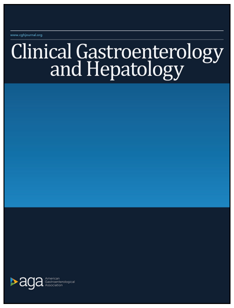


Gender-Equity Model for Liver Allocation using Artificial Intelligence (GEMA-AI) for waiting list liver transplant prioritization

Antonio Manuel Gómez-Orellana, MCS, Manuel Luis Rodríguez-Perálvarez, PhD, David Guijo-Rubio, PhD, Prof. Pedro Antonio Gutiérrez, PhD, Avik Majumdar, PhD, Prof Geoffrey W. McCaughan, PhD, Rhiannon Taylor, PhD, Prof. Emmanuel A. Tsochatzis, PhD, Prof. César Hervás-Martínez, PhD

PII: S1542-3565(24)01135-2
DOI: https://doi.org/10.1016/j.cgh.2024.12.010
Reference: YJCGH 59758

To appear in: *Clinical Gastroenterology and Hepatology*
Accepted Date: 19 December 2024

Gender-Equity Model for Liver Allocation using Artificial Intelligence (GEMA-AI) for waiting list liver transplant prioritization

GEMA-AI
Gender-Equity Model for Liver Allocation-AI

VS.

MELD-Na
Model for End-stage Liver Disease-Na

MELD 3.0
Model for End-stage Liver Disease 3.0

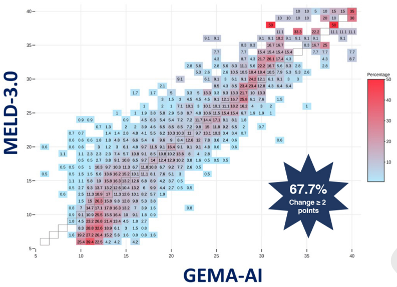
GEMA-Na
Gender-Equity model for Liver Allocation-Na

Cohort study including two populations of adult patients enlisted for elective LT

N=7,682 UK (2010-2020) Model training and internal validation	**N=1,638** Australia (1998-2020) External validation

Primary outcome: Mortality or delisting for sickness at 90 days

Transition from MELD 3.0 to GEMA-AI



DEATHS AVOIDED

1/13

1/11

Clinical Gastroenterology and Hepatology

Title

Gender-Equity Model for Liver Allocation using Artificial Intelligence (GEMA-AI) for waiting list liver transplant prioritization

Short Title

GEMA-AI for liver transplant prioritization

Authors

1- Antonio Manuel Gómez-Orellana*, MCS[1,2].

2- Manuel Luis Rodríguez-Perálvarez*, PhD[3,4,5].

3- David Guijo-Rubio, PhD[1,2,6].

4- Prof. Pedro Antonio Gutiérrez, PhD[1,2].

5- Avik Majumdar, PhD[7,8,9,10].

6- Prof Geoffrey W McCaughan, PhD[7,8,11].

7- Rhiannon Taylor, PhD[12].

8- Prof. Emmanuel A. Tsochatzis**, PhD[13].

9- Prof. César Hervás-Martínez**, PhD[1,2].

*Joint first authors, **Joint senior authors

Affiliations

1- Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain. Address: Campus Universitario de Rabanales, Albert Einstein Building. Ctra. N-IV, Km. 396. 14071, Córdoba, Spain.

2- Instituto Maimónides de Investigación Biomédica de Córdoba (IMIBIC), Córdoba, Spain. Address: Av. Menéndez Pidal, s/n, Poniente Sur, 14004 Córdoba, Spain.

3- Department of Hepatology and Liver Transplantation, Hospital Universitario Reina Sofía, IMIBIC, Córdoba, Spain. Address: Avda. Menéndez Pidal s/n, 14014, Córdoba, Spain.

4- Department of Medicine, University of Córdoba, Córdoba, Spain. Address: Calle Maria Virgen y Madre, 5, 14004, Cordoba, Spain.

5- Centro de investigación biomédica en red de enfermedades hepáticas y digestivas (CIBERehd), Madrid, Spain. Address: C/ Monforte de Lemos 3-5, 28029 Madrid, Spain.

6- Department of Signal Processing and Communications, University of Alcalá, Alcalá de Henares (Madrid), Spain. Address: Campus Universitario, Ctra. Madrid-Barcelona Km. 33, Alcalá de Henares (Madrid), 28805, Spain.

7- AW Morrow Gastroenterology and Liver Centre and Australian National Liver Transplant Unit, Royal Prince Alfred Hospital, Sydney, Australia.

8- Central Clinical School, The University of Sydney, Sydney, Australia.

9- Victorian Liver Transplant Unit, Austin Health, Melbourne, Australia.

10- The University of Melbourne, Melbourne, Australia. Address: 145 Studley Rd, Heidelberg, VIC 3084, Melbourne, Australia.

11- Liver Injury and Cancer Program, Centenary Institute, Sydney, Australia. Address: 50 Missenden Rd, Camperdown, NSW 2050, Sydney, Australia.

12- Department of Statistics and Clinical Studies, NHS Blood and Transplant, Stoke Gifford, Bristol, United Kingdom. Address: 500-600 North Bristol Park Northway, Bristol BS34 7QH, United Kingdom.

13- Sheila Sherlock Liver Unit and UCL Institute for Liver and Digestive Health, Royal Free Hospital, Pond Street, NW3 2QG, London, United Kingdom.

Grant support

“Agencia Estatal de Investigación (España)”, European Regional Development Fund, European Union, “Instituto de Salud Carlos III”, “Junta de Andalucía”, Ministry of Universities, Spanish Ministry of Research and Innovation.

Abbreviations

ACLF: acute-on-chronic liver failure

AI: artificial intelligence

ANN: artificial neural network

GEMA-AI: gender-equity model for liver allocation using artificial intelligence

GEMA-Na: gender-equity model for liver allocation corrected by serum sodium

Hc: Harrell's concordance statistic

INR: international normalized ratio

LT: liver transplantation

MELD-Na: model for end-stage liver disease corrected by serum sodium

ML: machine learning

RFH-GFR: Royal Free Hospital cirrhosis glomerular filtration rate

XAI: explainable artificial intelligence

Correspondence

Professor Emmanuel A. Tsochatzis, Sheila Sherlock Liver Unit and UCL Institute for Liver and Digestive Health, Royal Free Hospital, Pond Street, NW3 2QG, London, UK. Tel.: +44 2077 94500 extension 33575, Fax: +44 2074 726226. Email: e.tsochatzis@ucl.ac.uk

Disclosures

MLR-P has received lecture fees from Chiesi, Astellas, and Advanz Pharma, outside the present work. All other authors declare no competing interests.

Data Transparency Statement

The data used for model training and internal validation was extracted from the UK Transplant Registry, held by NHS Blood and Transplant. Deidentified participant data could be shared with an external investigator only after approval by NHS Blood and Transplant. For this purpose, proposals must be referred to the representative of NHS Blood and Transplant in this study, Rhiannon Taylor, by e-mail at rhiannon.taylor@nhsbt.nhs.uk. A signed confidentiality agreement would be required. The hybrid evolutionary algorithm used in this study is implemented in a Java framework for ANNs evolution called Neural Net Evolutionary Programming (NNEP), which is available at https://www.uco.es/grupos/ayrna/index.php/en/GEMA-AI.

Abstract

Background & Aims: We aimed to develop and validate an artificial intelligence score (GEMA-AI) to predict liver transplant (LT) waiting list outcomes using the same input variables contained in existing models.

Methods: Cohort study including adult LT candidates enlisted in the United Kingdom (2010-2020) for model training and internal validation, and in Australia (1998-2020) for external validation. GEMA-AI combined international normalized ratio, bilirubin, sodium, and the Royal Free Glomerular Filtration Rate in an explainable Artificial Neural Network. GEMA-AI was compared with GEMA-Na, MELD 3.0, and MELD-Na for waiting list prioritization.

Results: The study included 9,320 patients: training cohort n=5,762, internal validation cohort n=1,920, and external validation cohort n=1,638. The prevalence of 90-days mortality or delisting for sickness ranged 5.3%-6% across different cohorts. GEMA-AI showed better discrimination than GEMA-Na, MELD-Na and MELD 3.0 in the internal and external validation cohorts, with a more pronounced benefit in women and in patients showing at least one extreme analytical value. Accounting for identical input variables, the transition from a linear to a non-linear score (from GEMA-Na to GEMA-AI) resulted in a differential prioritization of 6.4% of patients within the first 90 days and would potentially save one in 59 deaths overall, and one in 13 deaths among women. Results did not substantially change when ascites was not included in the models.

Conclusions: The use of explainable machine learning models may be preferred over conventional regression-based models for waiting list prioritization in LT. GEMA-AI made more accurate predictions of waiting list outcomes, particularly for the sickest patients.

Keywords: eXplainable Artificial Intelligence; Machine Learning; Artificial Neural Networks; Liver Allocation; Gender; Disparities

Introduction

The historical imbalance between organ donors and potential candidates for liver transplantation (LT) requires to delineate strategies for organ allocation which allow to maximize donor utility while reducing the risk of mortality in the waiting list. The principle of urgency has prevailed over decades, with the sickest patients granted the first positions in the waiting list for earlier access to LT.[1] The current gold standard for ranking patients in the waiting list according to their mortality risk is the Model for End-stage Liver Disease corrected by serum sodium (MELD-Na), which combines four serum analytic and objective parameters, namely bilirubin, international normalized ratio (INR), creatinine and sodium.[2] However, despite the use of MELD-Na, the risk of mortality or delisting for sickness beyond safety thresholds for LT ranges from 9%-30% depending on the geographical area,[1] and is higher among women.[3]

In recent years, two alternative models have emerged to reduce mortality in the waiting list while addressing gender disparities for accessing LT. MELD 3.0 was developed and internally validated in the United States,[4] and the gender-equity model for liver allocation corrected by serum sodium (GEMA-Na) was trained and internally validated in the United Kingdom, and externally validated in Australia.[5] GEMA-Na was associated with a more pronounced discrimination benefit than MELD 3.0, probably owing to the replacement of serum creatinine with the Royal Free Hospital cirrhosis Glomerular Filtration Rate (RFH-GFR)[6] in the formula.[5] GEMA-Na has since been independently validated in Italian[7] and Spanish[8] transplant cohorts and has been adopted as the official transplant allocation system in Spain.

The methodology to design MELD-Na, MELD 3.0 and GEMA-Na was generalized additive Cox's regression which assumes a linear relationship between the analytical predictors and waiting list mortality. This linearity assumption is not met in clinical practice by any of the continuous variables contained in these scores, thus making it necessary to set lower and upper thresholds to define a range in which the relationship is linear. Such capping may result in less accurate predictions in individuals showing extreme analytical values, who are precisely the sickest patients requiring maximal prioritization. In addition, Cox models do not capture the complexity of the relationships between the covariates of the model and may overlook specific patterns or combinations of parameters associated with worse outcomes.

Machine Learning (ML) techniques, particularly Artificial Neural Networks (ANNs), have shown utility in hepatology,[10] although their implementation in clinical practice is not widely accepted by the medical community.[11] Indeed, ML algorithms are often criticized and referred to as "black box" models because of their difficult interpretability.[12] In recent years, eXplainable Artificial Intelligence (XAI)[13] has emerged under the rationale of giving equal relevance to performance and explainability, therefore granting acceptance for clinical use.[12] By using non-linear methods such as ANNs, empirical capping of continuous variables would not be required while patterns of variables associated with worse outcomes could be identified. A shallow ANN model composed of a reduced number of neurons in hidden layer enables its full explanation and interpretation about how outcome predictions are derived from specific inputs.

In this study, we propose a XAI model for waiting list LT prioritization using the same input variables which compose GEMA-Na.[5] The proposed model, Gender-Equity Model for Liver Allocation using Artificial Intelligence (GEMA-AI), was created using an ANN optimized by neuroevolution[14] and hybridization[15] and compared with existing LT allocation models. Finally, we provide a comprehensive explanation of GEMA-AI to ease the interpretation of its predictions, which is essential for its implementation in clinical practice.

Materials and Methods

Data sources, study population and outcomes

The population (n=9,320) comprised two cohorts of patients who were listed for LT in the United Kingdom and Australia. Briefly, a consecutive cohort of adult patients enlisted for elective LT in the United Kingdom Transplant Registry between April 1st, 2010, and March 31st, 2020, was used for model training and internal validation after random split of the database in a 3:1 ratio (n=5,762 and n=1,920, respectively). A second cohort of 1,638 patients included in the waiting list of LT in two Australian institutions from January 1st, 1998, to December 31st, 2020, was used for external validation. The data required to calculate the models used for comparison purposes (MELD-Na, MELD 3.0 and GEMA-Na) and to develop GEMA-AI was obtained at the inclusion in the waiting list. An overview of transplant policies in the United Kingdom and in the participating transplant institutions during the study period is provided in the appendix (p3).

After inclusion in the waiting list, patients were followed until transplantation, exclusion from the waiting list or death. The primary outcome of the study was a composite endpoint comprising death or exclusion from the waiting list due to clinical worsening, whichever occurred first, as a time-dependent outcome right-censored at 90 days after inclusion. The present study complies with the principles contained in the Declaration of Helsinki and was approved by the Andalusian ethics committee (Code 5412, 22/09/2022). This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.

GEMA-AI model development

To develop GEMA-AI we used ANNs[16] as ML approach based on Artificial Intelligence (AI). GEMA-AI was constructed from the training dataset using INR, bilirubin, sodium, and RFH-GFR as input variables. To optimize GEMA-AI during the training phase we used a hybrid approach combining neuroevolution[14] and hybridization.[15] We followed a XAI[13] perspective to provide the explanation and interpretation about how GEMA-AI derives prioritization scores. GEMA-AI score for each patient was rounded to the nearest integer and fitted to the range [6-40], which is used in LT prioritization. The methodology considered to develop GEMA-AI is fully described in appendix (pp3-11).

Performance evaluation of the GEMA-AI model

GEMA-AI model performance was compared with GEMA-Na, MELD-Na and MELD 3.0 in terms of discrimination, calibration, and re-classification. Discrimination was assessed by the Harrell’s concordance statistic (Hc), which is specific for time-dependent outcomes. To compare the discrimination ability of the different models, we used a one-shot non-parametric method specifically designed for right-censored outcomes,[17] in which resampling is not required. To evaluate the overall accuracy of the model we used the Brier score. Calibration was evaluated by the Greenwood-Nam-D’Agostino test, merging deciles if necessary to allow at least two events in each decile of risk. Calibration diagrams were plotted to graphically compare the agreement between the observed and predicted probabilities in each group. Re-classification refers to the ability of GEMA-AI to change the position of the patients in the waiting list compared with a reference model (MELD-Na, MELD 3.0 or GEMA-Na). A change of ≥2 points in the model score was considered clinically relevant as this could significantly impact on the probability of receiving a LT. To better understand the clinical impact of implementing GEMA-AI, we estimated the number of potential lives saved. First, we ranked the patients according to the score obtained from each model and considered the number of transplants performed within the first 90 days equal to the number of organs available. For each model, the prioritized group comprised those patients with the highest scores according to the number of available organs. To compare the re-classification benefit of GEMA-AI against another model, for instance MELD-Na, we identified the subgroup of patients who would be transplanted only under GEMA-AI prioritization or only under MELD-Na prioritization and calculated the risk of the primary outcome in both subgroups. The model which granted higher priority to patients with higher risk of the primary outcome was considered more appropriate for clinical use. Finally, the potential deaths avoided were calculated by subtracting the absolute number of patients experiencing the primary outcome in the GEMA-AI prioritized group, from the number of patients experiencing the primary outcome using each reference model, divided by the total number of patients experiencing the primary outcome.[5] In a sensitivity analysis, we assessed the performance of the model without taking into account the presence of moderate/severe ascites in the RFH-GFR equation (i.e. all patients were considered as not having ascites).

These analyses were performed in the training, internal validation, and external validation cohorts separately. Sub-analyses were performed in women for each cohort and in the subgroup of patients who showed at least one extreme analytical value in which a linear relationship with the primary outcome was not met: bilirubin >550 μmol/L, INR >3, RFH-GFR <20 mL/min, and serum sodium <122 mmol/L or >138 mmol/L.[5] As the number of patients in these subgroups was reduced, internal and external validation cohorts were merged for some analyses. For statistical tests, the significance level α=0.0500 was used. Analyses were performed using SPSS v27.0 and/or R v4.1.2.

Results

Clinical features of the study population including 7,682 patients from the United Kingdom for model training and internal validation, and 1,638 patients from Australia for external validation, are presented in table 1. The primary outcome occurred in 333 patients from the training cohort (5.8%), in 116 patients from the internal validation cohort (6.0%), and in 87 patients from the external validation cohort (5.3%). In the whole cohort, deaths accounted for 64.6% and delisting due to clinical deterioration accounted for 35.4% of the primary outcome events (n=346 and n=190, respectively). Minimum and maximum values of each input variable were obtained from the training dataset: serum bilirubin (2-870 μmol/L), INR (0.8-7.7), RFH-GFR (10.45-270.44 mL/min), and serum sodium (113-154 mmol/L), which were set as lower and upper thresholds for each of them to scale their values in the range [-1,1] (appendix p8) before entering the GEMA-AI model.

The mathematical definition of the GEMA-AI model is shown in table 2 and its graphical representation is illustrated in figure 1. The explanation and interpretation about how GEMA-AI derives prioritization scores is provided in appendix (pp12-14), being of particular interest the analysis of the sodium behavior due to its non-linear relationship with the primary outcome ("U" shape).

Table 3 shows the Hc for discrimination of GEMA-AI, GEMA-Na, MELD 3.0 and MELD-Na to predict the primary outcome. In the training cohort, GEMA-AI showed better discrimination (Hc=0.798) than MELD-Na (Hc=0.783;p=0.0424) and MELD 3.0 (Hc=0.770;p=0.0003). This superiority was consistent and more pronounced in the internal and external validation cohorts, with GEMA-AI obtaining the best discrimination capacity among women. When comparing GEMA-AI with GEMA-Na, there was no significant difference in the training cohort, but GEMA-AI performed better in the internal validation cohort (Hc=0.781 vs 0.766, p=0.0334) and in the external validation cohort (Hc=0.793 vs 0.774; p=0.003), with these differences being again more pronounced among women. The subanalysis according to the baseline liver disease is shown in the appendix (p27). GEMA-AI performed better than MELD-Na and MELD 3.0 in all cohorts, and better than GEMA-Na in the internal and external validation cohorts using the Brier score (appendix p28).

GEMA-AI and GEMA-Na showed good calibration in the internal and external validation cohorts (appendix p15). However, although MELD-Na and MELD 3.0 had adequate calibration in the external validation cohort, they showed poorer calibration in the internal validation cohort. The subgroups of women from the internal and external validation cohorts were combined to have a sufficient number of events to allow meaningful calibration. Again, GEMA-AI was well calibrated (appendix p16). Linear calibration diagrams of GEMA-AI and the reference models are shown in appendix (pp17-18).

Figures 2 and 3 and appendix (p19) show the re-classification diagrams of GEMA-AI vs the reference models, merging the internal and the external validation cohorts. GEMA-AI changed the position in the waiting list for a significant proportion of patients: a meaningful change of ≥2 score prioritization points occurred in 27.8% of patients (11.4% upgraded, 16.4% downgraded) when compared with GEMA-Na, in 67.7% of patients (39.2% upgraded, 28.5% downgraded) when compared with MELD 3.0, and in 61.5% of patients (33.4% upgraded, 28.1% downgraded) when compared with MELD-Na. During the first 90 days, a total of 3,725 transplant procedures were performed. Differential prioritization (ie, patients who would be transplanted only with GEMA-AI vs any of the other models) occurred in 6.4% (n=240) when compared with GEMA-Na, in 15.9% (n=594) when compared with MELD 3.0, and in 15.1% (n=561) when compared with MELD-Na. Clinical characteristics of patients differentially prioritized are shown in appendix (pp29-31). GEMA-AI prioritized more patients with moderate-severe ascites and worse renal function than any other model. GEMA-

AI attributed more weight to serum sodium and relatively less weight to bilirubin and INR. Of note, GEMA-AI prioritized more women than GEMA-Na (38.3% vs 28.8%;p=0.026) and MELD-Na (48.1% vs 29.9%;p<0.0001), but less than MELD 3.0 (36.4% vs 45.3%;p=0.002). The probability of the primary outcome was increased in patients differentially prioritized by GEMA-AI compared to MELD 3.0 (8.6% vs 2.5%;p<0.0001) and MELD-Na (8.4% vs 3.2%;p<0.0001), suggesting that sicker patients would get higher priority. Although numerically higher, this difference did not reach statistical significance when comparing GEMA-AI with GEMA-Na (6.7% vs 2.9%;p=0.054). The number of potential deaths avoided with the implementation of GEMA-AI would be one in 59 deaths overall compared with GEMA-Na (one in 13 deaths in women), one in 13 deaths overall compared with MELD 3.0 (one in 11 deaths in women), and one in 18 deaths overall compared with MELD-Na (one in 9 deaths in women).

Subsequently, we performed a subgroup analysis of the patients who showed at least one extreme analytical value. After merging the internal and the external validation cohorts (n=3,558), a total of 1,403 patients (39.4%) were eligible for this analysis. GEMA-AI obtained the highest discrimination (Hc=0.823), which was superior to that obtained by GEMA-Na (Hc=0.797;p=0.0362), MELD 3.0 (Hc=0.778;p=0.0189), and MELD-Na (Hc=0.769;p=0.0044). GEMA-AI was the only model with adequate calibration in this subset of patients (χ^2_7=5.04, p=0.6554) (appendix pp20-23).

Finally, we performed a sensitivity analysis where all patients were considered as not having ascites when estimating the RFH-GFR. This showed a non-significant decrease of the discrimination of the GEMA-AI score, which still performed better than MELD-Na and MELD 3.0 (appendix p32).

Discussion

In this study, using data from two different countries and LT allocation systems, we developed and validated the GEMA-AI score, which is a shallow ANN optimized by combining neuro-evolution and hybridization techniques. This is the first non-linear model aimed to waiting list prioritization according to the individual risk of short-term mortality or delisting for sickness. The interpretability of GEMA-AI with rational explanations of how prioritization scores are derived from covariates makes it attractive for implementation in clinical practice.

The components of the current scores available for waiting list prioritization provide objective and reproducible information about liver function (bilirubin, albumin, INR, and ascites) and renal function (serum sodium and RFH-GFR to a greater extent than serum creatinine), which in turn are associated with the probability of mortality or clinical deterioration resulting in transplant unsuitability.[18] However, this relationship is non-linear and could roughly follow two different patterns in real clinical practice. For most analytical parameters, there is a normality threshold below which the risk of the outcome remains almost unchanged. Above this threshold, there is a linear association with the outcome risk but at a certain point, for the highest values typically found in the sickest patients, the relationship with the outcome risk becomes exponential.[5] The second pattern is characteristic of serum sodium, and to a lesser extent of serum creatinine and bilirubin, and consists in a narrow range in which there is a linear relationship, but with exponential increase in the risk of mortality for both abnormally high and low values ("U" shape). To use these parameters in linear models such as additive Cox's regression, previous studies empirically established lower and upper bounds between which the relationship with the primary outcome is linear. However, this strategy may result in neglecting important prognostic information, particularly for the sickest patients who may require the first positions in the waiting list. This was confirmed in the subanalysis of patients with at least one variable outside the linear range, in which GEMA-AI was the only adequately calibrated model and showed the greatest advantage on discrimination, even over GEMA-Na.

Another advantage of non-linear methodologies, and particularly of ANNs, is their ability to identify patterns of combinations of values that are associated with an increased risk of death or delisting due to clinical worsening. While linear models give a fixed weight to each variable irrespective of its value or the value of other variables in the model, ANNs could capture specific combinations to modulate the weighting.[19] For instance, GEMA-AI identified the combination of severe sodium alterations and low RFH-GFR as a pattern associated with very high risk of the primary outcome, particularly in the presence of moderate-severe ascites, and therefore granted maximal priority to these patients (appendix p33). This observation is supported by previous studies in which the subgroup of patients with refractory ascites and hyponatremia were identified as insufficiently prioritized by the current linear models.[20,21] Although ANNs allow to consider a large number of variables to refine clinical predictions, it would also make the score more complex and prevent explainability. Our approach restricted the input variables to those already contained in previous scores and enabled full explainability, which could facilitate its implementation in clinical practice.

The implementation of GEMA-AI would have a significant impact on the waiting list composition, with a varying extent depending on the model in effect. The position in the waiting list would change by ≥2 score points in 27.8%-67.7% of patients. Differential prioritization between GEMA-AI vs MELD-Na and MELD 3.0 would be 15.1% and 15.9%, respectively, but to some extent, this could be explained because these models used different covariates: GEMA-AI included RFH-GFR and ascites instead of creatinine, and MELD 3.0 had sex and albumin. The comparison of GEMA-AI with GEMA-Na found differential prioritization in 6.4% of patients, which truly mirrored the impact of the transition from a linear to a non-linear methodology given that both models shared the same input variables. This may be clinically relevant in terms of potential lives saved and demonstrated that the use of ANNs allowed refined allocation decisions. An additional finding was that MELD 3.0 did not perform better than MELD-Na in these non-US cohorts, and that although it prioritized more women, these were not necessarily at higher risk of death.

There are subgroups of patients who could benefit the most from the implementation of this methodology. In recent years, many centers have incorporated Acute-on-Chronic Liver Failure (ACLF) grade 2 and 3, and severe acute alcoholic hepatitis unresponsive to corticosteroids as new indications for LT.[22] Such patients are extremely sick and their mortality risk without LT at 90 days is around 80%.[23] The analytic values of patients with ACLF and severe acute alcoholic hepatitis are often by far outside the bounds established for existing scores, resulting in inaccurate predictions of outcomes.[24] The use of GEMA-AI in these patients could provide them with the priority they deserve according to their actual risk of mortality. Another group of interest would be women. There is a historical gender imbalance for accessing LT according to which women have to wait longer to receive a liver graft, and they show increased risk of delisting for sickness.[3,25] GEMA-Na replaced creatinine by RFH-GFR and could amend this gender inequity[5] but in the present study it seems that GEMA-AI could make even more accurate predictions in women.

The GEMA-Na score was criticized for incorporating the presence of moderate/severe ascites in the calculation of RFH-GFR, which is one of its four components. The criticism concerns the lack of objectivity in assessing ascites. Although objective assessment of ascites can be accomplished with cross-sectional imaging, it is possible that this criticism might hinder implementation in certain countries. In order to alleviate these concerns, we tested an iteration of the GEMA-AI score which did not take into account the presence of ascites. We are pleased to report that even without ascites, the score still performed better than MELD-

Na and MELD 3.0. This iteration can be used in healthcare systems where the presence of ascites is considered too subjective for inclusion in an organ allocation score.

This study is not without limitations. In our efforts to develop an explainable model only including objective analytic variables widely available, we may have overlooked other relevant clinical or analytical parameters which could have improved the performance of the model. As with other transplant scores, GEMA-AI was designed to predict 90-days outcomes in the waiting list. Therefore, GEMA-AI should be reassessed at least every 3 months until transplant, or earlier than that upon a clinically meaningful change in the patient's physical condition. In addition, the model would require additional validations before implementing in allocation systems different from those used in the study. Patients with hepatocellular carcinoma and other MELD exceptions may require corrections of the score to allow equitable access to transplantation when using GEMA-AI. As the population of LT candidates may change over time, periodical performance tests should be performed to update the structure of the model if necessary. Finally, we lacked the data to evaluate the effect of implementing GEMA-AI on post-LT outcomes.

In conclusion, GEMA-AI made more accurate predictions of waiting list outcomes than the currently available models, and could alleviate gender disparities for accessing LT. The iteration of GEMA-AI without ascites also outperformed the MELD family scores and would be easier to implement in certain allocation systems. The implementation of GEMA-AI could save a significant number of lives and is considered feasible owing to the interpretability of the model.

Acknowledgments

The present study was supported by the “Agencia Estatal de Investigación (España)” (grant reference: PID2020-115454GB-C22 / AEI / 10.13039 / 501100011033), the Spanish Ministry of Research and Innovation, the “Instituto de Salud Carlos III”, and co-funded by the European Regional Development Fund (grant reference: FIS PI22/00312). AMG-O was supported by “Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía” (grant reference: PREDOC-00489). DG-R was supported by both the University of Córdoba through grants to Public Universities for the requalification of the Spanish university system of the Ministry of Universities, financed by the European Union - NextGenerationEU (grant reference: UCOR01MS) and by the "Agencia Estatal de Investigación (España)" MCIU/AEI/10.13039/501100011033 and European Union NextGenerationEU/ PRTR (grant reference: JDC2022-048378-I).

Tables

Table 1: Clinical features of the study population. Descriptive analysis of 9,320 patients enlisted for liver transplantation stratified in a derivation cohort from the United Kingdom (n=7,682) and an external validation cohort from Australia (n=1,638).

VARIABLE	**UNITED KINGDOM** **n=7,682**	**AUSTRALIA** **n=1,638**	**p**
Age	53.22 ± 11.55	53.52 ± 9.28	0.27
Sex (Women)	2,578 (33.6%)	432 (26.4%)	<0.001
Etiology of liver disease:			
Alcohol	2,783 (36.2%)	474 (28.9%)	<0.001
Hepatitis C	1,242 (16.2%)	681 (41.6%)	<0.001
NAFLD/cryptogenic	1,374 (17.9%)	203 (12.4%)	<0.001
Primary sclerosing cholangitis	808 (10.5%)	87 (5.3%)	<0.001
Primary biliary cholangitis	633 (8.2%)	142 (8.7%)	0.57
Ascites			
No	3,285 (42.8%)	616 (37.6%)	<0.001
Mild	1,986 (25.8%)	440 (26.9%)	
Moderate-severe	2,411 (31.4%)	582 (35.5%)	
Urea (mmol/L)	5.10 (IQR 3.9-7.1)	6 (IQR 4.0-8.0)	0.09
Creatinine (µmol/L)	80.13 ± 35.04	84.13 ± 39.14	<0.001
RFH-GFR (ml/min)	69.81 ± 25.04	66.51 ± 24.99	<0.001
INR	1.45 ± 0.44	1.63 ± 0.59	<0.001
Bilirubin (µmol/L)	43.95 (IQR 24-87)	53 (IQR 27-116)	<0.001
Na (mmol/L)	136.24 ± 4.65	136.17 ± 5.04	0.63
Albumin (g/L)*	31.86 ± 6.61	32.26 ± 6.91	0.034
MELD-Na	17.25 ± 6.44	18.90 ± 7.65	<0.001
MELD 3.0*	17.15 ± 6.29	18.80 ± 7.59	<0.001
GEMA-Na	17.65 ± 5.84	19.39 ± 6.93	<0.001
GEMA-AI	17.46 ± 5.62	19.13 ± 6.73	<0.001
Primary outcome**	449 (5.8%)	87 (5.3%)	0.40

NAFLD: Non-alcoholic fatty liver disease; RFH-GFR: Royal Free Hospital glomerular filtration rate; INR: international normalized ratio; MELD-Na: model for end-stage liver disease corrected by serum sodium; MELD 3.0: model for end-stage liver disease 3.0; GEMA-Na: gender-equity model for liver allocation corrected by serum sodium; GEMA-AI: gender-equity model for liver allocation using artificial intelligence.

* Albumin and MELD 3.0 were not available in 549 patients from the derivation cohort (7.14%).

** Mortality or delisting due to clinical deterioration within the first 90 days after inclusion in the waiting list.

Model equation and basis functions	Sign of the coefficient
$\text{GEMA-AI} = -25.210 \times B_1 + 21.395 \times B_2 + 10.592 \times B_3 - 7.108 \times B_4 + 29.981$	$B_1(-),B_2(+),B_3(+),B_4(-)$
$B_1 = B(-3.094 \times \text{bilirubin}^{*} + 2.620 \times \text{sodium}^{*} - 1.961 \times \text{INR}^{*} + 1.594 \times \text{RFH-GFR}^{*} - 3.427)$	$\text{bilirubin}^{*}(-),\text{sodium}^{*}(+),\text{INR}^{*}(-),\text{RFH-GFR}^{*}(+)$
$B_2 = B(-7.175 \times \text{RFH-GFR}^{*} + 2.448 \times \text{sodium}^{*} + 1.136 \times \text{bilirubin}^{*} - 6.631)$	$\text{RFH-GFR}^{*}(-),\text{sodium}^{*}(+),\text{bilirubin}^{*}(+)$
$B_3 = B(15.596 \times \text{sodium}^{*} + 4.996 \times \text{RFH-GFR}^{*} - 5.524)$	$\text{sodium}^{*}(+),\text{RFH-GFR}^{*}(+)$
$B_4 = B(13.977 \times \text{bilirubin}^{*} + 11.911 \times \text{sodium}^{*} + 9.137 \times \text{RFH-GFR}^{*} - 4.792 \times \text{INR}^{*} + 3.645)$	$\text{bilirubin}^{*}(+),\text{sodium}^{*}(+),\text{RFH-GFR}^{*}(+),\text{INR}^{*}(-)$

The result of each basis function (B_1, B_2, B_3 and B_4) is calculated according to equation 3 in appendix (p4). Basis functions and their input variables (INR^{*}, bilirubin^{*}, sodium^{*} or RFH-GFR^{*}) are shown ordered according to the absolute value of their coefficients. The * means that each input variable was previously scaled in the range [−1,1] (appendix p8). A positive coefficient is represented by (+), whereas (−) represents a negative coefficient.

Table 3: Harrell's concordance statistics for the distinct models in each cohort of the study.

	GEMA-AI	GEMA-Na	MELD 3.0	MELD-Na			
Cohort	Hc	Hc	Hc	Hc	p value[1]	p value[2]	p value[3]
Training (whole; n=5,762)	0.798 (0.772-0.824)	0.796 (0.769-0.823)	0.770 (0.740-0.800)	0.783 (0.755-0.810)	0.4945	**0.0003**	**0.0424**
Training (women; n=1,955)	0.824 (0.785-0.864)	0.821 (0.781-0.860)	0.766 (0.718-0.815)	0.784 (0.739-0.829)	0.3872	**0.0001**	**0.0027**
Internal validation (whole; n=1,920)	0.781 (0.732-0.829)	0.766 (0.715-0.818)	0.720 (0.657-0.784)	0.742 (0.686-0.797)	**0.0354**	**0.0006**	**0.0023**
Internal validation (women; n=623)	0.826 (0.747-0.905)	0.802 (0.716-0.888)	0.763 (0.660-0.867)	0.779 (0.688-0.871)	**0.0487**	0.0578	**0.0196**
External validation (whole; n=1,638)	0.793 (0.741-0.846)	0.774 (0.720-0.827)	0.749 (0.696-0.802)	0.745 (0.690-0.800)	**0.0030**	**0.0005**	**0.0024**
External validation (women; n=432)	0.836 (0.751-0.921)	0.796 (0.698-0.895)	0.732 (0.625-0.839)	0.714 (0.592-0.835)	**0.0143**	**0.0024**	**0.0048**

Hc=Harrell's concordance statistic (95% Confidence Interval). Each cohort was analyzed as a whole and also the subgroup of women separately. Albumin data was not available for 413 and 136 patients from the training and the internal validation cohorts, respectively, and were excluded from comparisons of MELD 3.0 vs GEMA-AI. GEMA-Na=Gender-Equity Model for liver Allocation corrected by serum sodium. MELD-Na=Model for End-stage Liver Disease corrected by serum sodium. P values highlighted with boldface denote statistically significant differences.

[1] p values of the discrimination comparison of GEMA-AI vs GEMA-Na.

[2] p values of the discrimination comparison of GEMA-AI vs MELD 3.0.

[3] p values of the discrimination comparison of GEMA-AI vs MELD-Na.

Figure legends

Figure 1 Graphical representation of GEMA-AI. The * means that each input variable was previously scaled in the range $[-1,1]$ as expressed in equation 9 in appendix (p8). B_1, B_2, B_3 and B_4 represent the four basis functions according to equation 3 in appendix (p4). The GEMA-AI score is calculated as defined in equation 2 in appendix (p4), then the score is rounded to the nearest integer and fitted to the range $[6,40]$.

Figure 2 Re-classification diagram showing patients prioritization of GEMA-AI vs MELD-Na merging the internal and the external validation cohorts (n=3,558). The number in each box represents the percentage of GEMA-AI for a specific MELD-Na score value. The diagonal, matching score values of both models, is represented by a gray frame. Above diagonal values represent lower GEMA-AI scores compared with MELD-Na, whereas below diagonal values higher GEMA-AI scores compared with MELD-Na.

Figure 3 Re-classification diagram showing patients prioritization of GEMA-AI vs GEMA-Na merging the internal and the external validation cohorts (n=3,558).

References

1. **Tschuor C,Ferrarese A,Kuemmerli C,**et al. Allocation of liver grafts worldwide - Is there a best system?. J Hepatol 2019;**71**:707-18.

2. Kim WR,Biggins SW,Kremers WK,et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med 2008;**359**:1018-26.

3. Cullaro G,Sarkar M,Lai JC. Sex-based disparities in delisting for being "too sick" for liver transplantation. Am J Transplant 2018;**18**:1214-9.

4. Kim WR,Mannalithara A,Heimbach JK,et al. MELD 3.0: The Model for End-stage Liver Disease Updated for the Modern Era. Gastroenterology 2021;**161**:1887-95.

5. **Rodríguez-Perálvarez ML,Gómez-Orellana AM**,Majumdar A,et al. Development and validation of the Gender-Equity Model for Liver Allocation (GEMA) to prioritise candidates for liver transplantation: a cohort study. Lancet Gastroenterol Hepatol 2023;**8**:242-52.

6. Kalafateli M,Wickham F,Burniston M,et al. Development and validation of a mathematical equation to estimate glomerular filtration rate in cirrhosis: The royal free hospital cirrhosis glomerular filtration rate. Hepatology 2017;**65**:582-91.

7. Marrone G,Giannelli V,Agnes S,et al. Superiority of the new sex-adjusted models to remove the female disadvantage restoring equity in liver transplant allocation. Liver Int 2024;**44:**103-112.

8. Rodríguez-Perálvarez ML,De la Rosa G,Gómez-Orellana AM,et al. GEMA-Na and MELD 3·0 severity scores to address sex disparities for accessing liver transplantation: a nationwide retrospective cohort study. eClinicalMedicine 2024;**74**:102737.

9. Harrell FE. Cox Proportional Hazards Regression Model. In: Regression Modeling Strategies. Springer New York, 2001;465-507.

10. Schattenberg JM,Chalasani N,Alkhouri N. Artificial Intelligence Applications in Hepatology. Clin Gastroenterol Hepatol 2023;**21**:2015-2025.

11. Berry P,Kotha S. The fundamental importance of exploring risks alongside the benefits in the application of artificial intelligence. J Hepatol 2023;**80**:e223-e225

12. The Lancet Respiratory Medicine. Opening the black box of machine learning. Lancet Respir Med 2018;**6**:801.

13. Barredo-Arrieta A,Díaz-Rodríguez N,Del Ser J,et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;**58**:82-115.

14. Stanley K,Clune J,Lehman J,et al. Designing neural networks through neuroevolution. Nat Mach Intell 2019;**1**:24-35.

15. Yao X. Evolving artificial neural networks. Proc IEEE Inst Electr Electron Eng 1999;**87**:1423-47.

16. Bishop CM. Neural Networks for Pattern Recognition. New York: Oxford University Press, 1995.

17. Kang L,Chen W,Petrick NA,et al. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. Stat Med 2015;**34**:685-703.

18. Ginès P,Krag A,Abraldes JG,et al. Liver cirrhosis. Lancet 2021;**398**:1359-76.

19. Rajkomar A,Dean J,Kohane I. Machine Learning in Medicine. N Engl J Med 2019;**380**:1347-58.

20. Sersté T,Gustot T,Rautou PE,et al. Severe hyponatremia is a better predictor of mortality than MELDNa in patients with cirrhosis and refractory ascites. J Hepatol 2012;**57**:274-80.

21. Heuman DM,Abou-Assi SG,Habib A,et al. Persistent ascites and low serum sodium identify patients with cirrhosis and low MELD scores who are at high risk for early death. Hepatology 2004;**40**:802-10.

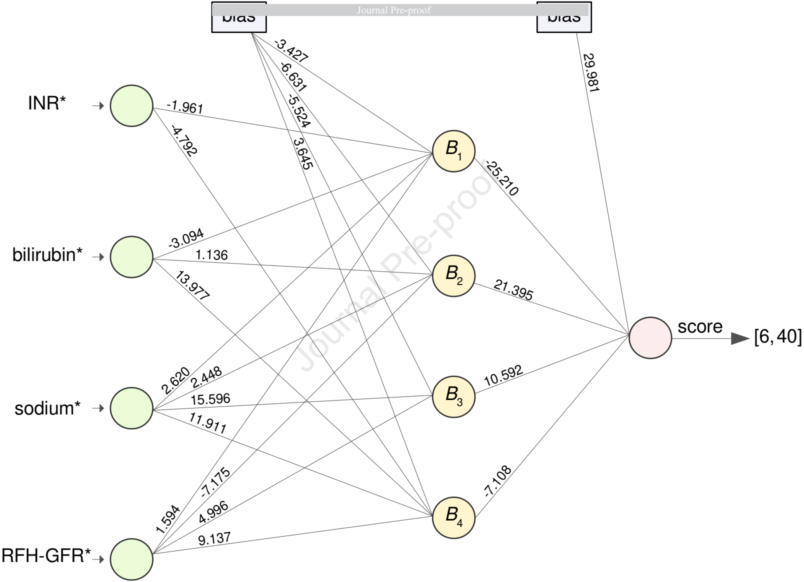
22. Rodríguez-Perálvarez ML,Gómez-Bravo MA,Sánchez-Antolín G,et al. Expanding indications of Liver Transplantation in Spain: Consensus statement and recommendations by the Spanish Society of Liver Transplantation. Transplantation 2021;**105**:602-7.

23. Artru F,Louvet A,Ruiz I,et al. Liver transplantation in the most severely ill cirrhotic patients: A multicenter study in acute-on-chronic liver failure grade 3. J Hepatol 2017;**67**:708-15.

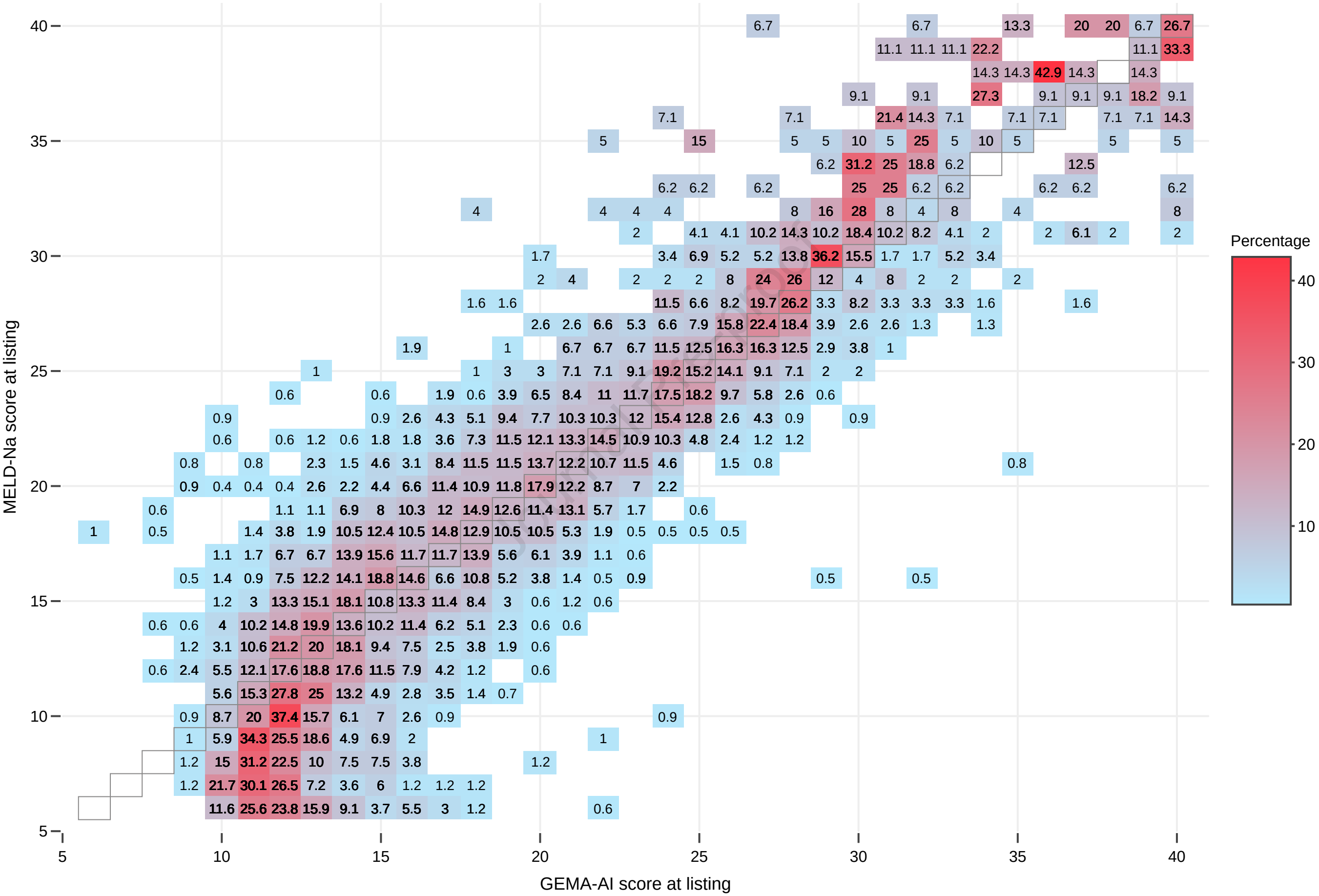
24. Hernaez R,Liu Y,Kramer JR,et al. Model for end-stage liver disease-sodium underestimates 90-day mortality risk in patients with acute-on-chronic liver failure. J Hepatol 2020;**73**:1425-33.

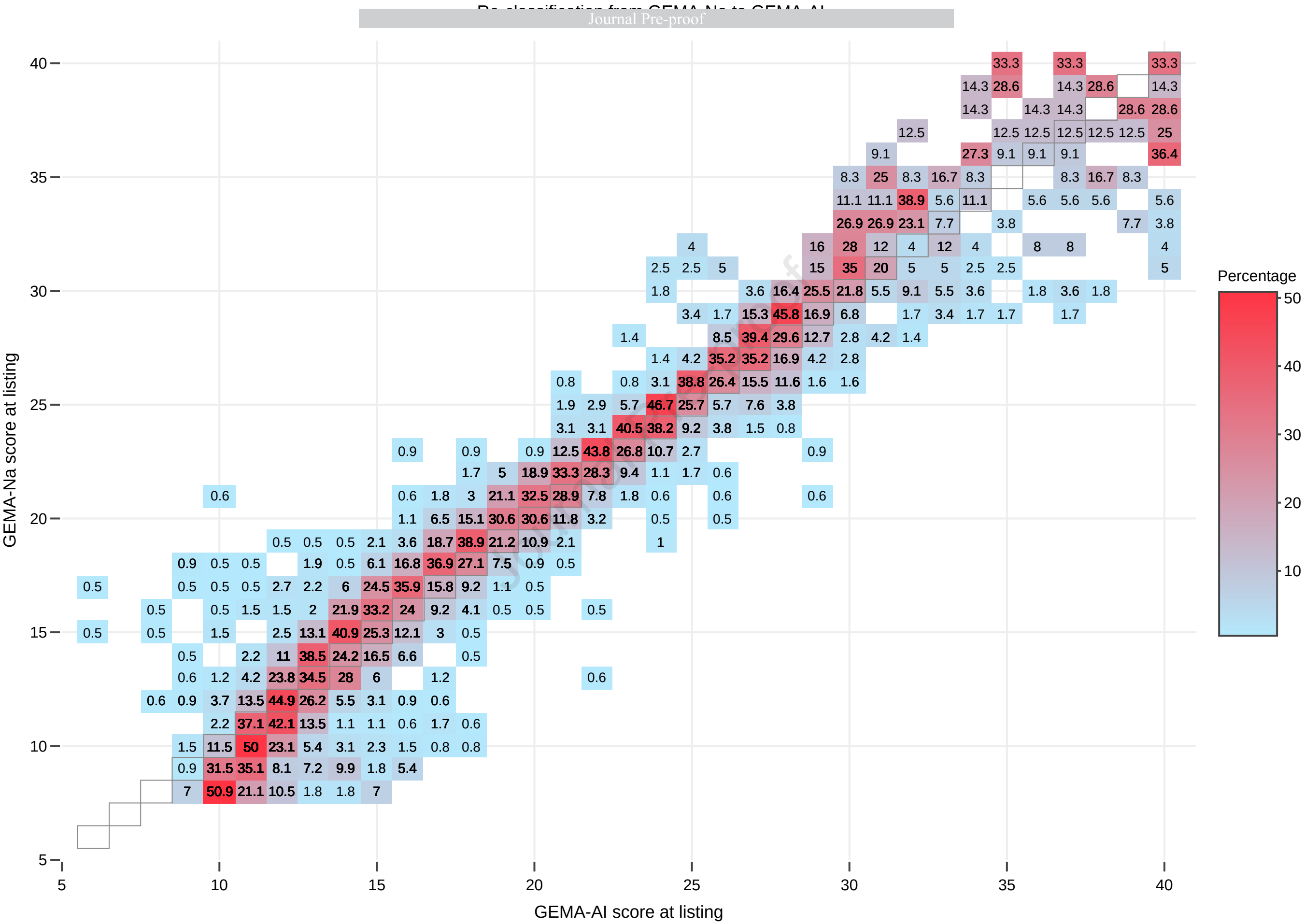
25. Allen AM,Heimbach JK,Larson JJ,et al. Reduced Access to Liver Transplantation in Women: Role of Height, MELD Exception Scores, and Renal Function Underestimation. Transplantation 2018;**102**:1710-6.

Author names in bold designate shared co-first authorship.


bias
bias
-3.427
-6.631
-5.524
3.645
29.981
INR*
-1.961
-4.792
B1
-25.210
bilirubin*
-3.094
1.136
13.977
B2
21.395
score
[6, 40]
sodium*
2.620
2.448
15.596
11.911
B3
10.592
RFH-GFR*
1.594
-7.175
4.996
9.137
B4
-7.108

Re-classification from MELD-Na to GEMA-AI




Re-classification from GEMA-Na to GEMA-AI
GEMA-Na score at listing
GEMA-AI score at listing
5
10
15
20
25
30
35
40
Percentage
10
20
30
40
50
33.3 33.3 33.3
14.3 28.6 14.3 28.6 14.3
14.3 14.3 14.3 28.6 28.6
12.5 12.5 12.5 12.5 12.5 12.5 25
9.1 27.3 9.1 9.1 9.1 36.4
8.3 25 8.3 16.7 8.3 8.3 16.7 8.3
11.1 11.1 38.9 5.6 11.1 5.6 5.6 5.6 5.6
26.9 26.9 23.1 7.7 3.8 7.7 3.8
4 16 28 12 4 12 4 8 8 4
2.5 2.5 5 15 35 20 5 5 2.5 2.5 5
1.8 3.6 16.4 25.5 21.8 5.5 9.1 5.5 3.6 1.8 3.6 1.8
3.4 1.7 15.3 45.8 16.9 6.8 1.7 3.4 1.7 1.7 1.7
1.4 8.5 39.4 29.6 12.7 2.8 4.2 1.4
1.4 4.2 35.2 35.2 16.9 4.2 2.8
0.8 0.8 3.1 38.8 26.4 15.5 11.6 1.6 1.6
1.9 2.9 5.7 46.7 25.7 5.7 7.6 3.8
3.1 3.1 40.5 38.2 9.2 3.8 1.5 0.8
0.9 0.9 0.9 12.5 43.8 26.8 10.7 2.7 0.9
1.7 5 18.9 33.3 28.3 9.4 1.1 1.7 0.6
0.6 0.6 1.8 3 21.1 32.5 28.9 7.8 1.8 0.6 0.6 0.6
1.1 6.5 15.1 30.6 30.6 11.8 3.2 0.5 0.5
0.5 0.5 0.5 2.1 3.6 18.7 38.9 21.2 10.9 2.1 1
0.9 0.5 0.5 1.9 0.5 6.1 16.8 36.9 27.1 7.5 0.9 0.5
0.5 0.5 0.5 0.5 2.7 2.2 6 24.5 35.9 15.8 9.2 1.1 0.5
0.5 0.5 1.5 1.5 2 21.9 33.2 24 9.2 4.1 0.5 0.5 0.5
0.5 0.5 1.5 2.5 13.1 40.9 25.3 12.1 3 0.5
0.5 2.2 11 38.5 24.2 16.5 6.6 0.5
0.6 1.2 4.2 23.8 34.5 28 6 1.2 0.6
0.6 0.9 3.7 13.5 44.9 26.2 5.5 3.1 0.9 0.6
2.2 37.1 42.1 13.5 1.1 1.1 0.6 1.7 0.6
1.5 11.5 50 23.1 5.4 3.1 2.3 1.5 0.8 0.8
0.9 31.5 35.1 8.1 7.2 9.9 1.8 5.4
7 50.9 21.1 10.5 1.8 1.8 7

What you need to know

Background: Existing models for liver transplant wating list prioritization are built on generalized additive Cox regression, which is a linear methodology that neglects prognostic information coming from extreme analytical values.

Findings: Using artificial intelligence in data from two different countries and organ allocation systems, we developed and validated the GEMA-AI score, which outperforms previous linear models to predict waiting list outcomes.

Implications for patient care: GEMA-AI is the first externally validated and fully explainable machine learning model which could avoid a meaningful number of deaths, particularly among historically disadvantaged groups of patients including women, individuals with ascites, and to the sickest patients showing extreme analytical values.

Contents

1 Supplementary Methods

1.1 Overview of liver transplant policies in the United Kingdom and Australia during the study period

There were similarities and differences of liver transplant practices between the United Kingdom and Australia during the study period that should be noted. Deceased donation was the major source of donors for adults in both countries, being living donors anecdotical. Regarding liver transplant candidates, there was a male predominance in both countries, although this was more pronounced in the participating Australian centres (73.6% men, 26.4% women) than in the UK (66.4% men, 33.6% women). The most frequent ethnicity was Caucasian (88% in the UK and 80.1% in Australia) followed by Asian (7.7% in the UK and 14.4% in Australia), and Black (6.8% in the UK and 2.5% in Australia). Allocation of donors for adult elective liver transplantation was centre based in both countries. Donors were offered to centres rather than to named individuals. If the local centre declined the offer for all patients on its list, the organ would be offered to centres in other regions. The United Kingdom End Stage Liver Disease (UKELD) score was used in the UK and the MELD score was used in Australia during the study period. In 2018, the National Liver Offering Scheme (NLOS) was introduced in the UK, which is an algorithm aimed to balance survival on the waiting list and survival after liver transplantation [1]. In Australia, the "Share 35" policy, which grants national priority to patients with MELD score above 35, was adopted in 2016. Hepatocellular carcinoma was the most frequent MELD exception in both countries and Milan criteria were observed during the earlier period of the study. In the UK, the alpha-fetoprotein model [2] replaced Milan criteria in 2014. In Australia, the UCSF criteria were implemented in 2010 [3]. Downstaging to Milan criteria was allowed both in the UK and Australia and bridging therapies were performed in the waiting list unless technically unfeasible. Finally, the probability of death or delisting for sickness at 90 days in the UK was 5.8% while in the participating Australian institutions was 5.3%.

1.2 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) [4] promotes the development of explainable models that enable end-users to understand their mechanisms and predictions while maintaining their prediction accuracy, and is broadly acknowledged as an utmost characteristic for the adoption of Machine Learning (ML) models based on Artificial Intelligence (AI), especially in healthcare.

Although interpretable by design models (ie, transparent models) are desirable in clinical practice, they have limitations for modeling non-linear problems, where a XAI model can overcome such limitations. Besides, interpretability can also be considered as a design factor in ML models development that are not interpretable by design without loss of performance [5].

We used post-hoc XAI techniques for model explanation and understanding. In addition, we considered some constraints in the model structure as an attempt to develop an intrinsically interpretable model.

1.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) [6] are ML models based on AI that can infer complex learning rules or non-linear relationships from data using mathematical non-linear functions, often referred to as basis functions, arranged and interconnected in a layered structure. Usually, this structure is mainly composed of one input layer, at least one hidden layer, and one output layer, which are linked by connections between neurons of contiguous layers. Specifically, in feedforward ANNs the input layer feeds data to the model, which is non-linearly transformed by basis functions in the hidden layers and, finally, the output layer obtains the model prediction.

To develop GEMA-AI, we addressed the problem from the point of view of binary classification, with the class $\mathcal{C}_+$ representing the patients who experienced the primary outcome within the 90 days in the waiting list (ie, primary outcome = 1), and the class $\mathcal{C}_-$ representing the patients who did not (ie, primary outcome = 0). Hence, the problem was formulated as follows:

$$D = \Big\{ (\mathbf{x}_i, y_i) \,; i = 1, 2, \dots, n \Big\}, \tag{1}$$

where D is the training dataset, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^4$ represents the vector composed of the four input variables of the i-th patient, being $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, with x_{i1}, x_{i2}, x_{i3}, and x_{i4} standing for INR_i, $\mathrm{bilirubin}_i$, sodium_i, and $\mathrm{RFH\text{-}GFR}_i$, respectively; $y_i \in \{0, 1\}$ is the target class to predict (ie, class $\mathcal{C}_-$ or class $\mathcal{C}_+$, respectively); and n is the number of patients in the dataset.

As aforementioned, with the aim of developing an intrinsically interpretable model, favoring its transparency without loss of performance, we considered the following two constraints in the model structure: 1) the number of hidden layers was fixed to one (ie, a shallow model); 2) the maximum number of basis functions in the hidden layer was fixed to four. Hence, the ANN classification model we considered for the problem being tackled was composed of three layers: one input layer with four neurons (one for each analytical predictor), one hidden layer, and one output layer with one linear neuron to predict the class of each patient. So, the model output was defined as follows:

$$f(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{m} \beta_j B_j(\mathbf{x}_i, \mathbf{w}_j), \quad (1 \leq m \leq 4), \tag{2}$$

where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)$; $B_j(\mathbf{x}_i, \mathbf{w}_j)$ stands for each basis function in the hidden layer that non-linearly transforms the input vector $\mathbf{x}_i$ of the i-th patient; $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$ represent the synaptic weights (coefficients) of the connections between neurons of the hidden and the output layers, β_0 being the bias; $\mathbf{w}_j^T = (w_{j0}, w_{j1}, \dots, w_{j4})$ are the coefficients of the connections between neurons of the input and the hidden layers, w_{j0} being the bias; and m represents the number of basis functions in the hidden layer.

As for the basis functions in the hidden layer, the Sigmoidal Unit (SU) [7] was considered. An ANN model using SUs as basis functions represents an additive model of non-linear transformations able to learn problems that are non-linearly separable. Following the notation used in equation 2, SU basis function was formulated as follows:

$$B_j(\mathbf{x}_i, \mathbf{w}_j) = \frac{1}{1 + e^{-\left(w_{j0} + \sum_{k=1}^{4} w_{jk} x_{ik}\right)}}, \quad j = 1, ..., m, \quad (0 \leq B_j(\mathbf{x}_i, \mathbf{w}_j) \leq 1)\,. \tag{3}$$

Therefore, and according to equation 2, the optimization of the proposed ANN model consisted in estimating from the training dataset D the number m of basis functions in hidden layer along with their connections, and the values of the parameters $\mathbf{W}$ and $\boldsymbol{\beta}$, that is, the model structure and synaptic weights.

Finally, to obtain the predicted class of each patient, the model output (equation 2) was transformed into a probability using the following softmax transformation:

$$g(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta}) = \frac{\exp(f(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta}))}{1 + \exp(f(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta}))}, \quad (0 \leq g(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta}) \leq 1)\,, \tag{4}$$

where $g(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta})$ represents the probability that patient $\mathbf{x}_i$ belongs to the positive class, whereas $1 - g(\mathbf{x}_i, \mathbf{W}, \boldsymbol{\beta})$ is the probability of not belonging to the positive class.

1.4 Neuroevolution

The optimization of ANNs involves both structure and synaptic weights, being a challenging task due to the numerous local minima they present as a consequence of their convoluted and complex error surface [8]. Backpropagation [9] and other gradient-descent-based algorithms (local optimization techniques) are most commonly used for ANNs optimization. However, these algorithms present two important shortcomings. They often get stuck in the local minima presented in the error surface of the ANNs, which makes it necessary to run the process several times. Moreover, most of them only adjust the synaptic weights of the ANNs in an iterative process, not their structure. Hence, taking into account that there is no systematic procedure to automatically design a specific near-optimal structure, many model architectures have to be empirically evaluated to identify the best possible configuration.

Neuroevolution [10] is a powerful alternative approach to overcome the above referred drawbacks which leverages Evolutionary Algorithms (EAs) to formulate the optimization task as the evolution of both synaptic weights and structure in the environment defined by the problem being addressed (ie, the training dataset D). The evolution of ANNs structure provides an automatic way to discover new structures, allowing ANNs to dynamically adapt to their environment by inferring learning rules from it [8], that is, relationships between input variables. The simultaneous evolution of synaptic weights and structure represents a more efficient way of optimization leading to better results. In addition, EAs maintain a population of ANNs during evolution, enabling large exploration. Therefore, we used neuroevolution for ANNs optimization, as it is a rather competitive and robust approach that performs global and adaptive optimization with the goal of more efficiently exploring the search space and, thus, finding better performance ANNs.

1.5 Hybridization

The efficiency of ANNs optimization can be improved by hybridizing neuroevolution approach and gradient-descent-based algorithms. We considered a hybrid optimization approach by incorporating a gradient-descent-based method in the evolution (ie, a hybridization that combines global and local search capabilities throughout the evolution).

1.6 Hybrid Evolutionary Algorithm

EAs are AI-based metaheuristics that emulate natural evolution, and have shown their robustness and flexibility to address the optimization of complex problems, even performing competitively with state-of-the-art reinforcement learning algorithms [11]. EAs use a search scheme based on a population of individuals (ANNs in this case), which are evolved simultaneously throughout the evolutionary process (search space exploration) by means of mechanisms inspired in nature. The evolution is based on randomly generated decisions, and a fitness function evaluates ANNs to guide the process toward better performance ANNs.

The pseudocode of the Hybrid Evolutionary Algorithm (HEA) used for ANNs optimization is presented in Algorithm 1 and described below.

Algorithm 1 Hybrid Evolutionary Algorithm

```
create a random population of size $Np$ of ANNs, using the structure defined in equation 2
$i = 0$
repeat
    calculate the fitness of every ANN
    rank the ANNs according to their fitness
    if $i \mod 20 = 0$ then
        apply local search optimization to the best ANN
    end if
    replicate the best 10% of the ANNs that substitutes the worst 10%
    perform parametric mutation to the best 10% of the ANNs
    perform structural mutation to the remaining 90% of the ANNs
    $i = i + 1$
until the stopping criteria are met
return the best optimized ANN
```

The HEA starts creating a random population of ANNs. After that, the population is optimized through the evolution, which is an iterative process including the following actions: calculation of ANNs fitness (equation 8) and their ranking, optimization of the best ANN applying local search, substitution of the worst 10% of ANNs by a replica of the best 10% of ANNs, and parametric and structural mutations of ANNs. The HEA finishes the evolution when the stopping criteria are fulfilled, returning the best optimized ANN. Crossover is not considered, as it may be inefficient for ANNs optimization [12, 13].

The main actions related to the evolution of ANNs are described bellow:

- Parametric mutation: updates the synaptic weights of the connections of each ANN by adding Gaussian noise, and whose variance gradually decreases as evolution progresses to control the strength of changes. Hence, this type of mutation optimizes the values of $\mathbf{W}$ and $\boldsymbol{\beta}$ parameters of each ANN (equation 2, more details are provided in Section 1.9).
- Structural mutation: aims to optimize the structure of ANNs (ie, number of neurons in hidden layer and their connections to contiguous layers) enabling the HEA to explore and discover new structures while providing ANNs with the way of dynamically adapting to their environment as evolution takes place. This type of mutation also aims to maintain a diverse population of ANNs, favoring large exploration of the search space while avoiding getting stuck in possible local minima. To this end, the HEA applies the following types of structural mutations: neuron deletion, neuron addition, neuron fusion, connection deletion and connection addition. So, structural mutation optimizes for each ANN the number m of basis functions in hidden layer along with their connections (equation 2, more details are provided in Section 1.9).
- Local search optimization: applies local optimization (ie, fine tuning of the values of $\mathbf{W}$ and $\boldsymbol{\beta}$ parameters estimated throughout parametric mutation) every 20 generations to the best ANN of the population. The aim of this scheme of optimization is a compromise between efficacy and efficiency, since one ANN is optimized each time, but throughout the evolution, thus intensifying the exploitation of the global search to be able to find better and near-optimal ANNs. The method used for this task is the *iRprop+* algorithm [14], which is an adaptive gradient-based algorithm that implements a weight-backtracking scheme. The *iRprop+* is considered a robust algorithm with minimal parameterization

and has shown great performance in optimizing ANNs of arbitrary structure (more details can be found in Section 1.10).

To evaluate the ANNs, one could consider cross-entropy as a measure of performance, which is commonly used in classification problems. However, in imbalanced classification problems, some classes are underrepresented in the training dataset, and cross-entropy can mask a poor performance for the minority class, which is precisely the most important one. In our case, patients who experienced the primary outcome within the 90 days in the waiting list (class $\mathcal{C}_+$) are much less frequent than patients who did not (class $\mathcal{C}_-$). The Minimum Sensitivity (MS) metric [15] was proposed to correctly estimate the performance in the minority class, as it calculates the minimum value of the sensitivity among all classes. In this way, in order to increase the ANNs performance taking into account the minority class, we used a continuous adaptation of the MS metric based on the maximum cross-entropy individually calculated for each class (ie, the maximum error obtained among both classes).

On the one hand, the cross-entropy of the positive class $\mathcal{C}_+$ was defined as:

$$\begin{aligned} L_{\mathcal{C}_+}(\mathbf{X},\mathbf{W},\boldsymbol{\beta}) &= -\frac{1}{n_{\mathcal{C}_+}}\sum_{\mathbf{x}_i|y_i=1}\left(\left(y_i\log\left(g\left(\mathbf{x}_i,\mathbf{W},\boldsymbol{\beta}\right)\right)\right)+\left(1-y_i\right)\log\left(1-g\left(\mathbf{x}_i,\mathbf{W},\boldsymbol{\beta}\right)\right)\right) \\ &= -\frac{1}{n_{\mathcal{C}_+}}\sum_{\mathbf{x}_i|y_i=1}\log\left(g\left(\mathbf{x}_i,\mathbf{W},\boldsymbol{\beta}\right)\right), \end{aligned} \tag{5}$$

where $\mathbf{X}^T = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, and $n_{\mathcal{C}_+}$ denotes the number of patients from the positive class.

Similarly, the cross-entropy of the negative class $\mathcal{C}_-$ was defined as:

$$L_{\mathcal{C}_-}(\mathbf{X},\mathbf{W},\boldsymbol{\beta}) = -\frac{1}{n_{\mathcal{C}_-}}\sum_{\mathbf{x}_i|y_i=0}\log\left(1-g\left(\mathbf{x}_i,\mathbf{W},\boldsymbol{\beta}\right)\right), \tag{6}$$

where $n_{\mathcal{C}_-}$ denotes the number of patients from the negative class.

Then, the maximum cross-entropy error function was formulated as follows:

$$L_{\text{Max}}(\mathbf{X},\mathbf{W},\boldsymbol{\beta}) = \max\left\{L_{\mathcal{C}_+}(\mathbf{X},\mathbf{W},\boldsymbol{\beta}), L_{\mathcal{C}_-}(\mathbf{X},\mathbf{W},\boldsymbol{\beta})\right\}. \tag{7}$$

Therefore, as the goal of the HEA is to optimize the ANNs by maximizing their performance throughout the evolution, the fitness function used by the HEA to assess ANNs performance was formulated as a strictly decreasing transformation of the maximum cross-entropy error function:

$$A(\mathbf{X},\mathbf{W},\boldsymbol{\beta}) = \frac{1}{1+L_{\text{Max}}(\mathbf{X},\mathbf{W},\boldsymbol{\beta})}, \quad (0 < A(\mathbf{X},\mathbf{W},\boldsymbol{\beta}) \leq 1). \tag{8}$$

1.7 Experimental setup

Since this study aimed to propose one ANN model for waiting list liver transplantation (LT) prioritization, we performed a large exploration and exploitation (hybridizing global and local search) to find a competitive and interpretable model.

The settings used for the main parameters of the HEA involved in the optimization of the ANNs were as follows. The size of ANNs population was Np=$1,000$ to perform an extensive exploration, and for the random initialization of each ANN: the range to select the initial number of neurons was $[1, 3]$, the range to initialize the synaptic weights of the connections between input and hidden layers was $[-10, 10]$, whereas $[-5, 5]$ was the range to initialize the synaptic weights for the connections between hidden and output layers. For the evolution of each ANN: the range to both select the number of connections and neurons to add or delete by structural mutation was $[1, 2]$, with 4 being the maximum number of neurons in hidden layer that an ANN could have. Regarding the *iRprop+* local optimizer, applied every 20 generations to the best ANN of the population, the paremeter configuration was $\eta^+ = 1.2$, $\eta^- = 0.2$, $\Delta_0 = 0.0125$, $\Delta_{min} = 0$, $\Delta_{max} = 50$ and *Epochs* $= 500$. As for the stopping criterion, it was set at a maximum of 500 generations or a maximum of 10 generations without improving the best fitness, whichever occurred first.

Prior to the evolution, the input variables were scaled in the interval $[-1, 1]$ using the following expression:

$$x^* = 2 \cdot \frac{x - \min(x)}{\max(x) - \min(x)} - 1, \tag{9}$$

where x is the input variable being scaled, and $\min(x)$ and $\max(x)$ are the minimum and the maximum values of x from the training dataset, respectively.

We performed 40 independent runs of the HEA due to its stochastic search and also to carry out a large exploration. The optimized ANN that obtained, out of all local optimizations from all runs, the best discrimination by means of the highest Area under the ROC Curve (AUC) in the training dataset was selected as the GEMA-AI model.

1.8 GEMA-AI output transformation

To compare the performance of GEMA-AI with the reference models for LT prioritization, the output of GEMA-AI (defined in equation 2) was transformed to follow the same distribution of values than GEMA-Na, MELD 3.0 and MELD-Na (ie, $[6, 40]$). To this end, and using the training dataset, the $25th$ and $75th$ quartiles of the GEMA-AI outputs were matched to those of the MELD-Na, as it is the most frequently used model for LT allocation. Then, the values of the coefficients of $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_m)$ corresponding to the linear part of the model were transformed accordingly. GEMA-AI score for each patient was rounded to the nearest integer and fitted to the range $[6, 40]$.

1.9 Parametric and structural mutations

The goal of mutation mechanisms is to maintain a diverse population of ANNs and to properly exploit them during evolutionary process. Each ANN in the population is mutated independently. To control the intensity of ANNs mutations throughout their evolution, each ANN in the population has an associated temperature, which is expressed as follows:

$$T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}) = 1 - A(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}), \qquad 0 \leq T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}) < 1, \tag{10}$$

where $A(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta})$ is the fitness function defined in equation 8 .

Concerning the parametric mutation, it is applied to the best 10% of the ANNs in the population. It updates the synaptic weights of the connections of each ANN by adding Gaussian noise and whose variance gradually decreases as a function of the associated temperature to the ANN being mutated, to control the intensity of changes as its evolution progresses. Hence, this adaptive variance enables the HEA to shift from exploring ANNs to exploiting them.

On the one hand, and following the notation used in equation 2, the synaptic weights of the connections linking the input and hidden layers are altered as described next:

$$w_{jk}(t+1) = w_{jk}(t) + \xi_1(t), \quad k = 1, \ldots, 4, \quad j = 1, \ldots, m, \tag{11}$$

where $\xi_1(t) \in N(0,\ \alpha_1(t) \cdot T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}))$ is a random number obtained from a one dimension normal distribution of 0 mean and $\alpha_1(t) \cdot T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta})$ variance. The objective is to gradually decrease the intensity of the parametric mutations applied to each ANN as its performance increases. This adaptive mutation is also controlled using the parameter $\alpha_1(t)$, which will be explained later.

On the other hand, the synaptic weights of the connections linking the hidden and output layers are updated as described next:

$$\beta_j(t+1) = \beta_j(t) + \xi_2(t), \quad j = 1, \ldots, m, \tag{12}$$

where $\xi_2(t) \in N(0,\ \alpha_2(t) \cdot T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}))$ is similar to $\xi_1(t)$, but using $\alpha_2(t)$ as the variance control parameter.

Once each ANN is parametrically mutated, its fitness is calculated again, and then, using a simulated annealing strategy [16], the mutation is finally accepted or discarded. Specifically, considering ΔA as the difference between the fitness of the ANN before and after being mutated, the mutation is always accepted if $\Delta A \geq 0$. Conversely, if the fitness of the ANN mutated is worse than before being mutated, the mutation is accepted with a probability given by $\exp(\Delta A / T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}))$.

The mentioned $\alpha_1(t)$ and $\alpha_2(t)$ parameters are aimed to determine the intensity of parametric mutations through the evolution of ANNs, enabling them to infer learning rules from their environment. To this end, both parameters are dynamically adapted through evolutionary process to avoid getting stuck in possible local minima and to speed up the convergence of ANNs as long as the conditions of the search space are appropriate. The adaptation of $\alpha_1(t)$ and $\alpha_2(t)$ is defined as follows:

$$\alpha_k(t) = \begin{cases} (1+\lambda) \cdot \alpha_k(t) & \text{if} \quad A(\mathbf{X}, \mathbf{W}_g, \boldsymbol{\beta}_g) > A(\mathbf{X}, \mathbf{W}_{g-1}, \boldsymbol{\beta}_{g-1}) \ \forall g \in \{t, t-1, \ldots, t-\rho\} \\ (1-\lambda) \cdot \alpha_k(t) & \text{if} \quad A(\mathbf{X}, \mathbf{W}_g, \boldsymbol{\beta}_g) = A(\mathbf{X}, \mathbf{W}_{g-1}, \boldsymbol{\beta}_{g-1}) \ \forall g \in \{t, t-1, \ldots, t-\rho\} \\ \alpha_k(t), & \text{otherwise} \end{cases} \tag{13}$$

where $k \in \{1, 2\}$, $A(\mathbf{X}, \mathbf{W}_g, \boldsymbol{\beta}_g)$ corresponds to the fitness of the best ANN in generation g, and λ and ρ are parameters that manage the change. The values used for these parameters were $\alpha_1(0) = 0.5$, $\alpha_2(0) = 1$, $\lambda = 0.1$, and $\rho = 10$. The reason for using the scheme defined in equation 13 is following explained: a successful generation indicates that the best current ANN is better than the best ANN of the previous generation. When this happens ρ consecutive times it means that the best ANNs could be located at the region of the search space being explored. In this situation, the intensity of mutations is raised with the purpose of finding better and near-optimal ANNs. On the contrary, if the best ANN remains the same for ρ times, the intensity of mutations is decreased. Otherwise, the intensity of mutations remains unchanged.

As for structural mutation, which is applied to the remaining 90% of the ANNs in the population, it consists in modifying ANNs structure by deleting or adding hidden neurons and their connections linking the input and output layers. Hence, this type of mutation enables the HEA to perform a large exploration of the search space, maintaining the diversity in the population, and allows ANNs to dynamically adapt to their environment.

The HEA considers five types of structural mutations: neuron deletion, neuron addition, neuron fusion, connection deletion and connection addition. These types of structural mutations are consecutively applied to each ANN in the population using a probability given by $T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta})$. If none of the five structural mutations is applied because of the probability, one of them is randomly chosen and applied. Then, the fitness of the ANN mutated is calculated again.

The structural mutations related to the connections of ANNs are described next:

- Connection addition. A new connection is created between two randomly chosen neurons from contiguous layers using a random synaptic weight. This mutation is firstly applied to connect neurons between the input and hidden layers, and then between neurons of the hidden and output layers.
- Connection deletion. A connection between neurons of contiguous layers is chosen at random and deleted. This mutation is also performed on all contiguous layers.

The number of connections involved in both mutations is calculated as $\Delta_{min} + u \cdot T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}) \cdot [\Delta_{max} - \Delta_{min}]$, where u represents a number generated at random in the range $[0, 1]$, and Δ_{min} and Δ_{max} are user-defined values (specified in Section 1.7) that represent the minimum and maximum number of connections that can be mutated at a time, respectively.

Finally, the structural mutations related to the neurons of ANNs are described below:

- Neuron addition. A new neuron is added in the hidden layer and connected to two neurons chose at random, one of them from the input layer and the other from the output layer. As for the synaptic weights, they are randomly chosen from user-defined ranges $[-I, I]$ and $[-O, O]$ (specified in Section 1.7) for the connections from input to hidden and from hidden to output layers, respectively.
- Neuron deletion. One hidden neuron is selected at random and deleted along with its connections.
- Neuron fusion. Two hidden neurons, a and b, are chosen at random and fused into a single hidden neuron c. Their connections in common are maintained, recalculating the synaptic weights in the following way:
$$\beta_c = \beta_a + \beta_b, \qquad w_{ck} = \frac{w_{ak} + w_{bk}}{2}. \tag{14}$$
The connections that are not common to the neurons being fused are inherited by c with a probability of 0.5, and their synaptic weights remain unchanged.

The number of neurons involved in neuron addition and deletion is obtained as $\Delta_{min} + u \cdot T(\mathbf{X}, \mathbf{W}, \boldsymbol{\beta}) \cdot [\Delta_{max} - \Delta_{min}]$, with Δ_{min} and Δ_{max} being user-defined values (described in Section 1.7) indicating the minimum and maximum number of neurons that can be mutated at a time, respectively.

Finally, all the mutations applied to each ANN are accepted if the mutated ANN is valid. On the contrary, if the mutated ANN is invalid, the mutations are discarded, and another parametric or structural mutation is chosen at random and applied to the original ANN, avoiding the use of repair mechanisms.

1.10 Local search optimization

The purpose of local optimization is to fine adjust the synaptic weights of the ANNs connections, which are estimated throughout parametric mutation, with the aim of finding better and near-optimal ANNs. This optimization is applied each 20 generations to the best ANN of the population. The aim of this scheme of optimization is a compromise between efficacy and efficiency, since one ANN is optimized each time, but throughout the evolution, thus intensifying the exploitation of the global search. The method used for local optimization is the *iRprop+* algorithm [14], which is an adaptive gradient-based algorithm that implements a weight-backtracking scheme.

The *iRprop+* algorithm uses individual step-sizes for updating each ANN connection (ie, the change amount for a particular synaptic weight is individually adapted during the optimization process with the aim of minimizing variations and maximizing the update change). *iRprop+* requires a minimal parameterization, which controls the step-size adaption along the optimization process, as detailed next:

- Δ_0: initial step-size.
- Δ_{min}: step-size lower bound.
- Δ_{max}: step-size upper bound.
- η^+: step-size increment factor.
- η^-: step-size reduction factor.

The value of each parameter is given in Section 1.7. After each weight update, *iRprop+* decides whether or not to discard the update using both the sign of the partial derivative and the evolution of the ANN error, thus taking advantage of local and global information, respectively. The *iRprop+* is considered a robust algorithm with respect to its parameters, and has shown great performance in optimizing ANNs of arbitrary structure.

2 Supplementary Results

2.1 GEMA-AI model explanation

GEMA-AI (table 2 of the manuscript) comprises four basis functions (B_1 to B_4), each one with its own coefficient or synaptic weight ($-25.210, 21.395, 10.592$ and -7.108, respectively), and the bias (29.981). Each basis function interacts with input variables by a sigmoidal function defined in equation 3, which performs a non-linear transformation of the input variables. For example, basis function B_3 interacts with sodium* and RFH-GFR* (the * means that input variable was previously scaled using equation 9) using the coefficients 15.596 and 4.996, respectively, and the bias -5.524, as follows:

$$\begin{aligned} &B_3([\text{sodium}^*,\text{RFH-GFR}^*],[15.596,4.996,-5.524]) = \\ &= \frac{1}{1+e^{-(15.596\times\text{sodium}^*+4.996\times\text{RFH-GFR}^*-5.524)}}, \quad (0 \le B_3 \le 1)\,. \end{aligned} \tag{15}$$

Therefore, GEMA-AI is a linear model of non-linear basis functions whose specific contribution is detailed below.

As shown in equations 3 and 15, the result of each basis function B_i is in the range $[0, 1]$. Therefore, as the result of B_i tends to 1 its contribution increases and, conversely, its contribution decreases as the result tends to 0. Thus, the higher the absolute value of the coefficient of B_i, the greater the contribution of B_i to the GEMA-AI score. B_1 is the basis function with the highest contribution in GEMA-AI model, whereas B_4 provides the lowest contribution, as their coefficients are -25.210 and -7.108, respectively (table 2 of the manuscript). In addition, the positive coefficients (+) of B_2 and B_3 mean that increments in the result of each basis function lead to increments in the GEMA-AI score. Conversely, the negative coefficients ($-$) of B_1 and B_4 lead to reductions in the GEMA-AI score as their results increase.

Noteworthy, B_1 and B_4 comprise information of all input variables, whereas B_2 and B_3 include only some of them ([RFH-GFR*, sodium*, bilirubin*] and [sodium*, RFH-GFR*], respectively). Besides, the contribution of each input variable differs from one basis function to another according to the absolute value of the coefficient. For instance, bilirubin* has the highest contribution in B_1 and B_4, and the lowest in B_2. For RFH-GFR*, the highest contribution is in B_2 and the lowest in B_1. This aspect highlighted the non-linear relationship between the input variables and the risk of the primary outcome, so that each basis function focuses on specific interactions between input variables. This non-linearity was studied by deriving smoothing splines from the training cohort [17].

Supplementary table 1 summarizes the effect of each input variable in each basis function and, consequently, their effect on the GEMA-AI score. The way in which a particular variable impact on the GEMA-AI score depends on the sign of two coefficients: that of the input variable and that of the basis function. For instance, in the basis function B_1, which shows itself a negative coefficient ($-$), the positive coefficients (+) of sodium* and RFH-GFR* indicate that the higher the values of these variables, the higher the result of B_1 and therefore, GEMA-AI score decreases. The opposite effect occurs with bilirubin* and INR* due to their negative coefficients: the higher their values, the lower the result of B_1, which leads to increments in the GEMA-AI score. Besides, the basis function B_3, which shows itself a positive coefficient, the positive coefficients of sodium* and RFH-GFR* mean a direct relationship with the GEMA-AI score, contrary to the effect observed in the basis function B_1.

As bilirubin* increases, the result of B_1 decreases and that of B_2 increases, contributing in both cases to

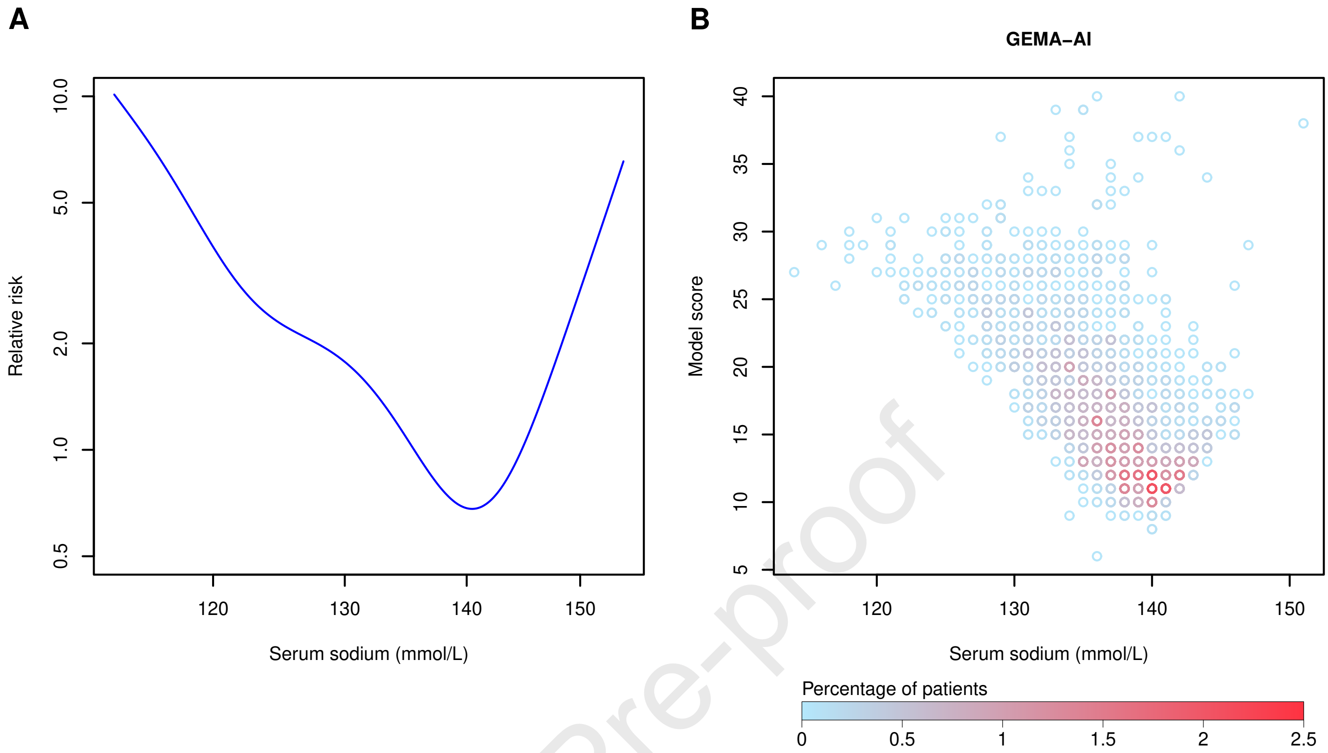
Supplementary Table 1: Behavior of input variables associated with each basis function and its effect on GEMA-AI score.

	Basis function											
	$B_1(-)$			$B_2(+)$			$B_3(+)$			$B_4(-)$		
Input variable	Sign	B_1	GEMA-AI	Sign	B_2	GEMA-AI	Sign	B_3	GEMA-AI	Sign	B_4	GEMA-AI
bilirubin*	$(-)$	↓	↑	$(+)$	↑	↑				$(+)$	↑	↓
INR*	$(-)$	↓	↑							$(-)$	↓	↑
sodium*	$(+)$	↑	↓	$(+)$	↑	↑	$(+)$	↑	↑	$(+)$	↑	↓
RFH-GFR*	$(+)$	↑	↓	$(-)$	↓	↓	$(+)$	↑	↑	$(+)$	↑	↓

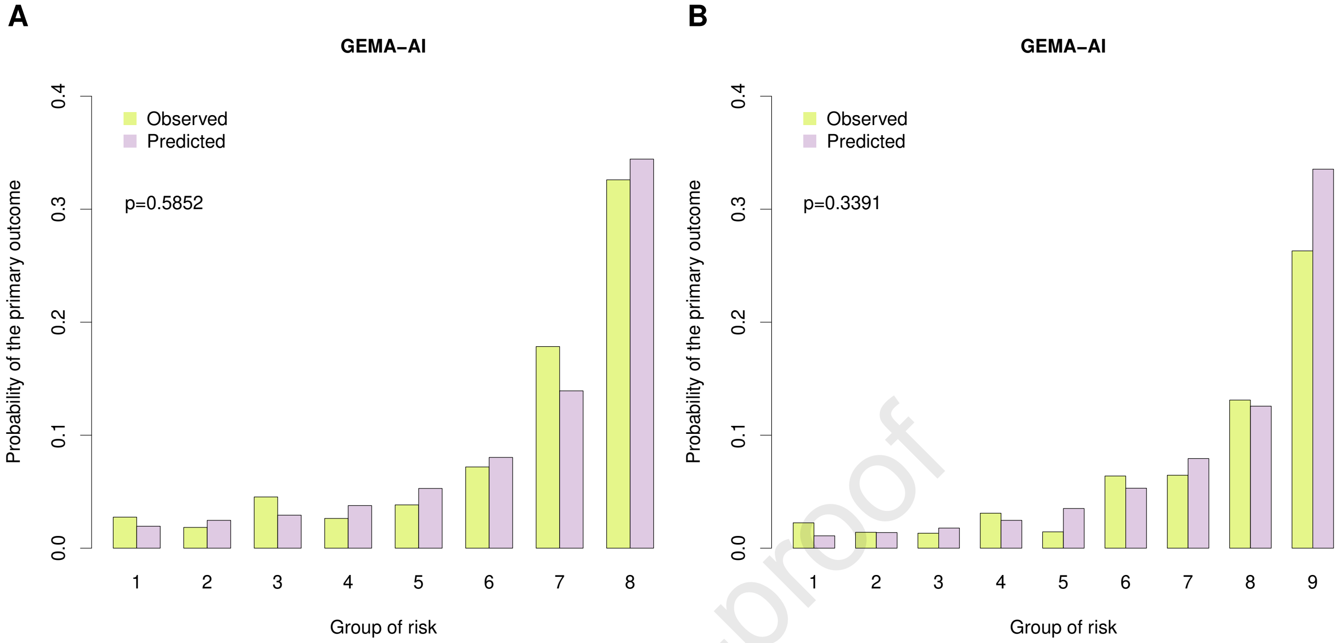
The * means that each input variable was previously scaled in the range $[-1, 1]$ using equation 9. A positive coefficient is represented by $(+)$, whereas $(-)$ represents a negative coefficient. ↑ means that B_i result or GEMA-AI score increments as the value of the input variable increases, the opposite behavior is represented by ↓ .

higher GEMA-AI score. The opposite behavior of bilirubin* is observed in B_4, causing reductions in GEMA-AI score, so it can be said that bilirubin* in B_4 behaves as a modulation in GEMA-AI score increments. Concerning INR*, when its value increases, the result of B_1 and B_4 decrease, thus increasing the GEMA-AI score. Regarding the RFH-GFR*, as its value increases, the result of B_1 and that of B_4 increase, whereas the result of B_2 decreases, however, in all three cases GEMA-AI score decreases (supplementary table 1). Conversely, the result of B_3 increases and so does the GEMA-AI score. Considering the absolute value of the coefficient of each basis function, a decrease of the GEMA-AI score would be expected, with RFH-GFR* in B_3 behaving as a modulation of this reduction.

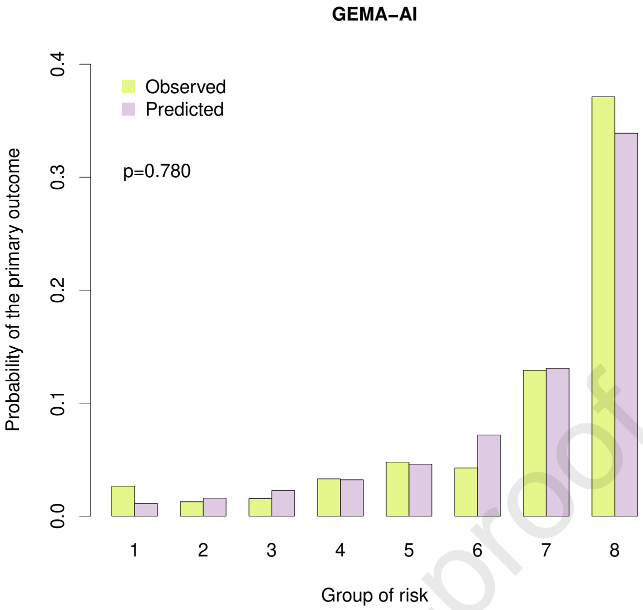
The analysis of the behavior of sodium* is particularly interesting due to its "U" shape relationship with the primary outcome, which is decreasing monotonous for sodium values lower than 140 mmol/L and increasing monotonous from this value (supplementary figure 1A). As sodium* increases, the result of the four basis functions increases and especially in B_3 due to its higher coefficient of sodium*. However, GEMA-AI score only increases in B_2 and B_3, decreasing in B_1 and B_4 (supplementary table 1). The reason for this behavior is that for intermediate values of sodium, GEMA-AI score decreases as a consequence of B_1 and B_4. Conversely, for extreme values, either high or low, GEMA-AI score increases as a consequence of B_2 and especially B_3. The basis functions of GEMA-AI allow to infer the non-linear relationship between sodium and the primary outcome (supplementary figure 1B) whereas linear models cannot. As shown in supplementary figures 11, 12, and 13, MELD-Na, MELD 3.0 and GEMA-Na scores reach a plateau at a sodium value of 140 mmol/L whereas GEMA-AI score continue raising above this threshold.



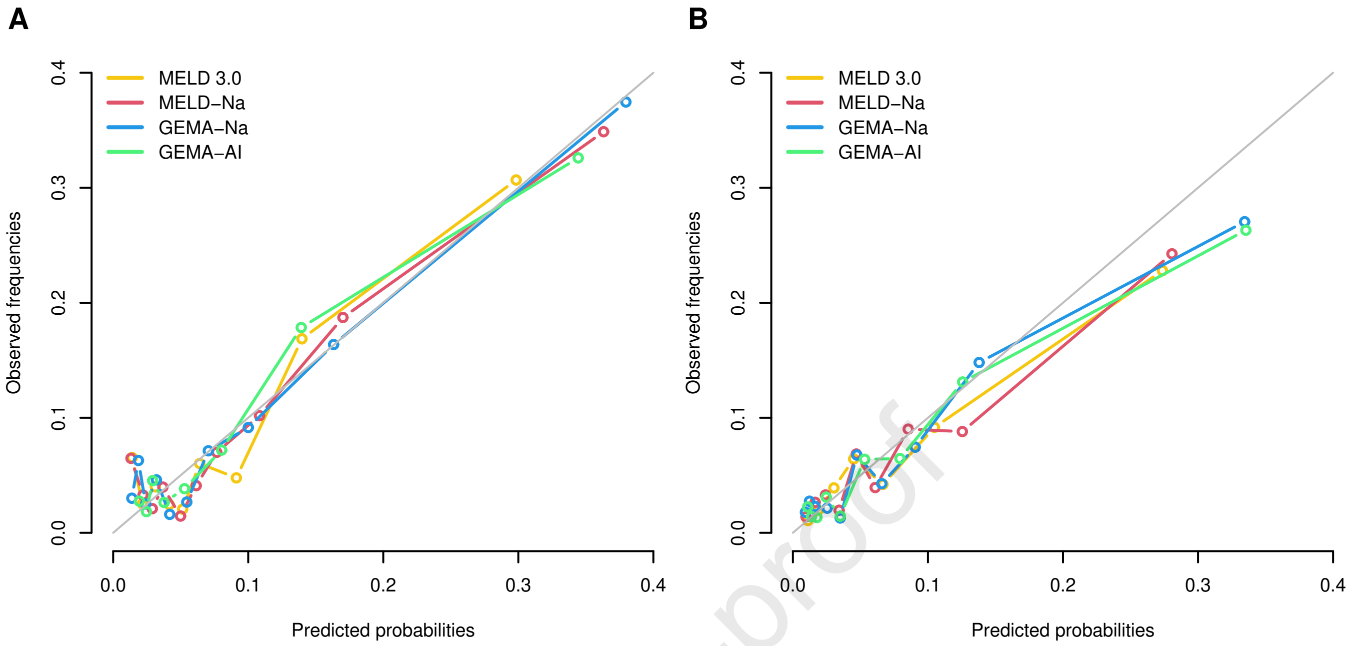
Supplementary Figure 1: Comparison of the non-linear relationship of sodium with the risk of the primary outcome vs the response of the GEMA-AI model to sodium. **(A)** Relationship between sodium and the risk of the primary outcome (smoothing spline plotted with the blue line) obtained from the training cohort (n=5, 762), using generalized additive models. **(B)** Response of the GEMA-AI model to sodium in the internal validation cohort (n=1, 920). Each circle represents a patient GEMA-AI score for a specific sodium value. The color of each circle denotes the percentage of patients matching both score and sodium value.



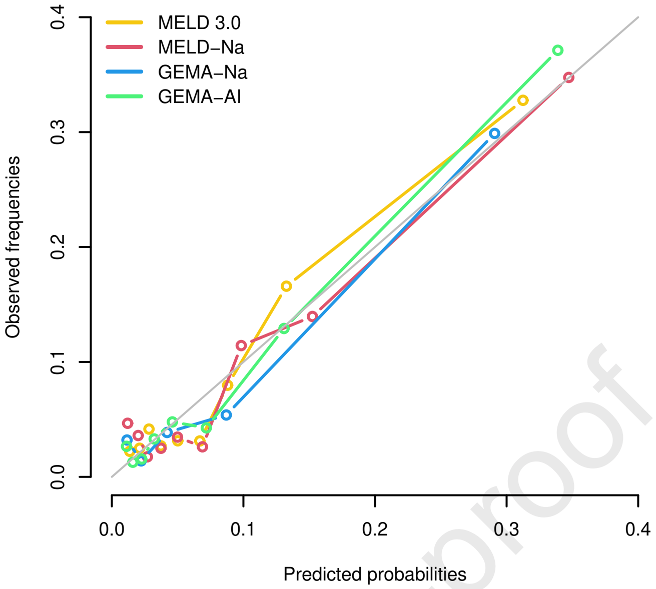
Supplementary Figure 2: Bar calibration diagrams of the GEMA-AI model. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p values correspond to the Greenwood-Nam-D'Agostino test. **(A)** Internal validation cohort (n=1, 920). **(B)** External validation cohort (n=1, 638).



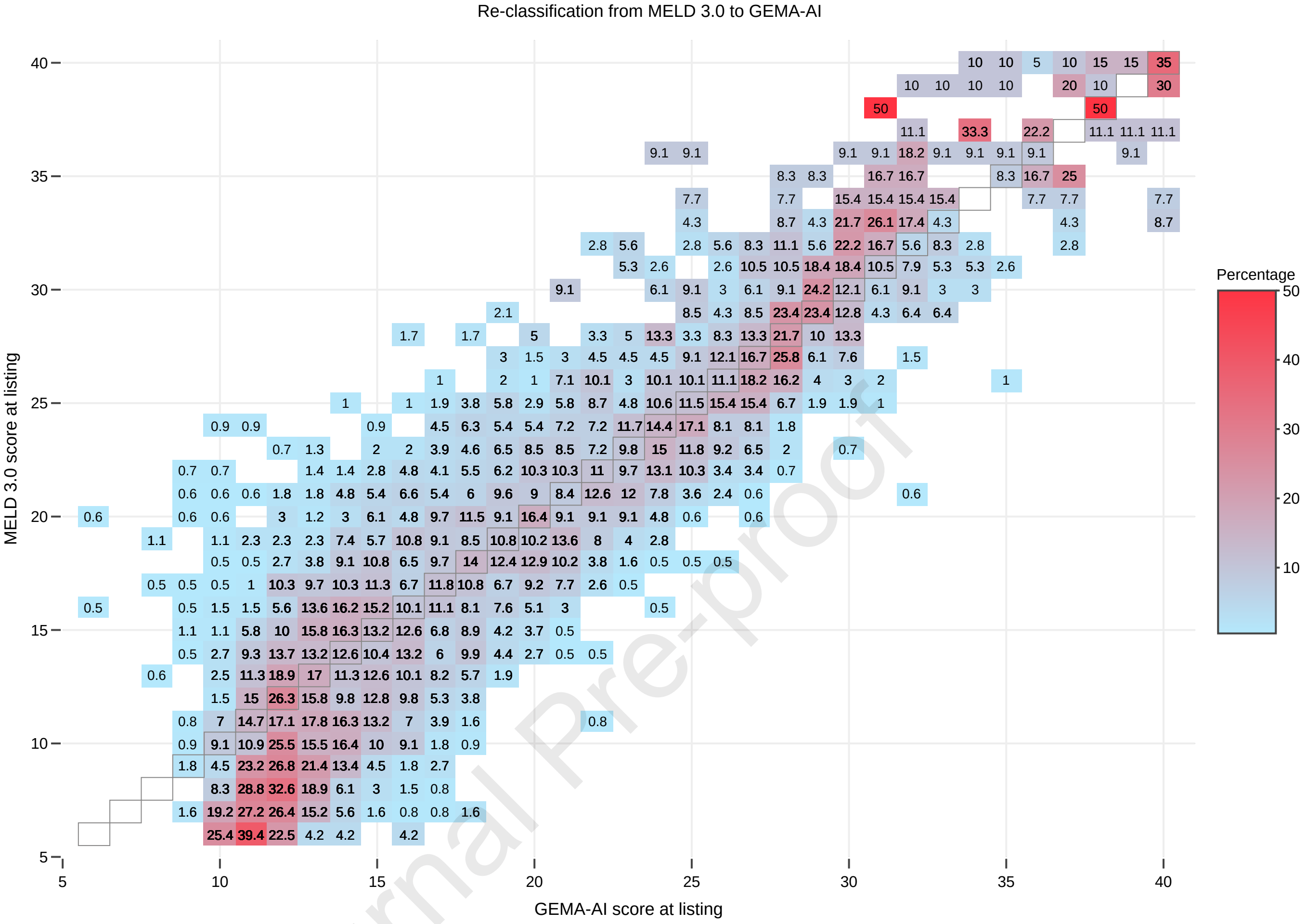
Supplementary Figure 3: Bar calibration diagram of the GEMA-AI model in women. In this analysis, the subgroups of women from the internal and external validation cohorts were combined to have a sufficient number of events to allow meaningful calibration (n=1, 055). The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p value corresponds to the Greenwood-Nam-D'Agostino test.



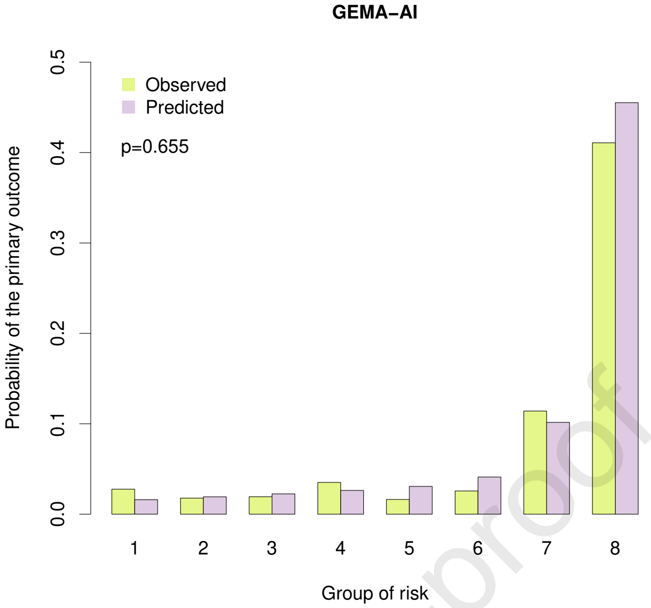
Supplementary Figure 4: Linear calibration diagrams of the GEMA-AI, GEMA-Na, MELD-Na and MELD 3.0 models. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. Each circle represents the intersection between the observed and predicted probabilities of each group. The gray diagonal line simulates a perfect calibration. GEMA-Na=Gender-Equity Model for liver Allocation corrected by serum sodium. MELD-Na=Model for End-stage Liver Disease corrected by serum sodium. **(A)** Internal validation cohort (n=1, 920), albumin data was not available for 136 patients from this cohort and were excluded from MELD 3.0 calibration analysis. **(B)** External validation cohort (n=1, 638).



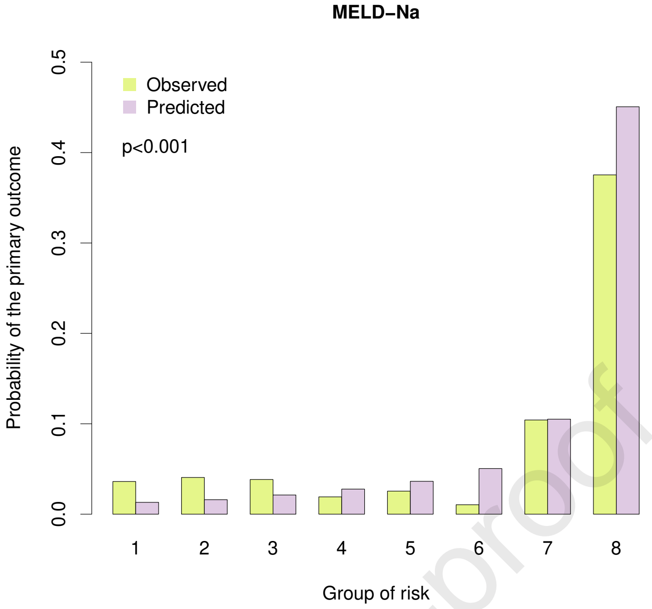
Supplementary Figure 5: Linear calibration diagrams of the GEMA-AI, GEMA-Na, MELD-Na and MELD 3.0 models in women. In this analysis, the subgroups of women from the internal and external validation cohorts were combined to have a sufficient number of events to allow meaningful calibration (n=1, 055). Albumin data was not available for 50 patients and were excluded from MELD 3.0 calibration analysis. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. Each circle represents the intersection between the observed and predicted probabilities of each group. The gray diagonal line simulates a perfect calibration. GEMA-Na=Gender-Equity Model for liver Allocation corrected by serum sodium. MELD-Na=Model for End-stage Liver Disease corrected by serum sodium.



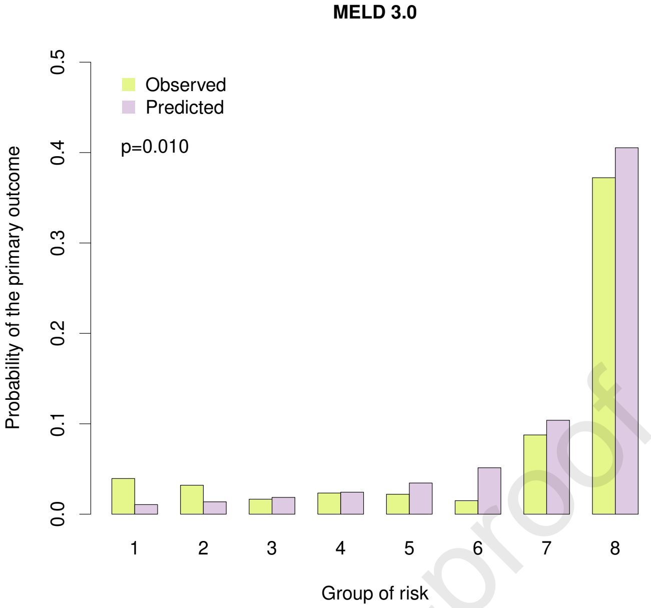
Supplementary Figure 6: Re-classification diagram showing patients prioritization of GEMA-AI vs MELD 3.0 merging the internal and the external validation cohorts (n=3, 422). The number in each box represents the percentage of GEMA-AI for a specific MELD 3.0 score value. The diagonal, matching score values of both models, is represented by a gray frame. Above diagonal values represent lower GEMA-AI scores compared with MELD 3.0, whereas below diagonal values higher GEMA-AI scores compared with MELD 3.0. Albumin data was not available for 136 patients and were excluded from this analysis.



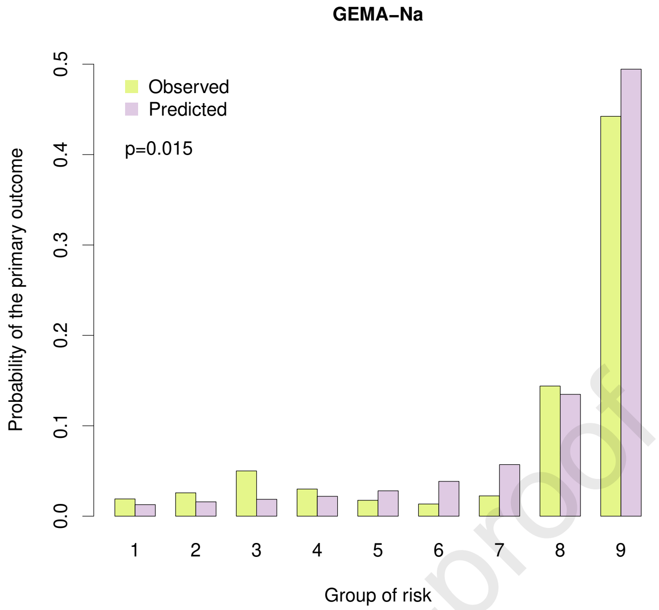
Supplementary Figure 7: Bar calibration diagram of the GEMA-AI model in patients with at least one extreme analytical value. The internal and the external validation cohorts were merged (n=3, 558) and a total of 1, 403 patients (39.4%) were eligible for this analysis. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p value corresponds to the Greenwood-Nam-D'Agostino test.



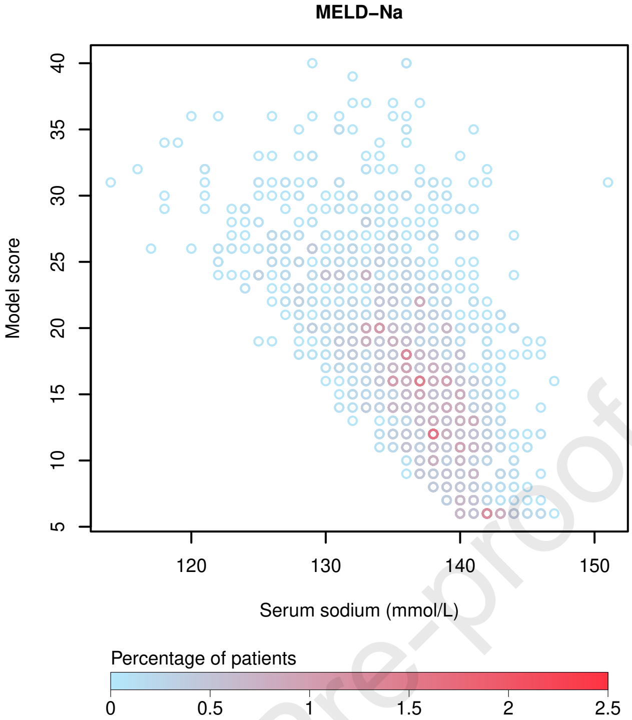
Supplementary Figure 8: Bar calibration diagram of the MELD-Na model in patients with at least one extreme analytical value. The internal and the external validation cohorts were merged (n=3, 558) and a total of 1, 403 patients (39.4%) were eligible for this analysis. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p value corresponds to the Greenwood-Nam-D'Agostino test.



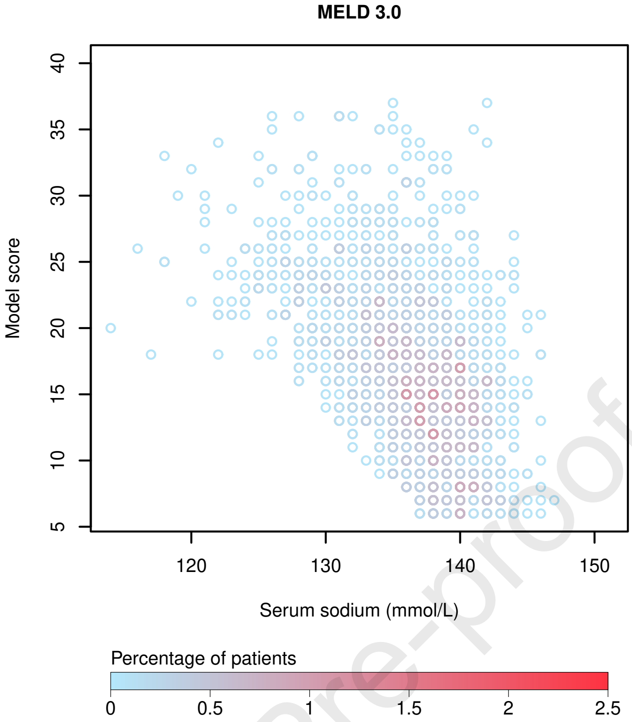
Supplementary Figure 9: Bar calibration diagram of the MELD 3.0 model in patients with at least one extreme analytical value. The internal and the external validation cohorts were merged (n=3, 422) and a total of 1, 337 patients (39.1%) were eligible for this analysis. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p value corresponds to the Greenwood-Nam-D'Agostino test. Albumin data was not available for 136 patients and were excluded from this analysis.



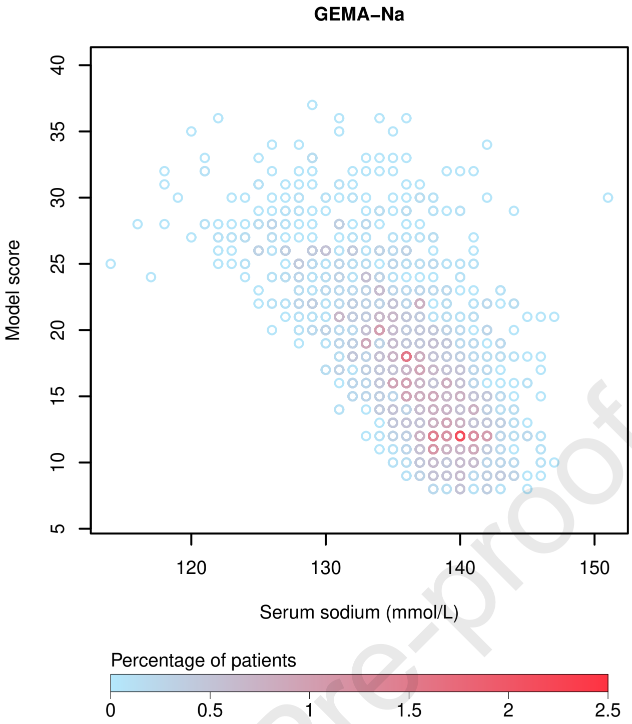
Supplementary Figure 10: Bar calibration diagram of the GEMA-Na model in patients with at least one extreme analytical value. The internal and the external validation cohorts were merged (n=3, 558) and a total of 1, 403 patients (39.4%) were eligible for this analysis. The observed and predicted probabilities for the primary outcome are showed. Deciles of risk were merged into risk groups if necessary to allow at least two events in each group. The p value corresponds to the Greenwood-Nam-D'Agostino test.



Supplementary Figure 11: Response of the MELD-Na model to sodium in the internal validation cohort (n=1, 920). Each circle represents a patient MELD-Na score for a specific sodium value. The color of each circle denotes the percentage of patients matching both score and sodium value.



Supplementary Figure 12: Response of the MELD 3.0 model to sodium in the internal validation cohort (n=1, 784). Each circle represents a patient MELD 3.0 score for a specific sodium value. The color of each circle denotes the percentage of patients matching both score and sodium value. Albumin data was not available for 136 patients and were excluded from this analysis.



Supplementary Figure 13: Response of the GEMA-Na model to sodium in the internal validation cohort (n=1, 920). Each circle represents a patient GEMA-Na score for a specific sodium value. The color of each circle denotes the percentage of patients matching both score and sodium value.

Supplementary Table 2: Harrells' concordance statistics and 95% confidence intervals (in brackets) for each model in the different subgroups of interest in the whole study cohort.

Cohort	n	MELD-Na	MELD 3.0	GEMA-Na	GEMA-AI
Alcoholic liver disease	3,257	0.736 (0.696-0.775)	0.734 (0.694-0.775)	0.753 (0.714-0.792)	0.758 (0.719-0.797)
Chronic hepatitis C	1,923	0.775 (0.724-0.826)	0.777 (0.726-0.829)	0.782 (0.731-0.833)	0.796 (0.747-0.845)
NAFLD/cryptogenic cirrhosis	1,577	0.761 (0.705-0.816)	0.756 (0.699-0.812)	0.781 (0.728-0.833)	0.773 (0.719-0.828)
Primary biliary cholangitis	775	0.762 (0.680-0.844)	0.743 (0.657-0.829)	0.788 (0.710-0.865)	0.801 (0.726-0.875)
Primary sclerosing cholangitis	895	0.823 (0.748-0.898)	0.799 (0.712-0.886)	0.855 (0.786-0.924)	0.856 (0.790-0.922)

NAFLD: Non-alcoholic fatty liver disease.

Supplementary Table 3: Brier scores for the distinct models in each cohort of the study.

	GEMA-AI	GEMA-Na	MELD 3.0	MELD-Na
Cohort	Brier Score	Brier Score	Brier Score	Brier Score
Training (whole; n=5,762)	0.0516 (0.0455-0.0577)	0.0510 (0.0447-0.0570)	0.0536 (0.0471-0.0602)	0.0528 (0.0467-0.0590)
Training (women; n=1,955)	0.0483 (0.0385-0.0582)	0.0480 (0.0382-0.0579)	0.0531 (0.0421-0.0642)	0.0519 (0.0417-0.0621)
Internal validation (whole; n=1,920)	0.0541 (0.0435-0.0646)	0.0534 (0.0429-0.0639)	0.0525 (0.0415-0.0635)	0.0549 (0.0443-0.0656)
Internal validation (women; n=623)	0.0447 (0.0283-0.0611)	0.0461 (0.0294-0.0627)	0.0498 (0.0315-0.0681)	0.0491 (0.0319-0.0663)
External validation (whole; n=1,638)	0.0439 (0.0342-0.0536)	0.0454 (0.0356-0.0553)	0.0472 (0.0370-0.0574)	0.0474 (0.0372-0.0576)
External validation (women; n=432)	0.0439 (0.0250-0.0627)	0.0452 (0.0259-0.0645)	0.0491 (0.0284-0.0698)	0.0493 (0.0285-0.0700)

Brier Score (95% Confidence Interval). Each cohort was analyzed as a whole and also the subgroup of women separately. Albumin data was not available for 413 and 136 patients from the training and the internal validation cohorts, respectively, and were excluded from MELD 3.0 Brier score calculations. GEMA-Na=Gender-Equity Model for liver Allocation corrected by serum sodium. MELD-Na=Model for End-stage Liver Disease corrected by serum sodium.

Supplementary Table 4: Clinical characteristics of prioritized patients according to the model used. The p values denote comparisons between patients differently prioritized by GEMA-Na vs GEMA-AI.

Variable	Transplanted both (n=3,485)	GEMA-Na transplanted (n=240)	GEMA-AI transplanted (n=240)	p value
Age	53.08 ± 10.53	49.74 ± 12.18	56.35 ± 9.41	<**0.001**
Sex (women)	1,253 (36.0%)	69 (28.8%)	92 (38.3%)	**0.026**
Ascites (Moderate-severe)	1,805 (51.8%)	64 (26.7%)	121 (50.4%)	<**0.001**
Urea (mmol/L)	6.7 (IQR 4.7-10.0)	4.1 (IQR 3.20-5.28)	8.5 (IQR 6.4-11.3)	<**0.001**
Creatinine (μmol/L)	95.80 ± 48.47	66.13 ± 19.44	106.60 ± 34.59	<**0.001**
RFH-GFR (mL/min)	54.69 ± 21.82	84.32 ± 24.84	46.35 ± 14.61	<**0.001**
INR	1.78 ± 0.60	1.68 ± 0.34	1.27 ± 0.29	<**0.001**
Bilirubin (μmol/L)	92 (IQR 49-180)	87.02 (IQR 63.50-132.50)	23 (IQR 16.07-33.00)	<**0.001**
Sodium (mmol/L)	132.82 ± 4.96	137.01 ± 3.12	135.02 ± 4.58	<**0.001**
Albumin (g/L)[1]	30.02 ± 6.50	28.73 ± 5.33	33.11 ± 5.45	<**0.001**
Primary outcome[2]	396 (11.4%)	7 (2.9%)	16 (6.7%)	0.054

p values highlighted with boldface denote statistically significant differences.
RFH-GFR: Royal-Free Hospital Glomerular Filtration Rate.
INR: International normalized ratio.

[1] Albumin was not available in 206 patients in this analysis.

[2] Mortality or delisting due to clinical deterioration within the first 90 days after inclusion in the waiting list.

Supplementary Table 5: Clinical characteristics of prioritized patients according to the model used. The p values denote comparisons between patients differently prioritized by MELD 3.0 vs GEMA-AI. Patients transplanted with missing albumin at waitlist inclusion were excluded (n=198).

Variable	Transplanted both (n=2,933)	MELD 3.0 transplanted (n=594)	GEMA-AI transplanted (n=594)	p value
Age	52.58 ± 10.55	47.68 ± 13.22	57.51 ± 8.44	<**0.0001**
Sex (women)	1,059 (36.1%)	269 (45.3%)	216 (36.4%)	**0.002**
Ascites (Moderate-severe)	1,479 (50.4%)	110 (18.5%)	348 (58.6%)	<**0.0001**
Urea (mmol/L)	6.30 (IQR 4.40-9.60)	3.60 (IQR 2.90-4.62)	8.40 (IQR 6.5-11.0)	<**0.0001**
Creatinine (μmol/L)	93.36 ± 47.48	57.98 ± 15.51	100.63 ± 25.69	<**0.0001**
RFH-GFR (mL/min)	56.59 ± 22.70	94.26 ± 25.55	46.02 ± 11.69	<**0.0001**
INR	1.84 ± 0.62	1.59 ± 0.33	1.36 ± 0.28	<**0.0001**
Bilirubin (μmol/L)	104.99 (IQR 59.00-198.51)	95 (IQR 70.96-141.07)	29 (IQR 20.00-43.94)	<**0.0001**
Sodium (mmol/L)	132.66 ± 5.13	137.94 ± 3.00	133.95 ± 3.86	<**0.0001**
Albumin (g/L)	29.64 ± 6.50	27.88 ± 5.11	32.81 ± 5.75	<**0.0001**
Primary outcome[1]	319 (10.9%)	15 (2.5%)	51 (8.6%)	<**0.0001**

p values highlighted with boldface denote statistically significant differences.
RFH-GFR: Royal-Free Hospital Glomerular Filtration Rate.
INR: International normalized ratio.

[1] Mortality or delisting due to clinical deterioration within the first 90 days after inclusion in the waiting list.

Supplementary Table 6: Clinical characteristics of prioritized patients according to the model used. The p values denote comparisons between patients differently prioritized by MELD-Na vs GEMA-AI.

Variable	Transplanted both (n=3,164)	MELD-Na transplanted (n=561)	GEMA-AI transplanted (n=561)	p value
Age	52.58 ± 10.56	48.07 ± 12.76	57.35 ± 9.09	<**0.0001**
Sex (women)	1,075 (34.0%)	168 (29.9%)	270 (48.1%)	<**0.0001**
Ascites (Moderate-severe)	1,619 (51.2%)	125 (22.3%)	307 (54.7%)	<**0.0001**
Urea (mmol/L)	6.40 (IQR 4.50-9.70)	3.70 (IQR 2.90-4.60)	8.80 (IQR 6.80-11.65	<**0.0001**
Creatinine (μmol/L)	94.79 ± 50.05	58.73 ± 15.02	106.08 ± 30.27	<**0.0001**
RFH-GFR (mL/min)	56.01 ± 22.33	91.15 ± 25.64	43.68 ± 11.64	<**0.0001**
INR	1.82 ± 0.61	1.63 ± 0.35	1.32 ± 0.23	<**0.0001**
Bilirubin (μmol/L)	100 (IQR 54.03-195.99)	82.08 (IQR 53.01-126.00)	29 (IQR 19.00-47.51)	<**0.0001**
Sodium (mmol/L)	132.53 ± 4.93	136.35 ± 2.86	135.67 ± 4.39	**0.002**
Albumin (g/L)[1]	29.89 ± 6.50	28.77 ± 5.46	32.05 ± 6.08	<**0.0001**
Primary outcome[2]	365 (11.5%)	18 (3.2%)	47 (8.4%)	<**0.0001**

p values highlighted with boldface denote statistically significant differences.
RFH-GFR: Royal-Free Hospital Glomerular Filtration Rate.
INR: International normalized ratio.

[1] Albumin was not available in 228 patients in this analysis.

[2] Mortality or delisting due to clinical deterioration within the first 90 days after inclusion in the waiting list.

Supplementary Table 7: Results of sensitivity analysis considering all patients as not having ascites when estimating RFH-GFR within the GEMA-AI equation.

	GEMA-AI	GEMA-AI without ascites	
Cohort	Hc	Hc	p value[1]
Training (whole; $n = 5,762$)	0.798 (0.772-0.824)	0.797 (0.771-0.823)	0.2222
Training (women; $n = 1,955$)	0.824 (0.785-0.864)	0.821 (0.781-0.860)	**0.0295**
Internal validation (whole; $n = 1,920$)	0.781 (0.732-0.829)	0.778 (0.729-0.827)	0.2538
Internal validation (women; $n = 623$)	0.826 (0.747-0.905)	0.827 (0.748-0.906)	0.5399
External validation (whole; $n = 1,638$)	0.793 (0.741-0.846)	0.789 (0.736-0.841)	**0.0111**
External validation (women; $n = 432$)	0.836 (0.751-0.921)	0.830 (0.745-0.916)	0.1646

Hc=Harrell's concordance statistic (95% Confidence Interval). Each cohort was analyzed as a whole and also the subgroup of women separately. P values highlighted with boldface denote statistically significant differences.

[1] p values of the discrimination comparison of GEMA-AI vs. GEMA-AI without ascites.

Supplementary Table 8: Data from selected patients in the database who had extreme analytical values and died awaiting liver transplantation or were excluded from the waiting list due to clinical worsening. The prioritization scores according to the different models evaluated are shown.

Variable	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Creatinine (μmol/L)	145	65	400	266	194
Sodium (mmol/L)	142	153	128	139	153
Bilirubin (μmol/L)	845.08	40.00	419.98	10.94	338.07
INR	1.8	1.7	1.1	1.4	2.1
RFH-GFR (mL/min)	24.61	50.15	13.15	16.38	24.80
Ascites (Moderate-severe)	Yes	Yes	No	Yes	Yes
MELD-Na	32	16	35	21	34
MELD 3.0	37	16	34	21	35
GEMA-Na	34	19	34	22	33
GEMA-AI	40	24	38	24	40

INR: International normalized ratio.
RFH-GFR: Royal-Free Hospital Glomerular Filtration Rate.

References

1. Elisa Allen, Rhiannon Taylor, Alexander Gimson, and Douglas Thorburn. Transplant benefit-based offering of deceased donor livers in the United Kingdom. *Journal of Hepatology*, 81(3):471–478, sep 2024. URL https://doi.org/10.1016/j.jhep.2024.03.020.

2. Christophe Duvoux, Françoise Roudot-Thoraval, Thomas Decaens, Fabienne Pessione, Hanaa Badran, Tullio Piardi, Claire Francoz, Philippe Compagnon, Claire Vanlemmens, Jérome Dumortier, Sébastien Dharancy, Jean Gugenheim, Pierre-Henri Bernard, René Adam, Sylvie Radenne, Fabrice Muscari, Filomena Conti, Jean Hardwigsen, Georges-Philippe Pageaux, Olivier Chazouillères, Ephrem Salame, Marie-Noelle Hilleret, Pascal Lebray, Armand Abergel, Marilyne Debette-Gratien, Michael D Kluger, Ariane Mallat, Daniel Azoulay, Daniel Cherqui, and Liver Transplantation French Study Group. Liver Transplantation for Hepatocellular Carcinoma: A Model Including α-Fetoprotein Improves the Performance of Milan Criteria. *Gastroenterology*, 143(4):986–994.e3, oct 2012. URL https://doi.org/10.1053/j.gastro.2012.05.052.

3. Francis Y. Yao, Linda Ferrell, Nathan M. Bass, Jessica J. Watson, Peter Bacchetti, Alan Venook, Nancy L. Ascher, and John P. Roberts. Liver Transplantation for Hepatocellular Carcinoma: Expansion of the Tumor Size Limits Does Not Adversely Impact Survival. *Hepatology*, 33(6):1394–1403, jun 2001. URL https://doi.org/10.1053/jhep.2001.24563.

4. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, Jun. 2020. ISSN 1566-2535. URL https://doi.org/10.1016/j.inffus.2019.12.012.

5. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. URL https://doi.org/10.1038/s42256-019-0048-x.

6. Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, 1996. URL https://dl.acm.org/doi/10.5555/235248.

7. R. P. Lippmann. Pattern classification using neural networks. *IEEE Communications Magazine*, 27(11): 47–50, Nov 1989. ISSN 1558-1896. URL https://doi.org/10.1109/35.41401.

8. Xin Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, Sep. 1999. ISSN 1558-2256. URL https://doi.org/10.1109/5.784219.

9. Geoffrey E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1):185–234, Sep. 1989. ISSN 0004-3702. URL https://doi.org/10.1016/0004-3702(89)90049-0.

10. Kenneth O. Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1):24–35, Jan. 2019. ISSN 2522-5839. URL https://doi.org/10.1038/s42256-018-0006-z.

11. Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution Strategies as a Scalable Alternative to Reinforcement Learning, 2017. Preprint at https://arxiv.org/abs/1703.03864.

12. X. Yao and Y. Liu. A new evolutionary system for evolving artificial neural networks. *IEEE Transactions on Neural Networks*, 8(3):694–713, May 1997. URL https://doi.org/10.1109/72.572107.

13. P.J. Angeline, G.M. Saunders, and J.B. Pollack. An evolutionary algorithm that constructs recurrent neural networks. *IEEE Transactions on Neural Networks*, 5(1):54–65, Jan. 1994. ISSN 1045-9227. URL https://doi.org/10.1109/72.265960.

14. Christian Igel and Michael Hüsken. Empirical evaluation of the improved rprop learning algorithms. *Neurocomputing*, 50:105–123, Jan. 2003. ISSN 0925-2312. URL https://doi.org/10.1016/S0925-2312(01)00700-7.

15. Juan Carlos Fernandez Caballero, Francisco José Martinez, César Hervas, and Pedro Antonio Gutierrez. Sensitivity Versus Accuracy in Multiclass Problems Using Memetic Pareto Evolutionary Neural Networks. *IEEE Transactions on Neural Networks*, 21(5):750–770, May 2010. ISSN 1045-9227. URL https://doi.org/10.1109/TNN.2010.2041468.

16. S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598): 671–680, May 1983. ISSN 0036-8075. URL https://doi.org/10.1126/science.220.4598.671.

17. Manuel Luis Rodríguez-Perálvarez, Antonio Manuel Gómez-Orellana, Avik Majumdar, Michael Bailey, Geoffrey W McCaughan, Paul Gow, Marta Guerrero, Rhiannon Taylor, David Guijo-Rubio, César Hervás-Martínez, and Emmanuel A Tsochatzis. Development and validation of the Gender-Equity Model for liver Allocation (GEMA) to prioritise candidates for liver transplantation: a cohort study. *The Lancet Gastroenterology & Hepatology*, 8(3):242–252, Mar. 2023. ISSN 2468-1253. URL https://doi.org/10.1016/S2468-1253(22)00354-5.

Analytical parameters

Creatinine*	100	µmol/L
Bilirubin*	120	µmol/L
INR*	1.8	
Sodium*	140	mmol/L
Urea*	20	mmol/L
Gender*	Female	
Age*	56.00	Years
Ascites* (Moderate or Severe)	Yes	
RFH-GFR**	30.71	mL/min

** Values required for calculations.*
*** The RFH-GFR model is automatically calculated.*

GEMA-AI predictors value after applying thresholds

INR	Bilirubin (µmol/L)	Sodium (mmol/L)	RFH-GFR (mL/min)
1.8	120	140	30.71

	GEMA-AI	
	27	