On improving clinical scoring systems: Multiple Myeloma case study

Đorđe Čikić[1,*,†], Dragana Drašković[2,†] and Dragan Janković[1,†]

[1]University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 4, 18000 Niš, Serbia
[2]University of Niš, Faculty of Medicine, Blvd. Dr Zorana Đinđića 81, 18000 Niš, Serbia

Abstract

Clinical scoring systems are very important in efficient prognosis of patient's health. This study represents an experiment on a dataset of Multiple Myeloma patients and the hypothesis is that current scoring systems in Multiple Myeloma used by clinicians can be improved. Experiment used AI-based framework called AutoScore that produces a scoring system that can be tuned by medical experts. Scoring system that resulted from this experiment showed a significant increase in correlation with Overall Survival compared to scoring systems currently in use and proposed in previous studies.

Keywords

Scoring system, Multiple Myeloma, Artificial Intelligence

1. Introduction

Clinical scoring systems are widely used in today's medicine, because they take complex patient information with many variables and turn it into one number that is easy to interpret. Their main purpose is risk stratification: the patients are divided into groups (i.e. low, intermediate and high risk) that have clearly different probabilities for some defined outcome, whether that outcome is death in hospital, progression of the disease, or long-term survival. But the role of the scores is not only prognosis. A score that is well validated can also help the physician to decide how intensive the therapy should be, it makes the communication with patients and their families easier, and it gives a reproducible base for dividing the participants in randomized controlled trials and in big multicenter registries [1, 2]. If we compare them with complicated "black-box" models, the additive scores have one clear advantage — their calculation is transparent. Every input variable gets a weight, and the sum is connected directly with a risk estimate. This transparency is important, because it makes the clinical acceptance easier and it allows external validation in different populations [3].

Representative scores across different fields of medicine include: CHA_2DS_2-VASc, which quantifies annual thromboembolic risk in non-valvular atrial fibrillation [4]; GRACE, which estimates six-month post-discharge mortality across the full spectrum of acute coronary syndrome [5]; APACHE II, which uses twelve physiological variables together with age and previous health status to predict the mortality in the intensive care unit [6]; and SOFA, which tracks sequential organ-failure trajectory across six systems in critically ill patients [7]. Many of the scoring systems used are developed many years ago when modern day problems did not exist or were negligible (i.e. pollution, obesity, new diseases).

Multiple myeloma is a clonal malignant disease of the terminally differentiated plasma cells, which accumulate in the bone marrow and produce a monoclonal immunoglobulin. Its typical clinical signs — hypercalcaemia, renal dysfunction, anemia and osteolytic bone lesions (the CRAB criteria) — are a consequence of the end-organ damage that is caused by the malignant clone and by its secretory

SQAMIA 2026:13th Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications, September 16–18, 2026, Niš, Serbia

*Corresponding author.
†These authors contributed equally.
djordje.cikic@elfak.ni.ac.rs (Đ. Čikić); draskovic.gaga@gmail.com (D. Drašković); dragan.jankovic@elfak.ni.ac.rs (D. Janković)
0009-0003-2931-3596 (Đ. Čikić); 0000-0002-4739-9261 (D. Drašković); 0000-0003-1198-0174 (D. Janković)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-19

CEUR-WS.org/Vol-4336/paper6.pdf

products [8]. In the last two decades there was a real therapeutic revolution, and today we have proteasome inhibitors, immunomodulatory drugs, anti-CD38 monoclonal antibodies, autologous stem-cell transplantation, CAR-T cell therapy and bispecific antibodies, but in spite of all this, for the majority of the patients multiple myeloma still stays incurable [8]. This incurability is additionally complicated by a very big biological heterogeneity: some patients reach long remissions that last for decades, while the others progress already after few months. Overall survival (OS), defined as the time from the diagnosis or from the beginning of the treatment until death from any cause, and Time to progression (TTP), defined as the time until the documented progression of the disease or death, are the main endpoints that are used to measure the benefit of the treatment and to compare the prognosis between trials and everyday practice. For this reason the reliable prediction of OS and TTP already in the moment of diagnosis is clinically very important: it helps in the selection of a risk-adapted therapy, it is the base for the counseling about realistic expectations of the outcome, and it gives the clinical trials the possibility to balance the treatment procedures according to the baseline prognosis.

This paper presents a research on currently used Multiple Myeloma scoring systems and depending on the result either tuning them or inventing a totally new scoring system. Apart from Introduction in this Section I, the paper is organized in 4 more sections: Section II describes the multiple myeloma scoring systems currently used and related work; In Section III, the dataset and method used are defined; Results and discussion are in Section IV; While Section V presents the conclusion and future work that will be done in this field.

2. Multiple myeloma scoring systems

The development of the formal prognostic staging in multiple myeloma started in 1975, with the Durie–Salmon staging system, that connects the myeloma cell mass — estimated from the hemoglobin concentration, the serum calcium, the level of the monoclonal protein and the radiographic extent of the osteolytic lesions — with three prognostic stages [9]. Even if it was widely accepted, the Durie–Salmon system depended on a subjective interpretation of the radiographic findings. The International Staging System (ISS), published in 2005 by Greipp et al. solved this problem and reduced the prognostic information to only two laboratory values that are easily available: the serum beta-2 microglobulin and the serum albumin. The ISS was derived from 10,750 previously untreated patients from 17 institutions, and it defined three stages, with median overall survival of 62, 44 and 29 months for the stages I, II and III [10]. The Revised ISS (R-ISS), published in 2015 by Palumbo et al. brought in the tumor biology, because it added the serum lactate dehydrogenase (LDH) and the high-risk cytogenetic abnormalities del(17p), t(4;14) and t(14;16) to the variables of the ISS. The new four-stage system showed a stronger prognostic gradient and it was recommended for the use in all future clinical trials [11]. The Second Revision of the ISS (R2-ISS), developed inside the HARMONY project of the European Myeloma Network (EMN) and reported by D'Agostino et al. in 2022 on 10,843 patients, refined the risk stratification additionally, because it added the gain or amplification of 1q as a fourth cytogenetic variable, and so it gave a four-group system that separates better the big intermediate-risk population that the R-ISS could not separate well enough [12]. The R2-ISS was after that externally validated in real-world populations in Australia and New Zealand [13], in Denmark [14] and from Balkans [15] where C-index from in all three studies ranged from 0.60 to 0.69, so the need for a further improvement stays open.

In parallel, there is now also a research direction that tries to go further than the established staging systems. The Real-World International Staging System (RW-ISS), reported in 2025 on the data of 2,929 Danish patients with newly diagnosed multiple myeloma, adds to the R2-ISS frame some patient-level variables that are usually missing in the clinical-trial cohorts — i.e. age over 70 years and the performance status (PS). It shows a better prognostic discrimination exactly in the population-based, real-world conditions. It reached a clearly higher C-index of 0.708 [16]. On the level of tumor biology, Ren et al. constructed the Ubiquitin Proteasome Pathway Risk Score (UPPRS), which is a nine-gene expression signature, obtained by LASSO-penalized Cox regression. A machine-learning scoring system that combines the UPPRS and the ISS stage gave better prediction of the overall survival than

both the ISS and the R-ISS, at least in the setting where it was derived [17]. Other authors added new biological information, for example the PET/CT radiomics features [18], the whole-genome and immune-microenvironment signatures [19], or the circulating plasma cells that, when combined with the R2-ISS, additionally separated the highest-risk patients [20].

3. Method and Dataset

The hypothesis of this paper is that the established and approved scoring systems currently used by clinicians can be improved using artificial intelligence.

The method consists of training a machine learning model to predict Overall Survival or Time to Progression and then extract the important features. The most important features will represent variables in the new scoring system. In order to implement this idea, this paper will use a semi-automated framework for building new scoring systems called AutoScore [3].

3.1. AutoScore framework

AutoScore is a data-driven framework designed for the development of interpretable clinical risk scores. It combines automatic variable selection with a simple point-based scoring system, allowing the creation of practical prediction tools from large clinical datasets. The framework was proposed to address one of the major limitations of machine learning models in healthcare, namely their limited interpretability. While complex machine learning algorithms often achieve high predictive performance, their results are not always easy to understand and apply in daily clinical practice. AutoScore aims to bridge this gap by converting machine learning outputs into transparent risk scores that can be used without advanced computational tools.

This framework has three different use-cases. The first one implemented is AutoScore for binary outcomes, i.e. predicting if patient will live or not, if patient will live for the next 2 years etc. It uses random forest for variable ranking and logistic regression for converting predictors into integer points. Next one is AutoScore-Ordinal for predicting multiclass outcomes that have a hierarchy. There are many examples of this in medicine where there is a scoring outcome that implies that some scores are worse or better than the other (i.e. low, intermediate and high risk). The third one is AutoScore-Survival which predicts to time-to-event outcomes, where censoring must be handled properly and risk changes over time. Instead of logistic regression, it relies on survival models, such as Cox proportional hazards, to derive time-dependent risk scores that still end up in an interpretable, point-based form. This makes it suitable for problems like predicting long-term mortality or readmission, where classic binary classification is not enough [3, 21, 22].

The AutoScore framework consists of several consecutive steps, starting from raw clinical data and ending with a compact risk score. AutoScore has an option to split the data into train and test using k-fold cross validation (k=5 in this study) which is recommended for datasets with fewer than 10000 rows. In the first stage, machine learning techniques, most commonly random forest, are used to rank candidate predictor variables according to their importance. This enables the identification of the most informative variables from a potentially large set of clinical and laboratory features. Compared with traditional manual variable selection, this approach is more objective and may identify predictors that would otherwise be overlooked. After the ranking process, only the most relevant variables are retained for further model development. Figure 1. shows an example of the feature importance from this study.

The next step involves the categorization of continuous variables. Since point-based scoring systems are easier to interpret when based on discrete categories, continuous measurements must be transformed into clinically meaningful intervals. Instead of selecting arbitrary thresholds, AutoScore searches for cut-offs that balance discrimination performance with interpretability. This procedure improves the usability of the final score while preserving most of the predictive information contained in the original variable. Since the goal of this framework is the interpretability, a medical expert can adjust these thresholds according to their previous experience resulting in a more interpretable yet accurate prediction.

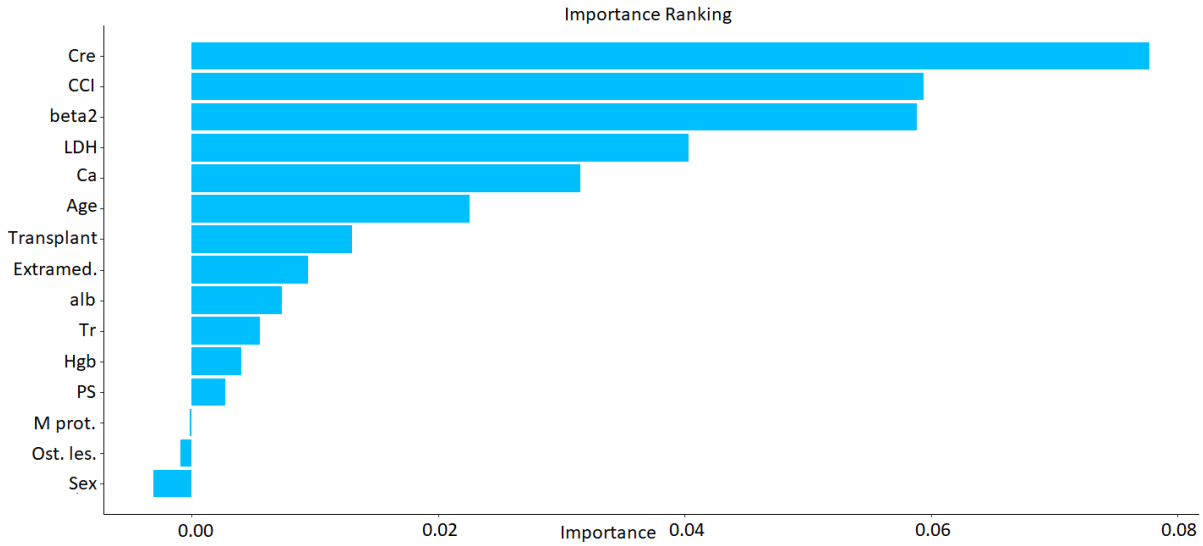


Figure 1: An output of the first module from AutoScore – Feature importance. All input features are displayed on y-axis and their importance on x-axis.

In the third step, AutoScore fits a simple logistic regression model on the selected and discretized variables to estimate their regression coefficients. These coefficients are then rescaled and rounded into integer points, so that each variable category contributes a small whole number to the total risk score. Categories associated with higher risk get more points, and protective categories can get zero or even negative points, depending on the implementation choice. The result is a table where clinicians can sum points across variables and map the total to an estimated probability of the outcome.

Depending on the use-case mentioned before, these points should be read differently. First of all, the points are positively correlated with higher risk of the outcome provided by user, but in binary classification these points do not represent anything meaningful. Together with the scoring system, AutoScore for binary classification provides the user with a threshold; if the sum of the points is greater that the threshold, the binary outcome is positive. For a Survival use-case, AutoScore proposes a scoring system ranging from *0* to *max_score* – a parameter of the framework, where this number of points does not represent a probability of survival, but it correlates with it.

Finally, there is a testing module at the end of the pipeline that allows users to calculate Area Under Curve (AUC) for every model created. For Survival use-case, AutoScore provides a probability distribution over time. The authors developed an R package that standardizes the entire workflow.

3.2. Dataset

Author 2, a physician at the Hematology Clinic of the Clinical Center Niš, collected data from patients with multiple myeloma following approval by the Ethics Committee for the use of these data in research on the treatment of multiple myeloma. The dataset has been fully anonymized and is available to qualified researchers upon submission of a justified request. All requests will be subject to review and approval by the Ethics Committee of the Clinical Center Niš before access to the data is granted. Dataset consists of Multiple Myeloma patients' records dating from 2016 up to 2026. Each record is tied to one patient and keeps information about initial state of the patient when diagnosed and tracks their responses to medical treatment. If patient's disease is progressing, there is information for TTP. Similarly, if patient dies there is information on how many months passed from diagnosing the disease to death - Overall Survival. Since all previously mentioned scoring systems use initial state of the patient to predict OS, this study focuses only on initial state and not medical treatment and responses. Dataset tracks multiple parameters of the patient's initial state:

- Age

- Sex
- Year of diagnosis
- Charlson Comorbidity Index (CCI)
- Eastern Cooperative Oncology Group - ECOG (PS)
- Clinical stadium - Durie&Salmon criteria (CS – DS)
- International Staging System (ISS)
- Type of imunnoglobulin or M protein
- Albumins (alb)
- β2 microglobulin
- Lactate Dehydrogenase (LDH)
- Creatinine (Cre)
- Calcium (Ca)
- Hemoglobin (hgb)
- Number of thrombocytes (Tr)
- FISH cytogenetic abnormalities
- Does patient have the osteolytic lesion?
- Does patient have the extramedullary disease?
- Did patient do the transplantation of stem cells?
- Is patient alive?

3.2.1. Statistics of the dataset

There are 310 patients in total - 150 male and 160 female. Figure 2 and 3 show the basic statistics of the dataset. Figure 2 represents the box plot of the numerical parameters and Figure 3 represents the pie chart of the categorical values. Dataset with this size and demographics is common in related studies [15, 23].

3.2.2. Preprocessing of the dataset

The target columns of the dataset are Overall Survival (OS) and Time to progression (TTP). OS is a time period from diagnosis of the MM until the death of the patient, while TTP is a time between peaks of

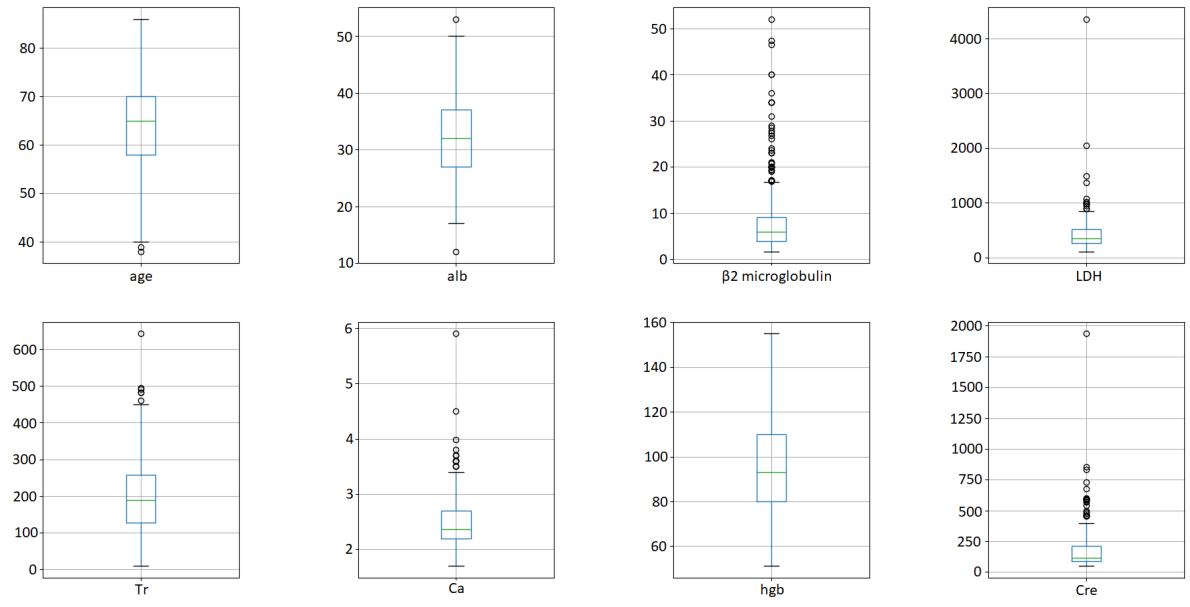


Figure 2: Box plot of 8 numerical features of the dataset.

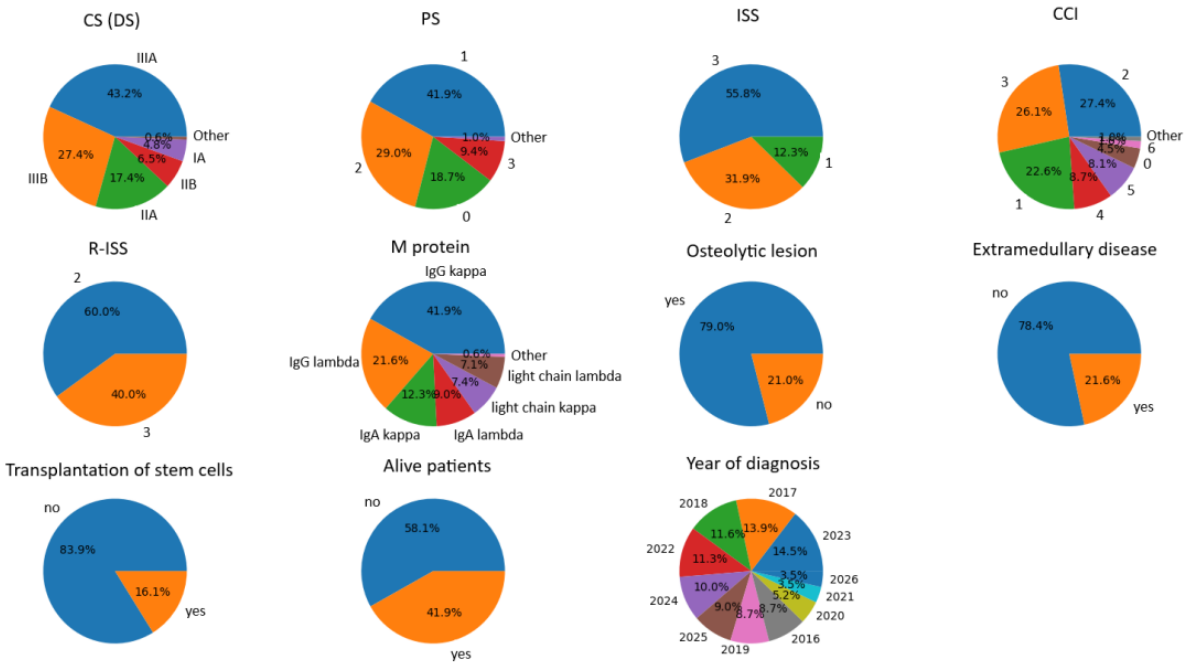


Figure 3: Distribution of categorical features of the dataset.

two active periods of the disease. A pie chart from Figure 3 shows that there are almost 42% of patients that are still alive, which means that they should be excluded from the analysis. Authors decided to include them, but in a separate analysis. The reason is to expand the dataset in order to potentially make a better predictor, because this scoring system will work with living people. Since there can be no information about the Overall Survival of the patients that are alive, authors made a simple solution to estimate their OS by calculating the difference between the current year and the year their disease was diagnosed. The same logic cannot be applied to TTP, so authors just excluded patients that did not have this information. Figure 4 shows the distribution of original OS, TTP and OS with estimated values. The distribution changed with adding a lot of higher values, because there are a lot of patients living with this disease and scoring system can now learn from patients that survived for more than 5 years, which is very useful for the predictor.

Authors identified another dimension of this problem. Clinicians often want to know if a patient will survive a fixed amount of time (months or years) and what is the optimal time point to predict. CCI is often used scoring system that outputs a probability of patient being alive in the next 10 years. This score is for general purpose, but Multiple Myeloma needs its own scoring system.

For this purpose, authors did another experiment that needs preprocessing of the dataset. As explained, for every patient there should be information if they will survive certain amount of months. Therefore, OS and TTP are preprocessed with thresholding resulting in a binary target column: “1” if patient died before fixed number of months and “0” if patient lived. Since there are no determined thresholds, authors experimentally determined that a range 2-50 months for OS and 2-36 months for TTP are more than enough for analysis of the threshold. This resulted in 82 new binary target columns (48 for OS and 34 for TTP).

Hypothesis of this study is that current scoring systems in use are less effective in predicting Overall Survival and Time to Progression, meaning that their correlation is low. Another hypothesis is that these current scoring systems can be improved using AI tools such as AutoScore or some custom-made AI software. Since TTP is more reliant on other factors that are not included in this study (therapeutic lines and responses to them), authors focused only on improving OS.

4. Results and Discussion

Metric used for determining the relationship between various scoring systems and Overall Survival is correlation, more precisely Pearson, Spearman and Kendall correlation [24, 25, 26]. The reason all three are used is that scoring systems in general are ordered and therefore Spearman and Kendall correlation naturally have a place in this comparison, beside regularly used Pearson correlation.

First of all, authors used this dataset to calculate correlations of existing scoring systems. Older scoring systems that are widely used - Durie&Salmon (DS) and ISS are included in this experiment. Beside that there are multiple revisions of ISS that are proposed, but only RW-ISS is included since it is the recently proposed. All of the graphs in Figure 5 will have the same structure: 6 graphs - 3 correlation types combined with 2 types of target columns (one is raw OS and TTP data which is not preprocessed with thresholding represented as a constant line, the other is data thresholded with different values of threshold explained in previous section). On X-axis is a value for threshold of months and on Y-axis is a correlation value. On the left is a correlation graph of OS with some scoring system and on the right is correlation graph of TTP with some scoring system.

Figure 5 shows the correlation between DS and OS (Figure 5.a), as well as DS and TTP (Figure 5.b). As seen from both graphs, DS is not a very good predictor of both OS and TTP, since its maximum correlation value (in absolute value) is around 0.17 which is very low correlation. Correlation of OS and TTP with ISS scoring system is shown in Figure 5.c and 5.d. In this graph, it is evident that ISS is either not a good predictor with maximum absolute correlation value of 0.27, but better than DS as it is expected. And the third scoring system experimented with is RW-ISS. It has mixed results, correlation with OS is better than ISS as it reaches -0.32.

All of these scoring systems are designed to predict survival of the patient, not the TTP, so it is reasonable why they fail to catch up with TTP values, but there is a lot of space for improving them in terms of predicting Overall Survival.

When using AutoScore framework, the first phase of building a scoring system is feature importance and the result for this dataset is shown in the second column of Table 1. From here, there are two options, one is to include first 5 features, because there is a big difference in importance between fifth and sixth feature and the other option is to include first 9 features.

It is very important to not have a large amount of features, because it increases the time to calculate the final score in a scoring system. In the second phase, which is a feature selection phase, every feature

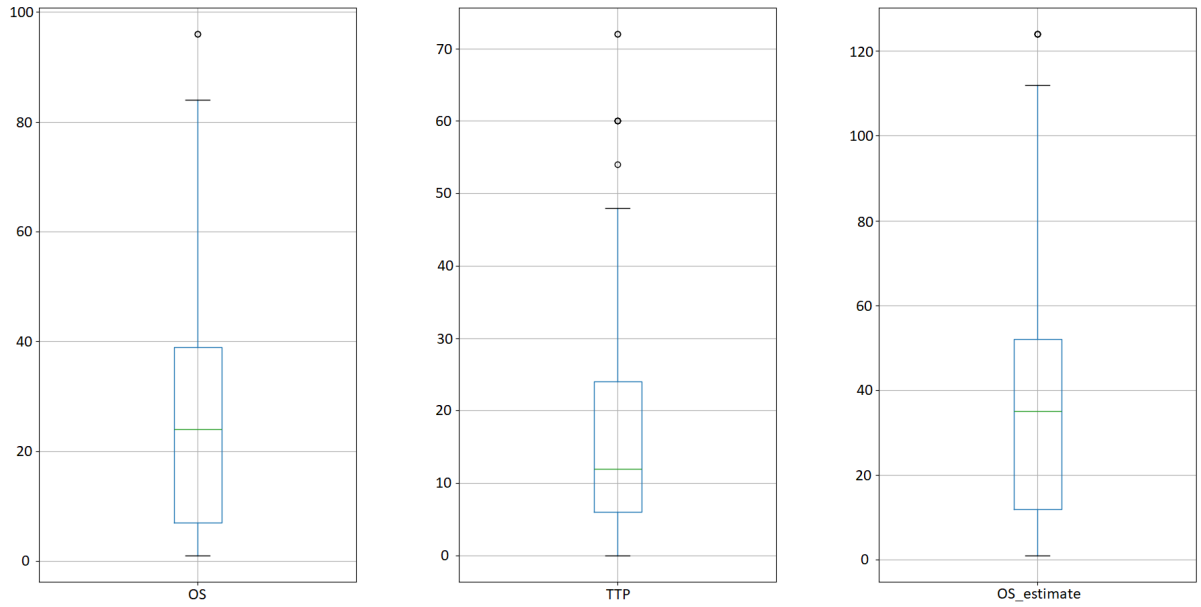


Figure 4: Distribution of original Overall Survival (OS), Time to progression (TTP) and Overall Survival with estimated values (OS_estimate)

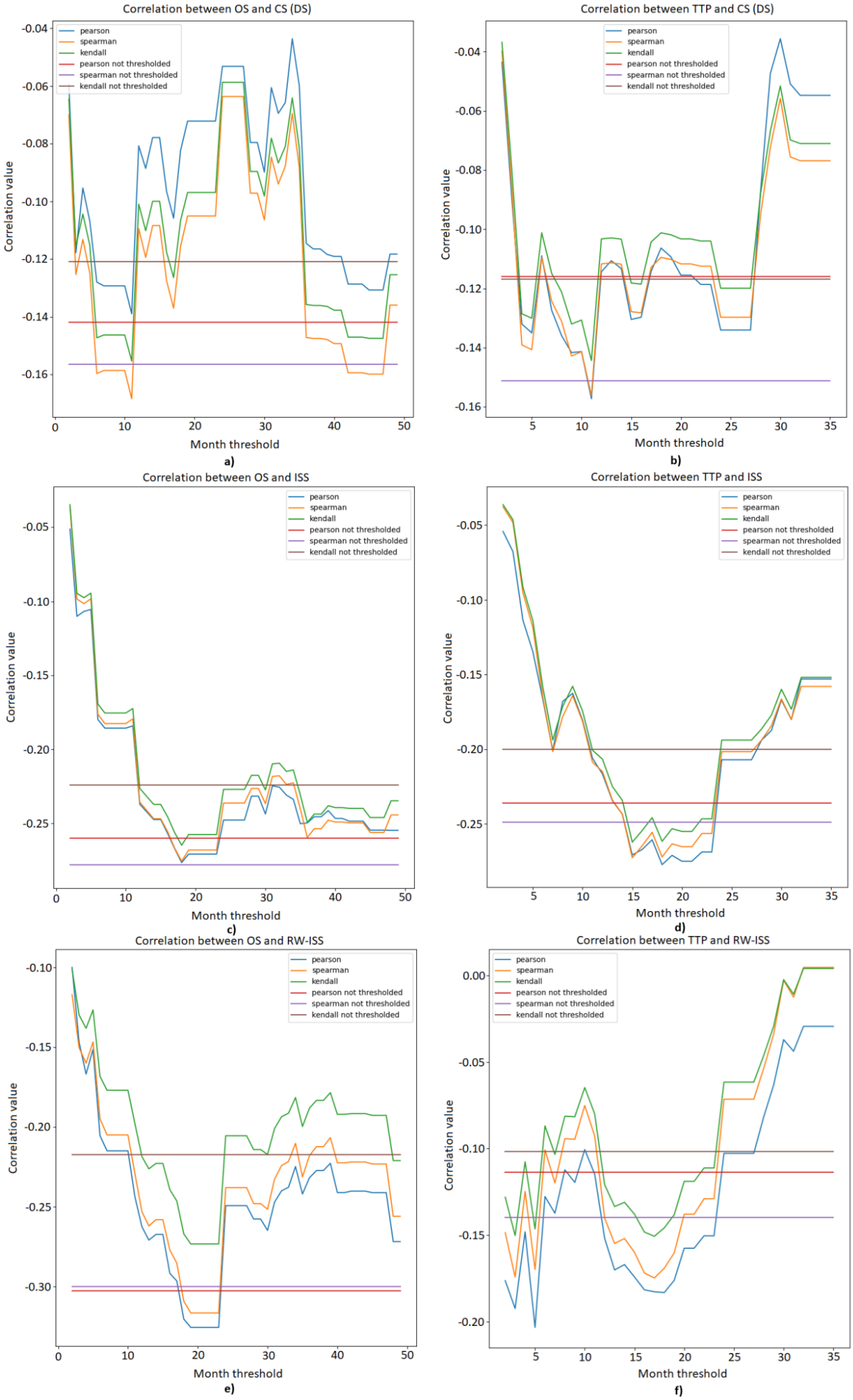


Figure 5: Correlation of OS and TTP with: DS – a) and b), ISS – c) and d) and RW-ISS – e) and f)

Table 1
Feature Importance sorted in descending order and AUC if that feature is selected

Feature	Importance	AUC if selected
CCI	0.078	0.640
LDH	300	0.675
Creatinine	0.066	0.721
Age	0.050	0.733
β2 microglobulin	0.040	0.734
Calcium	0.020	0.734
Stem cell transplant	0.018	0.731
Albumins	0.015	0.731
Number of thrombocytes	0.014	0.748
Hemoglobin	0.005	0.748
Other (total)	0.005	0.754

Table 2
Scoring system provided by AutoScore Survival with living patients included

Feature	Category/Interval	Scores
Creatinine	<65	17
	[65, 250)	0
	[250, 500)	3
	≥500	3
CCI	<1	0
	[1, 3)	8
	[3, 5)	13
	≥5	19
β2 microglobulin	<2.5	0
	[2.5, 11)	9
	[11, 24)	16
	≥24	11
LDH	<200	0
	[200, 250)	11
	[250, 550)	10
	[550, 800)	19
	≥800	26
Age	<45	0
	[45, 65)	1
	[65, 80)	8
	≥80	22

is included starting from the most important one and AUC is calculated. From the third column of Table 1, it is evident that only first option is viable, because AUC decreases after sixth variable and the AUC score when 9 variables are included is 0.015 higher than if 5 features are included, which is not beneficial if it makes the scoring system more complex. In the third phase, AutoScore generates a scoring system based on its algorithm, though it can be tuned in the fourth phase. Table 2 is a result of the scoring system after tuning with AUC of 0.736.

The same format of result is provided from AutoScore binary that authors used on thresholded OS values. There are 48 different experiments with the same number of scoring systems for different number of month threshold. Figure 6 sums that up in a correlation graph. On the left is a graph that shows the correlation of AutoScore proposed scoring systems with OS without estimated values and on the right is the same graph but OS has estimated values for living patients. Correlation reaches -0.55 in binary classification and -0.42 in survival analysis. Beside correlation, this scoring system achieved

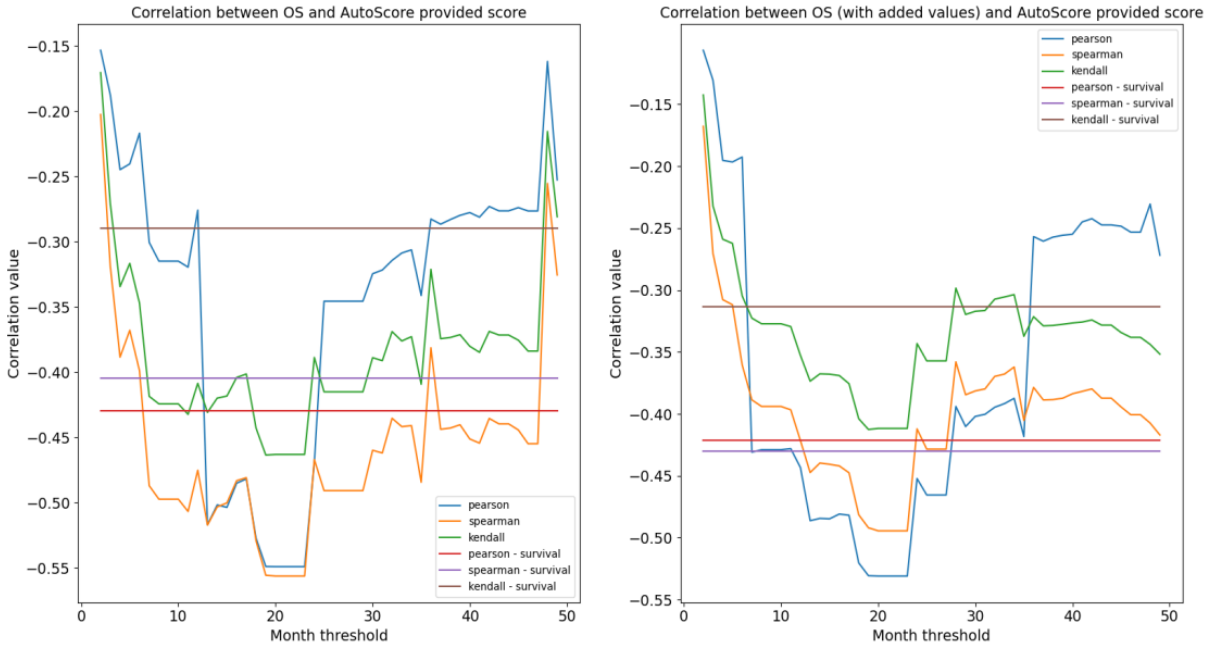


Figure 6: Correlation of OS and the scoring system AutoScore proposed

C-index of 0.722. These two graphs shows negligble difference between dataset where OS was estimated for living patients and dataset where living patients are excluded from the analysis. Either way, this represents a great improvement when comparing to scoring systems already in use and those that are recently proposed.

5. Conclusion and future work

This paper presented a method of improving Multiple Myeloma scoring systems using AI-based software AutoScore.

One of the main hypotheses of this study is that current Multiple Myeloma scoring systems in use have a low correlation with Overall Survival and Time to progression, which is proven with two different scores used by medical experts and one recently proposed have low correlation.

The other hypothesis is that AI-based algorithms can design new or improve existing scoring systems that have higher correlation with Overall Survival. This hypothesis is proven with AutoScore framework that produced a scoring system that substantially increased correlation of OS and AutoScore proposed score.

Future work will include therapeutic lines and responses in Multiple Myeloma, that will produce an even better scoring system and also include TTP. Along with that, authors will work on developing a new algorithm for constructing medical scoring systems that can outperform existing frameworks and algorithms.

Acknowledgments

This work has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [Grant Number: 451-03-34/2026-03/200102].

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] G. Collins, J. Reitsma, D. Altman, K. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement, European Journal of Clinical Investigation 32 (2015) 55–63. doi:10.1111/eci.12376.

[2] K. Moons, D. Altman, Y. Vergouwe, P. Royston, Prognosis and prognostic research: application and impact of prognostic models in clinical practice, BMJ 338 (2009) b606–b606. doi:10.1136/bmj.b606.

[3] F. Xie, B. Chakraborty, M. E. H. Ong, B. A. Goldstein, N. Liu, Autoscore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records, JMIR Med Inform 8 (2020) e21798. doi:10.2196/21798.

[4] G. Y. Lip, R. Nieuwlaat, R. Pisters, D. A. Lane, H. J. Crijns, Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation, Chest 137 (2010) 263–272.

[5] K. Eagle, M. Lim, O. Dabbous, K. Pieper, R. Goldberg, F. Van de werf, S. Goodman, P. Steg, J. Gore, A. Budaj, A. Avezum, M. Flather, K. Fox, A validated prediction model for all forms of acute coronary syndrome: Estimating the risk of 6month postdischarge death in an international registry, JAMA : the journal of the American Medical Association 291 (2004) 2727–33. doi:10.1001/jama.291.22.2727.

[6] W. Knaus, E. Draper, D. Wagner, J. Zimmerman, Apache ii: a severity of disease classification system, Critical care medicine 13 (1985) 818–29. doi:10.1097/00003465-198603000-00013.

[7] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. Mendonça, H. Bruining, C. Reinhart, P. Suter, L. Thijs, The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), Intensive Care Medicine 22 (1996) 707–710. doi:10.1007/BF01709751.

[8] S. V. Rajkumar, Multiple myeloma: 2024 update on diagnosis, risk-stratification, and management, American Journal of Hematology 99 (2024) 1802–1824. doi:10.1002/ajh.27422.

[9] B. G. M. Durie, S. E. Salmon, A clinical staging system for multiple myeloma correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival, Cancer 36 (1975) 842–854. doi:10.1002/1097-0142(197509)36:3<842::AID-CNCR2820360303>3.0.CO;2-U.

[10] P. R. Greipp, J. S. Miguel, B. G. Durie, J. J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J. A. Child, H. Avet-Loiseau, R. A. Kyle, J. J. Lahuerta, H. Ludwig, G. Morgan, R. Powles, K. Shimizu, C. Shustik, P. Sonneveld, P. Tosi, I. Turesson, J. Westin, International staging system for multiple myeloma, Journal of Clinical Oncology 23 (2005) 3412–3420. doi:10.1200/JCO.2005.04.242, pMID: 15809451.

[11] A. Palumbo, H. Avet-Loiseau, S. Oliva, H. M. Lokhorst, H. Goldschmidt, L. Rosinol, P. Richardson, S. Caltagirone, J. J. Lahuerta, T. Facon, S. Bringhen, F. Gay, M. Attal, R. Passera, A. Spencer, M. Offidani, S. Kumar, P. Musto, S. Lonial, M. T. Petrucci, R. Z. Orlowski, E. Zamagni, G. Morgan, M. A. Dimopoulos, B. G. Durie, K. C. Anderson, P. Sonneveld, J. San Miguel, M. Cavo, S. V. Rajkumar, P. Moreau, Revised international staging system for multiple myeloma: A report from international myeloma working group, Journal of Clinical Oncology 33 (2015) 2863–2869. doi:10.1200/JCO.2015.61.2267, pMID: 26240224.

[12] M. D'Agostino, D. Cairns, J. Lahuerta, R. Wester, U. Bertsch, A. Waage, E. Zamagni, M. V. Mateos, D. Dall'Olio, N. van de Donk, G. Jackson, S. Rocchi, H. Salwender, J. Creixenti, B. Holt, G. Castellani, F. Bonello, A. Capra, E. Mai, P. Sonneveld, Second revision of the international staging system

(r2-iss) for overall survival in multiple myeloma: A european myeloma network (emn) report within the harmony project, Journal of Clinical Oncology 364 (2022). doi:10.1200/JCO.21.02614.

[13] J. Tan, C. Wellard, E. Moore, P. Mollee, R. Rajagopal, H. Quach, S. Harrison, E. McDonald, P. Ho, H. Prince, B. Augustson, P. Campbell, Z. McQuilten, E. Wood, A. Spencer, The second revision of the international staging system (r2-iss) stratifies progression-free and overall survival in multiple myeloma: Real world data results in an australian and new zealand population, British Journal of Haematology 200 (2022). doi:10.1111/bjh.18536.

[14] C. Brieghel, M. Nørgaard Andersen, N. Grønne Jørgensen, T. Silkjær, T. Lund, M. Hinge, H. Gregersen, R. Schou Pedersen, A. Szabo, C. Niemann, P817: Validation of the second revised international staging system (r2-iss) in a large population-based, nationwide cohort of patients with newly diagnosed multiple myeloma, HemaSphere 7 (2023) e01814d6.

[15] E. Kastritis, E. Katodritou, S. Badelita, J. Bila, G. Cengiz Seval, Z. Cvetkovic, D. Coriu, E. Spanoudakis, D. Dalampira, A. Sretenovic, A. Sevastoudi, A. Bojan, M. Mitrovic, C. Danaila, M. Gavriatopoulou, M. Roussou, M. A. Dimopoulos, M. Beksac, E. Terpos, P956: Validation of the second revision of the international staging system (r2-iss) for overall survival in multiple myeloma in a real world cohort: An analysis by the balkan myeloma study group (bmsg), HemaSphere 7 (2023) e420064a. doi:10.1097/01.HS9.0000970728.42006.4a.

[16] C. Brieghel, T. Slordahl, M. Andersen, A. Szabo, c. u. Niemann, S. Thorsteinsdottir, A real-world international staging system (rw-iss) for patients with newly diagnosed multiple myeloma, Blood Cancer Journal 15 (2025). doi:10.1038/s41408-025-01268-y.

[17] L. Ren, B. Xu, J. Xu, J. Li, J. Jiang, Y. Ren, P. Liu, A machine learning model to predict survival and therapeutic responses in multiple myeloma, International Journal of Molecular Sciences 24 (2023). doi:10.3390/ijms24076683.

[18] B. Ni, G. Huang, H. Huang, T. Wang, X. Han, L. Shen, Y. Chen, J. Hou, Machine learning model based on optimized radiomics feature from 18f-fdg-pet/ct and clinical characteristics predicts prognosis of multiple myeloma: A preliminary study, Journal of Clinical Medicine 12 (2023). doi:10.3390/jcm12062280.

[19] F. Maura, E. M. Boyle, D. Coffey, K. Maclachlan, D. Gagler, B. Diamond, H. Ghamlouch, P. Blaney, B. Zicchedda, A. Cirrincione, M. Chojnacka, Y. Wang, A. Siegel, J. Hoffman, D. Kazandjian, H. Hassoun, E. Guzman, S. Mailankody, U. Shah, O. Landgren, Genomic and immune signatures predict clinical outcome in newly diagnosed multiple myeloma treated with immunotherapy regimens, Nature Cancer 4 (2023) 1–15. doi:10.1038/s43018-023-00657-1.

[20] B. Chu, Y.-T. Wang, S. Gao, L. Shi, M.-Q. Lu, L.-J. Fang, Q.-Q. Xiang, Y. Chen, M.-Z. Wang, L.-F. Wang, K. Sun, J. Yang, F. Duan, L. Bao, R2-ISS staging combined with circulating plasma cells improves risk stratification for newly diagnosed multiple myeloma: a single-center real-world study, Ann Hematol 103 (2024) 3677–3690.

[21] S. E. Saffari, Y. Ning, F. Xie, B. Chakraborty, V. Volovici, R. Vaughan, M. Ong, N. Liu, Autoscore-ordinal: an interpretable machine learning framework for generating scoring models for ordinal outcomes, BMC Medical Research Methodology 22 (2022). doi:10.1186/s12874-022-01770-y.

[22] F. Xie, Y. Ning, H. Yuan, B. A. Goldstein, M. E. H. Ong, N. Liu, B. Chakraborty, Autoscore-survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data, Journal of Biomedical Informatics 125 (2022) 103959. doi:10.1016/j.jbi.2021.103959.

[23] Y. Wang, T. Lan, C. Zhou, T. Xu, P. Liu, Comprehensive evaluation of clinical prognostic parameters in a real-world cohort of 812 patients with multiple myeloma, Clinical Cancer Bulletin 4 (2025) 4.

[24] K. Pearson, Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character (1896) 253–318. doi:10.1098/rsta.1896.0007.

[25] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1904) 72–101.

[26] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93. doi:10.1093/biomet/30.1-2.81.