft tissue sarcoma, Eur J Cancer. 49 (2) (2013 Jan) 377–385.

[7] P. Jettoo, G. Tan, C.H. Gerrand, K.S. Rankin, Role of routine blood tests for predicting clinical outcomes in osteosarcoma patients, J Orthop Surg (Hong Kong). 27 (2) (2019) 1–6.

[8] T. Nakamura, R.J. Grimer, C.L. Gaston, M. Watanuki, A. Sudo, L. Jeys, The prognostic value of the serum level of C-reactive protein for the survival of patients with a primary sarcoma of bone, Bone Joint J. 95 ((B)3) (2013) 411–418.

[9] D.G. Menter, S.C. Tucker, S. Kopetz, A.K. Sood, J.D. Crissman, K.V. Honn, Platelets and cancer: a casual or causal relationship: revisited, Cancer Metastasis Rev. 33 (1) (2014) 231–269.

[10] V. Giannakeas, J. Kotsopoulos, M.C. Cheung, L. Rosella, J.D. Brooks, L. Lipscombe, et al., Analysis of platelet count and new cancer diagnosis over a 10-year period, JAMA Netw Open. 5 (1) (2022) e2141633.

[11] G.M. Bernardo Harrington, P. Cool, C. Hulme, A. Osman, J.R. Chowdhury, N. Kumar, et al., Routinely measured hematological markers can help to predict american spinal injury association impairment scale scores after spinal cord injury, Aug 28 [cited 2020 Nov 23], J Neurotrauma [ilnternet]. (2020), https://doi.org/10.1089/neu.2020.7144.

[12] W. McKinney, et al., Data structures for statistical computing in python, in: In Proceedings of the 9th Python in Science Conference, 2010, pp. 51–56.

[13] C.R. Harris, K.J. Millman, S.J. van der Walt, et al., Array programming with NumPy, Nature 585 (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.

[14] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in python, J Mach Learn Res. 12 (85) (2011) 2825–2830.

[16] L. Breiman, Random Forests, Mach Learn. 45 (1) (2001) 5–32.

[17] Thorn, James. Random Forest: Hyperparameters and how to fine-tune them [(Internet]. [cited 2022 Nov 21]. Available from: https://towardsdatascience.com/random-forest-hyperparameters-and-how-to-fine-tune-them-17aee785ee0d.

[18] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[19] SHAP. SHAP (SHapley Additive exPlanations) [Internet]. [cited 2022 Nov 21]. Available from: https://shap-lrjball.readthedocs.io/en/latest/index.html.

[20] Kim SH, Shin K, Moon SH, Jang J, Kim HS, Suh JS, Yang WI. Cancer Med. 2017 Jun; 6(6): 1311–1322. Published online 2017 May 11. doi: 10.1002/cam4.1022.

[21] A. Flanagan, J. Blay, J.V.M.G. Bovée, M. Bredelia, P. Cool, G. Nielsen, et al., Bone Tumours: Introduction. Soft Tissue and Bone Tumours WHO Classification of Tumours, 5th ed, World Health Organisation, 2019.

[22] G.J.S. Tan, C.H. Gerrand, K.S. Rankin, Blood-borne biomarkers of osteosarcoma: a systematic review, Pediatr Blood Cancer. 66 (1) (2019 Jan) e27462.