*JAMIA Open*, 2025, **8(2)**, ooaf020
https://doi.org/10.1093/jamiaopen/ooaf020
**Research and Applications**

Research and Applications

# Computer-assisted prescription of erythropoiesis-stimulating agents in patients undergoing maintenance hemodialysis: a randomized control trial for artificial intelligence model selection

**Lee-Moay Lim, MD[1,2], Ming-Yen Lin, PhD[1], Chan Hsu, MS[3], Chantung Ku, MS[3], Yi-Pei Chen, MS[1], Yihuang Kang, PhD[3,*], Yi-Wen Chiu, MD[1,2,4,*]**

[1]Division of Nephrology, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung 80708, Taiwan, [2]Faculty of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 80708, Taiwan, [3]Department of Information Management, National Sun Yat-sen University, Kaohsiung 80708, Taiwan, [4]The Master Program of AI Application in Health Industry, Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Kaohsiung 80424, Taiwan

*Corresponding authors: Yi-Wen Chiu, MD, 100 Tzyou First Road, Sanmin District, Kaohsiung City 80708, Taiwan (chiuyiwen@kmu.edu.tw, chiuyiwen@gmail.com) and Yihuang Kang, PhD, 70 Lienhai Road, Kaohsiung 80424, Taiwan (ykang@mis.nsysu.edu.tw, yihungkang@gmail.com)

Drs Y. Kang and Y.-W. Chiu contributed equally.

## Abstract

**Objective:** Machine learning (ML) algorithms are promising tools for managing anemia in hemodialysis (HD) patients. However, their efficacy in predicting erythropoiesis-stimulating agents (ESAs) doses remains uncertain. This study aimed to evaluate the effectiveness of a contemporary artificial intelligence (AI) model in prescribing ESA doses compared to physicians for HD patients.

**Materials and Methods:** This double-blinded control trial randomized participants into traditional doctor (Dr) and AI groups. In the Dr group, doses of ESA were determined by following clinical guideline recommendations, while in the AI group, they were predicted by the developed models named Random effects (REEM) trees, Mixed-effect random forest (MERF), Long short-term memory (LSTM) networks-I, and LSTM-II. The primary outcome was the capability to maintain patients' hemoglobin (Hb) value near 11 g/dL with a margin of 0.25 g/dL after treating the suggested ESA, with the secondary outcome being Hb value between 10 and 12 g/dL.

**Results:** A total of 124 participants were enrolled, with 104 completing the study. The mean Hb values were 10.8 and 10.9 g/dL in the AI and Dr groups, respectively, with 69.7% and 73.5% of participants in the respective groups maintaining Hb levels between 10 and 12 g/dL. Only the REEM trees model passed the non-inferiority test for the primary outcome with a margin of 0.25 g/dL and the secondary outcome with a margin of 15%. There was no difference in severe adverse events between the 2 groups.

**Conclusion:** The REEM trees AI model demonstrated non-inferiority to physicians in prescribing ESA doses for HD patients, maintaining Hb levels within the therapeutic target.

**ClinicalTrials.gov Identifier:** NCT04185519.

## Lay Summary

This study evaluated the effectiveness of Machine learning (ML) models in predicting erythropoiesis-stimulating agent (ESA) doses for anemia management in hemodialysis (HD) patients, compared to doctors' prescriptions. In this double-blinded randomized control trial, HD patients were assigned to either a traditional doctor-led (Dr) group using clinical guidelines or an artificial intelligence (AI) group utilizing ML models, including REEM trees, MERF, LSTM-I, and LSTM-II. The main goal was to keep patients' hemoglobin (Hb) levels close to 11 g/dL, with a secondary goal of keeping levels within the range of 10-12 g/dL. On average, Hb levels were 10.8 g/dL in the AI group and 10.9 g/dL in the doctor-led group. Around 70% of the AI group and 74% of the doctor group stayed within the 10-12 g/dL range. REEM trees model was the only AI model that achieved both main and secondary goals as well as doctors. There were no major differences in serious side effects between the 2 groups. In these results, the REEM trees model successfully matches doctors' prescriptions for ESA doses, ensuring Hb levels remain within the desired range, and proving its ability to manage anemia in HD patients.

**Key words:** artificial intelligence (AI); ESA prescription; hemodialysis; anemia; model selection.

## Background and significance

Erythropoiesis-stimulating agents (ESAs) and intravenous iron supplements are fundamental in managing anemia in end-stage kidney disease (ESKD) patients undergoing hemodialysis (HD).[1–3] Over the past decades, several randomized trials and meta-analyses have refined the target hemoglobin

**Received:** July 17, 2024; **Revised:** February 10, 2025; **Editorial Decision:** March 3, 2025; **Accepted:** March 5, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

(Hb) levels (10-12 g/dL) as normalization of Hb in ESKD patients has been associated with an increased risk of thromboembolic events.[1,4–8] Patients with significant Hb level variations are more prone to complications.[9] Thus, maintaining Hb within the target range is a key goal in ESKD patient care.

For optimized anemia management, the National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF-K/DOQI) developed and published guidelines, employing protocols or algorithms for ESA prescription.[10] The NKF-K/DOQI guidelines recommended implementing ESA therapy for Chronic Kidney Disease (CKD) patients with Hb levels lower than 11 g/dL to reduce symptoms and improve quality of life.[10] A target Hb of 11-12 g/dL was recommended while a concentration above 12 g/dL was not advised due to potential cardiovascular risks.[10] Monitoring Epoetin responses requires monitoring Hgb/Hct every 1-2 weeks, until stabilizing Epoetin doses and Hgb/Hct levels are achieved.[10] Most HD facilities use non-validated algorithmic anemia management protocols based on these guidelines.[11]

Machine learning (ML) and artificial intelligence (AI) algorithms have been increasingly employed in nephrology, including anemia management in HD patients.[12–15] Subsequently, various individualized algorithms and prediction models have been developed and tested by data to recommend suitable ESA doses in HD patients.[12,13,16–19] Among these, the artificial neural network (ANN) model has recently gained popularity for ESA dose-response prediction.[17,20,21] A previous study by Barbieri et al. and colleagues demonstrated that the ANN model with 2 layers of 10 neurons each was a reliable tool for predicting long-term outcomes in HD patients receiving ESA/Iron therapy.[20] Model inputs included the last 90 days of patient's medical history and the subsequent 90 days of darbepoetin/iron prescriptions.[20] A comparison between predicted and observed Hb concentrations during training and testing was conducted to evaluate the accuracy of model prediction.[20]

Anemia control assisted by ML/AI technologies shows promise, but the evidence base is limited, with only a few randomized clinical trials testing their effectiveness.[11,12,22] Many reported ML/AI models for individualized ESA dose recommendations do not consider the correlated nature of patient data. Specifically, a patient's records of ESA dose-response information form a sequence of events/outcomes during anemia treatment, which should be treated as longitudinal data that violates the assumptions of these ML/AI models.[23–25] Over the past decades, there has been significant progress in anemia control-assisted modeling, with collaboration between ML and medical research.[19,20,26–28] Therefore, the main goal of this study is to evaluate the effectiveness of 4 contemporary models in predicting ESA dose-response in HD patients, compared with physicians' prescriptions through a randomized control trial (RCT).

## Materials and methods

### Study participants and randomization

This double-blind, parallel RCT was conducted at a tertiary hospital dialysis clinic in Taiwan from July 2019 to July 2020 (ClinicalTrials.gov Identifier: NCT04185519). Approved by the Institutional Review Board of Kaohsiung Medical University Hospital (KMUHIRB-F(I)-20190094), the study adhered to the Declaration of Helsinki and CONSORT reporting guidelines. The trial enrolled ESKD patients over 20 years old undergoing regular HD (4 hours per session, 3 times per week). Inclusion criteria required a minimum of 6 months of laboratory data with at least 1 Hb level within 10-12 g/dL and receipt of at least 1 ESA prescription to maintain Hb levels in this range in the past 6 months, with consistent use of the same ESA brand for at least 6 months prior to enrollment. Exclusion criteria included recent blood transfusion, active bleeding, active infection or malignancy, or inability to adhere to the study protocol. Participants could withdraw if they received different ESA treatments, had major bleeding surgery, or underwent transfusion, chemotherapy, radiotherapy, or immunosuppressive therapy.

### Predictive models used in this study

Four AI models were developed for this study, including the bagged Regression trees with random effects (REEM) trees model,[29–31] Mixed-effect random forest (MERF),[32] Long short-term memory (LSTM) networks LSTM-I, and LSTM-II.[33,34] To better capture the relationship between ESA dosage and changes in Hb levels in different patients, we chose models that incorporate random effects because our dataset includes multiple records for individual patients. The random effects in the study quantify the difference between each subject's Hb value and the average among study subjects and the deviation of the average change in Hb value from the overall average change. Both the bagged REEM tree model and the MERF estimate random effects using the expectation-maximization algorithm. The expectation-maximization algorithm assumes some information in data has not been obtained but could be assumed the existence to obtain the model parameters from collected data. In the study, some information, such as the patient's nutrition, inflammatory status, and potential blood loss, were not usually collected in clinical practice, which may affect the Hb responses after treating ESAs.

The LSTM models comprised 2 main branches: a fixed-effect branch for general patterns and a random-effect branch for patient-specific dose-response patterns. The fixed-effect branch, composed of 2 dense layers, processed general information such as demographics and comorbidities. The random-effect branch, on the other hand, first extracted sequential patterns from lab tests and medications using an LSTM layer and a dense layer. The resulting sequence representation was then combined with an embedding vector of patient identity from a fully connected layer to create a comprehensive patient-specific representation. The final outputs from both branches were concatenated and fed into a final dense layer to predict changes in normalized Hb.

We collected laboratory data from individuals on regular HD who had previously used ESAs to maintain Hb levels between 10 and12 g/dL from January 1, 2015, to June 30, 2019. Using this data, along with ESA and iron supplement doses, Hb levels, demographic, and biochemical data from 305 HD patients (22 649 records with 67 features), we trained our models. (Tables S1 and S2) After applying exclusion criteria, 17 553 records from 263 patients (dated 2015/01/01 to 2019/06/30) were used, with holding out the final 6 months as a validation set ($N = 2431$) and earlier data ($N = 15\ 112$) used for training to avoid temporal leakage issues.[35] The Hb levels of the most records in the training set ($N = 16\ 678$, 71.8%) and the validation set ($N = 1903$, 71.4%) are between 10 and 12. There are 239 and 212 patients in the training and validation set, respectively, with an intersection of 188 patients. Employing ensemble ML techniques such as bootstrap aggregating, we predicted Hb levels based on administered ESA doses, achieving robust predictive accuracy with a mean absolute error (MAE) under 0.5 g/dL, validating their effectiveness in clinical settings.[26,36]

To accommodate the sequential nature of medical data, we employed distinct preprocessing strategies for REEM trees/MERF and LSTM models. For REEM trees and MERF, which are inherently less suited for handling sequential data, we augmented the feature set by incorporating information from the 3 preceding patient visits. This contextual enrichment aimed to mitigate the limitations of these models in capturing temporal dependencies.

In contrast, for LSTM models, we applied standard preprocessing techniques: normalization of continuous features to a 0-1 range and one-hot encoding of categorical features to ensure optimal performance. Patient data was then organized into sequences, with a maximum length of 12 records, based on patient ID.

To address missing values, we utilized a combined approach: "Last observation carried forward (LOCF)" for general missingness and "Next observation carried backward (NOCB)" for missing values in the initial records of each patient, which often arise due to varying examination intervals.

The 4 models are developed using REEM tree (version 0.90.4), ranger (version 0.14.1), tensorflow (version 2.11.0), and keras (version 2.11.0) in an environment with an Intel(R) Xeon(R) CPU E5-2650 and an NVIDIA GeForce GTX1080 Ti GPU.

## Study protocol

Our study protocol is shown in Figure 1A. After a 3-month screening period, all eligible subjects who signed informed consent were randomized into the control arm (Dr group) and the intervention group (AI group). To ensure an equal balance in sample size, we divided the patient pool into 22 blocks, with 6 subjects in each block (3 assigned for the AI group and 3 to the Dr group by randomization). Using SAS, we applied the random seed (6457149) to generate sequential group name assignments. The study period was 6 months, with 2 tests conducted each month.

The study utilized 4 AI models to evaluate ESA dose-response in HD patients. The models employed, based on increasing complexity, were bagged REEM trees (tests 1-3), MERF (tests 4-6), LSTM-I (tests 7-9), and LSTM-II (tests 10-12). Two ESA brands with equivalent potency, Recormon® (5000 IU Epoetin beta) and Nesp® (20 mcg darbepoetin alfa) were administered, represented by syringes with a maximum of 2 per week, adhering to local regulations.[37] Intravenous iron supplements were given to maintain ferritin levels between 200 and 500 ng/mL and transferrin saturation (TSAT) between 20% and 30%, monitored quarterly.

The primary outcome was to assess whether AI is not inferior to physicians in prescribing ESA doses to maintain Hb near 11 g/dL, with a margin of 0.25 g/dL. The secondary outcome aimed to determine the non-inferiority of AI in maintaining Hb within the target range of 10-12 g/dL, with a margin of 15%. In addition to Hb levels and ESA doses, all HD-related bio-information data from 6 consecutive months before enrollment were uploaded.

The study protocol was double-blind, ensuring patient safety (Figure 1B). ESA prescriptions from physicians and AI models were reviewed by a second blinded physician. If a prescription risked Hb deviation from 9 to 13 g/dL, it was deemed a failure and replaced. The study would terminate if failures exceeded 5% at any assessment point.

## Statistical analysis

The study conducted 4 sequential 2-armed ESA prescription experiments, with 3 tests for each model on the same subjects. We assumed no carryover effect of ESA dosage between tests. Initially, 62-119 participants were arranged to be included, based on 90% power, 2.5% type 1 error, a mean Hb level difference of 0.8 g/dL, a standard deviation of 0.3, a non-inferiority margin of 25%-33%, and a 90% response rate.

Subject characteristics and outcomes were described as mean ± SD or median (interquartile range) for continuous variables, and percentages for categorical variables. Group distributions and adverse events were analyzed using t-tests, Mann-Whitney tests, Chi-Square, and Fisher's Exact tests as appropriate. Missing data due to withdrawal were handled with LOCF, and outcomes from 3 visits were averaged per AI model and comparator.

The primary outcome was tested with a one-tailed independent samples t-test, assuming a 0.25 g/dL margin. The secondary outcome was tested using the Farrington-Manning test, one-tailed, with a 15% margin. Failed prescriptions were excluded from as-treated analyses but included in intention-to-treat analyses. Holm adjustment was applied to handle raising alfa errors by multiple statistical comparisons. AI was considered non-inferior to physicians in prescribing ESA doses to maintain Hb near 11 g/dL if the adjusted $P$-value was $<0.05$. Data management and statistical operations were performed using SAS (version 9.4) and RStudio (version 1.2.5042).

# Results

A total of 124 participants who signed informed consent were included for randomization, as shown in the participant flow diagram (Figure 2). Of these, 62 were randomized to the AI group and the remaining to the Dr group. During the study, 20 participants (15.5%) withdrew, with 50 and 54 patients completing the study in the AI and Dr groups, respectively. The primary reasons for dropouts in both groups were blood transfusion events ($N = 5$ in each group). In the Dr group, other withdrawal causes included referral to another HD clinic and joining a different clinical trial. In the AI group, 3 participants failed to follow the study protocol, and 2 received ESA of various brands. Figure S2 shows the number of participants and failed ESA prescriptions for each study test. There was neither a point nor an accumulated rate of failed ESA prescriptions higher than 5%.

The summaries of baseline characteristics of participants in the 2 groups demonstrate similarity in clinical and laboratory parameters post-randomization (Table 1). Both groups, averaging around 65 years old, comprised slightly more males and mirrored the disease distribution of the general HD population, with nearly 45% having diabetes mellitus and high rates of cardiovascular and cerebrovascular disorders. Most participants had undergone HD for years, with over 80% utilizing arteriovenous fistula as vascular access. The mean Hb level hovered around 11 g/dL, with over 60% receiving iron supplements. The average ESA dose before the study was

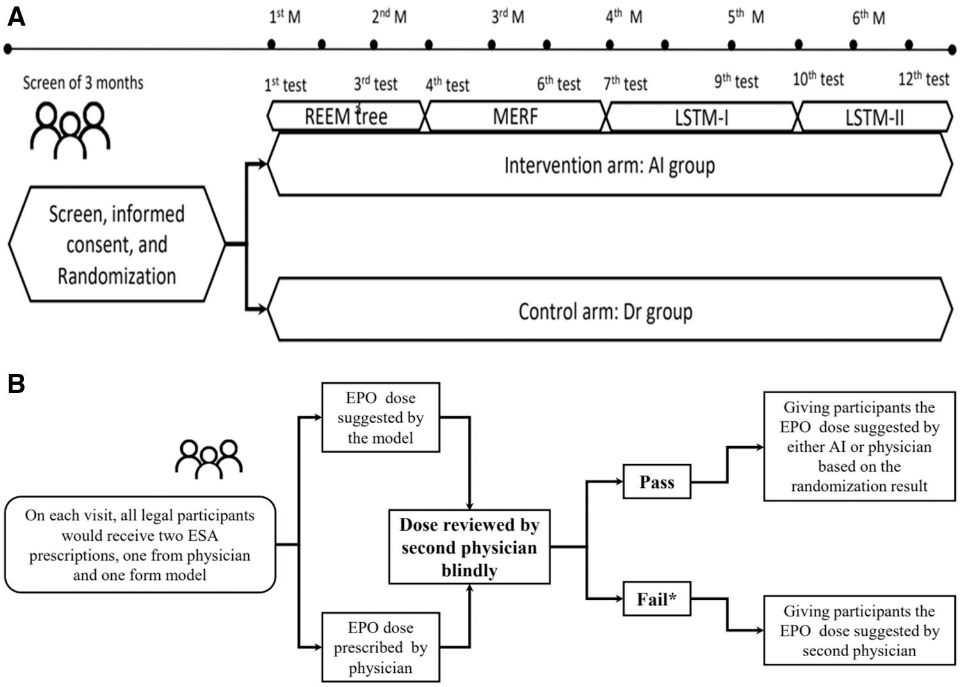

A
1st M
2nd M
3rd M
4th M
5th M
6th M
Screen of 3 months
1st test
3rd test
4th test
6th test
7th test
9th test
10th test
12th test
REEM tree
MERF
LSTM-I
LSTM-II
Intervention arm: AI group
Screen, informed consent, and Randomization
Control arm: Dr group
B
On each visit, all legal participants would receive two ESA prescriptions, one from physician and one form model
EPO dose suggested by the model
EPO dose prescribed by physician
Dose reviewed by second physician blindly
Pass
Fail*
Giving participants the EPO dose suggested by either AI or physician based on the randomization result
Giving participants the EPO dose suggested by second physician

**Figure 1.** (A) Study protocol: visit and assessment schedule. This study conducted 12 tests over 6 months to validate 4 candidate models: the bagged REEM tree model for the first 3 tests, the MERF model for the 4th to 6th tests, the LSTM-I model for the 7th to 9th tests, and the LSTM-II model for the 10th to 12th tests. (B) Study protocol: processes of giving EPO order on each test. Failure was defined as the second physician considering the ESA dose to have the potential to cause the subject's Hb levels to fall outside the range of 9-13 g/dL. The study would be terminated if the rate of failed prescriptions exceeded 5% at any test point or cumulatively.

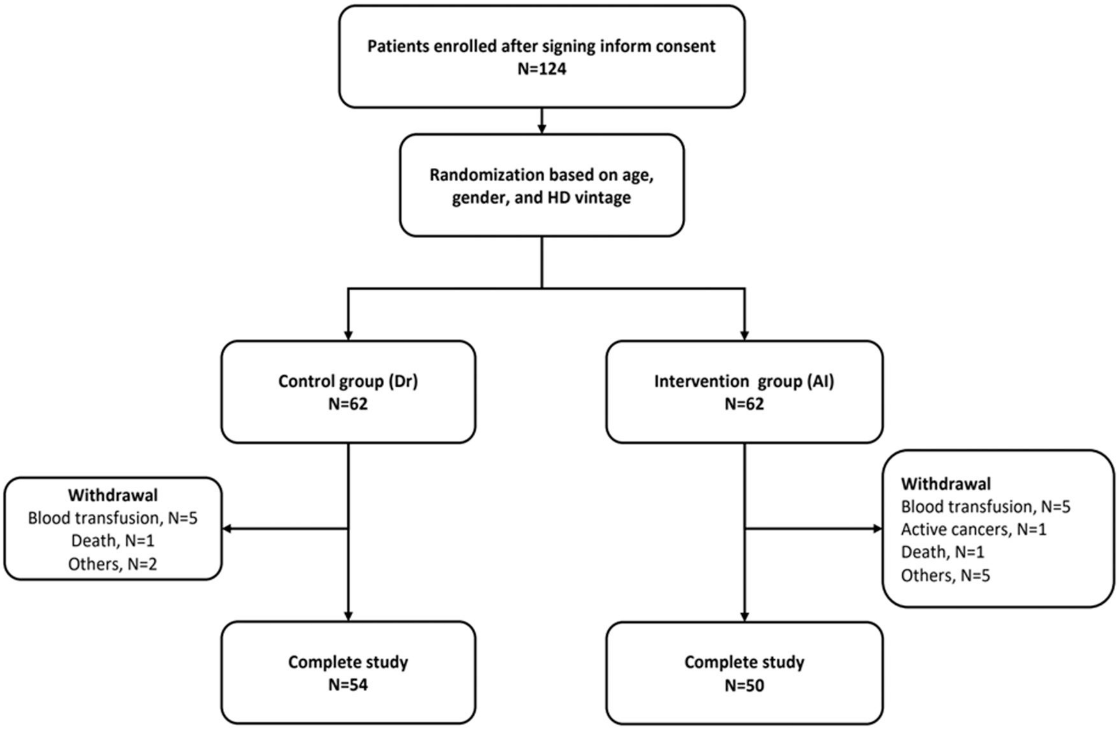


**Figure 2.** Study flow diagram.

comparable between the AI and Dr groups at 1.08 and 1.11 syringes per participant per week, respectively.

The mean Hb levels in both groups during the study are shown in Figure S3. Throughout the 12 tests of the study, the mean Hb levels of the AI and Dr groups were maintained in the range of 10.4-11.1 and 10.6-11.1 g/dL, respectively. In the first 3 tests, the AI group (REEM trees) had mean Hb levels that were not lower than those in the Dr group. However,

**Table 1.** Baseline characteristics of study participants.

| | AI group (*n* = 62) | Dr group (*n* = 62) | *P* value |
|---|---|---|---|
| Age, years | 66.9 (14.8) | 65.8 (14.2) | .69 |
| Gender, female, (%) | 27 (43.5) | 27 (43.5) | 1 |
| Vintage of dialysis, months | 68.5 (115.3) | 62.0 (123) | .52 |
| Causes of dialysis, *n* (%) | | | .42 |
| Diabetes mellitus | 28 (45.1) | 25 (40.3) | |
| Hypertension | 10 (16.1) | 15 (24.2) | |
| Chronic GN | 12 (19.4) | 15 (24.2) | |
| Others | 12 (19.4) | 7 (11.3) | |
| Vascular access, *n* (%) | | | .60 |
| Fistula | 52 (83.9) | 55 (88.7) | |
| Graft | 10 (16.1) | 7 (11.3) | |
| Blood pressure, mmHg | | | |
| Systolic | 144 ± 28 | 149 ± 29.5 | .36 |
| Diastolic | 80 ± 16 | 83 ± 13 | .22 |
| Comorbidity, *n* (%) | | | |
| Cerebrovascular disorder | 11 (17.7%) | 11 (17.7%) | 1 |
| Cardiovascular disorder[a] | 42 (67.7%) | 40 (64.5%) | .85 |
| 2° Hyperparathyroidism | 9 (14.5%) | 10 (16.1%) | 1 |
| HBV or HCV infection status | | | .34 |
| Both negative | 56 (90.3) | 60 (96.8) | |
| HBV positive only | 1 (1.6) | 0 (0) | |
| HCV positive only | 2 (3.2) | 0 (0) | |
| Both positive | 3 (4.8) | 2 (3.2) | |
| Biochemical study on enrollment | | | |
| Hemoglobin, g/dL | 10.8 ± 1.1 | 11.1 ± 0.8 | .10 |
| Hematocrit, % | 33.6 (3.4) | 34.3 (3.6) | .30 |
| MCV, pg/mL | 89.4 (7.0) | 89.7 (8.8) | .28 |
| Ferritin, ng/mL | 229.6 (316.5) | 264.4 (239.4) | .90 |
| Transferrin saturation, % | 29.6 (13.1) | 31.6 (8.5) | .53 |
| Albumin, g/dL | 3.9 ± 0.3 | 4.0 ± 0.4 | .14 |
| Calcium, mg/dL | 4.5 ± 0.4 | 4.6 ± 0.4 | .18 |
| Phosphorus, mg/dL | 4.8 ± 1.2 | 5.0 ± 1.3 | .40 |
| C-reactive protein, mg/dL | 2.7 (5.4) | 2.1 (3.8) | .78 |
| iPTH, pg/mL | 309.7 (339.6) | 332.3 (267.1) | .78 |
| URR, % | 74 (5) | 75 (7) | .93 |
| Iron supplement[b] | | | .35 |
| IV supplement, *n* (%) | 43 (69.4) | 38 (61.3) | |
| Oral supplement, *n* (%) | 0 (0) | 1 (1.6) | |
| None | 19 (30.6) | 23 (37.1) | |
| ESA prescription | | | |
| Nesp, *n* (%) | 31 (50.0) | 37 (59.7) | .37 |
| Mean dose, syringe per week[c] | 1.11 | 1.08 | .80 |

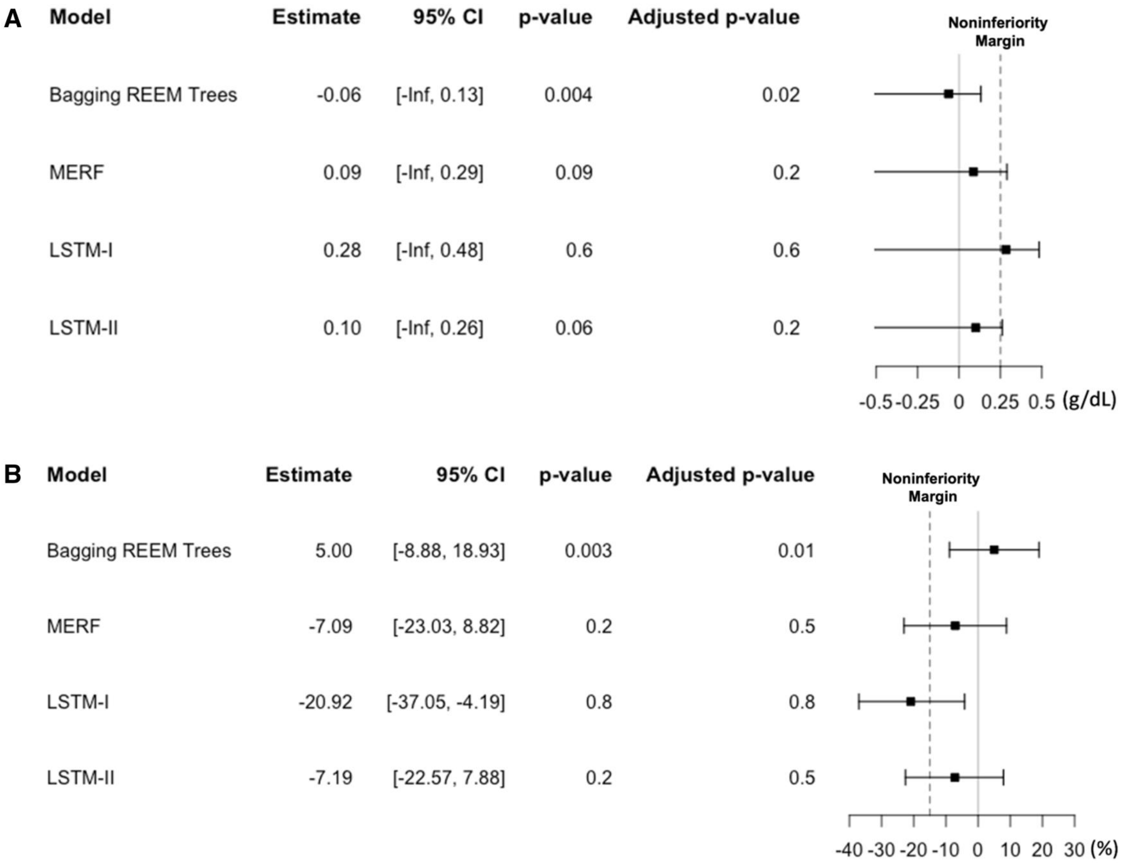
[a] The presence of a history or image study suggesting the diagnosis of cerebrovascular disorders.
[b] Available within 3 months before the study enrollment.
[c] The mean ESA doses of the last 2-week prescriptions before the study enrollment.

Abbreviations: AI = artificial intelligence; GN = glomerular nephritis; HBV = hepatitis B virus; HCV = hepatitis C virus; MCV = mean corpuscular volume; iPTH = intact parathyroid hormone; URR = urea reduction rate; IV = intravenous; ESA = erythropoietin stimulating agent.

the results were reversed in the subsequent tests, except for the last test of the LSTM-II model. Upon further comparison between the AI and Dr groups in each test, there was no significant difference between the 2 groups except for the 8th, 9th, and 10th tests (Figure S3).

Figure 3A and B display the primary and secondary outcomes of the study by the as-treated method. The AI and Dr groups had an absolute difference of Hb between 11 g/dL in the range of 0.7-1.0 and 0.7-0.9 g/dL, respectively, and 2 tests (second and third) in the AI group had a smaller mean difference than the Dr group. Regarding the primary outcome, only the first model passed the non-inferior test with an upper bound of <0.25 g/dL. The AI and Dr groups maintained their Hb between 10 and 12 g/dL in 62.5%-78.0% and 68.3%-81.5%, respectively, and 4 tests (2nd, 4th, 11th, and 12th tests) in the AI group had a greater percentage than the Dr group (Figure S4). When further conducting the group comparison in each test, there was no significant difference between AI and Dr groups over all 12 tests (Figure S4). Regarding the secondary outcome, again, only the first model passed the non-inferior test with an upper bound of <15%. The primary and secondary outcomes were the same if analyzed using the intention-to-treat method.

The AI and Dr groups show some noted characteristics in ESA prescription dose (Table 2). The average total ESA doses for AI and Dr groups were 1.24 and 1.31 syringes per participant per week, respectively. As expected, the AI and Dr groups had the average doses cycled up and down, while AI had a shorter cycle length than the Dr group. The AI group also had a similar average dose range (1.04-1.50 syringe/participant/week) as the Dr group (1.06-1.47 syringe/participant/week). However, considering the average ESA dose distribution, the AI group tended to have their doses located at both extremities while the Dr group was in the middle. The AI group, especially for the LSTM models, easily had their prescriptions of extreme doses, either zero or ceiling



A

| Model | Estimate | 95% CI | p-value | Adjusted p-value |
|---|---|---|---|---|
| Bagging REEM Trees | -0.06 | [-Inf, 0.13] | 0.004 | 0.02 |
| MERF | 0.09 | [-Inf, 0.29] | 0.09 | 0.2 |
| LSTM-I | 0.28 | [-Inf, 0.48] | 0.6 | 0.6 |
| LSTM-II | 0.10 | [-Inf, 0.26] | 0.06 | 0.2 |

B

| Model | Estimate | 95% CI | p-value | Adjusted p-value |
|---|---|---|---|---|
| Bagging REEM Trees | 5.00 | [-8.88, 18.93] | 0.003 | 0.01 |
| MERF | -7.09 | [-23.03, 8.82] | 0.2 | 0.5 |
| LSTM-I | -20.92 | [-37.05, -4.19] | 0.8 | 0.8 |
| LSTM-II | -7.19 | [-22.57, 7.88] | 0.2 | 0.5 |

**Figure 3.** (A) Primary outcome: AI versus Dr on the mean absolute difference between Hb and 11 g/dL by various models. The margin is set at 0.25 g/dL, assuming that the mean absolute difference between Hb and 11 g/dL is 0.75-0.8 g/dL in the dataset, allowing 0.20-0.25 g/dL as the tolerable space to keep Hb between 10 and 12 g/dL. Thus, the margin effect in this study is defined as 0.2/0.8-0.25/0.75, which is 25%-33%; we used the latter in the figure. (B) Secondary outcome: AI versus Dr on the portion of Hb between 10 and 12 g/dL by various models. The vertical dash line indicates the margin effect of the non-inferior test in this study. The margin is set at 15%, allowing the Hb portion between 10 and 12 g/dL to be higher than 60% since the mean value is around 75% in the training dataset.

syringes (4 syringes in 2 weeks and 6 syringes in 3 weeks). Four tests in average ESA dose (4th, 9th, 10th, and 11th) and seven (1st, 2nd, 4th, 9th, 10th, 11th, and 12th) in ESA prescription dose distribution had statistical differences between the AI and Dr group.

During the study period, 16 prescription failures (1.26%, 16/1348) were recorded, as shown in Figure S2, with 11 in the AI group and 5 in the Dr group, without a statistical difference between the groups ($P = .11$). Among the failed prescriptions, there were a total of 5 overdosages and 6 underdosages in the AI group, compared to 2 overdosages and 3 underdosages in the Dr group. Twenty-eight severe adverse events occurred during the clinical study without statistical difference between AI and Dr groups. Hospitalization was the most reported major adverse event, with 7 in the AI and 9 events in the Dr group ($P = .8$). Both AI and Dr groups had 5 blood transfusions and 1 death event ($P = 1$, by Fisher's Exact test).

## Discussion

We present the result of a double-blind, RCT in HD patients to test the performance of 4 AI models (bagged REEM trees, MERF, LSTM-I, and LSTM-II) in predicting ESA-dose response as compared to physicians' prescriptions. We discovered that the model, bagged REEM trees, is not inferior to the physician's decision in prescribing an ESA dose to keep Hb levels near 11 g/dL and maintain between 10 and 12 g/dL. The serious adverse effects (SAEs) between the 2 groups are similar, while the average ESA doses prescribed differ in some tests during the study. Compared with the Dr group, ML/AI models suggest more extreme ESA doses of either zero or maximal syringes.

In the 4 AI models, it is surprising that only the less complex one, bagged REEM trees, has passed the non-inferior test in both primary and secondary outcomes. All 4 models in this study passed the validation using MAE <0.5 g/dL as the training and model selection target. We assumed such prediction error was adequate to keep the Hb within the therapeutic range of 10-12 g/dL when aiming at 11 g/dL. It should be noted, however, that 3 of 4 models failed to pass the non-inferiority test, raising questions about the viability of using the common prediction error measurement for model selection, such as MAE or RMSE (root-mean-square error), in this study.[38] A possible explanation is that model improvement might weigh the Hb response by ESA dose equally, whether the Hb level is within 10-12 g/dL. In other words, the learned model might recognize no difference for increasing Hb from 9 to 10 g/dL with Hb from 10 to 11 g/dL when selecting a better model by minimizing the MAE/RMSE during ML/AI model development. We discovered it might be fragile to rely solely on standard ML/AI error measurement to select and

**Table 2.** ESA prescription dose between AI and Dr groups by test.

| Test | 1st[a] | 2nd[a] | 3rd | 4th[a,b] | 5th | 6th | 7th | 8th | 9th[a,b] | 10th[a,b] | 11th[a,b] | 12th[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prescription duration, weeks | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 |
| AI group | | | | | | | | | | | | |
| Participants, number | 57 | 55 | 52 | 52 | 52 | 52 | 52 | 50 | 50 | 50 | 50 | 50 |
| ESA dose, syringe | | | | | | | | | | | | |
| Total | 161 | 191 | 127 | 119 | 148 | 157 | 112 | 170 | 149 | 156 | 156 | 117 |
| Average[c] | 1.30 | 1.04 | 1.11 | 1.06 | 1.35 | 1.43 | 1.04 | 1.07 | 1.41 | 1.50 | 1.44 | 1.12 |
| Number of prescriptions by syringe, *n* (%) | | | | | | | | | | | | |
| 0 | 14 (23) | 13 (22) | 14 (25) | 20 (37) | 14 (26) | 10 (19) | 22 (42) | 16 (31) | 13 (25) | 9 (18) | 10 (20) | 21 (42) |
| 1 | 1 (2) | 7 (12) | 6 (11) | 3 (6) | 0 (0) | 4 (8) | 2 (4) | 4 (8) | 0 (0) | 1 (2) | 2 (4) | 0 (0) |
| 2 | 6 (10) | 3 (5) | 8 (15) | 5 (9) | 4 (8) | 1 (2) | 1 (2) | 1 (2) | 2 (4) | 3 (6) | 23 (4) | 0 (0) |
| 3 | 10 (17) | 7 (12) | 5 (9) | 0 (0) | 2 (4) | 3 (6) | 2 (4) | 4 (8) | 1 (2) | 1 (2) | 2 (4) | 2 (4) |
| 4 | 29 (48) | 2 (3) | 22 (40) | 26 (48) | 33 (62) | 35 (66) | 25 (48) | 0 (0) | 35 (69) | 36 (72) | 34 (68) | 27 (54) |
| 5 | — | 16 (27) | — | — | — | — | — | 7 (14) | — | — | — | — |
| 6 | — | 11 (19) | — | — | — | — | — | 19 (37) | — | — | — | — |
| Dr group | | | | | | | | | | | | |
| Participants in the study, *n* | 61 | 61 | 60 | 59 | 57 | 57 | 57 | 57 | 57 | 54 | 54 | 54 |
| ESA dose, syringe | | | | | | | | | | | | |
| Total | 153 | 233 | 168 | 170 | 160 | 164 | 148 | 224 | 138 | 137 | 135 | 115 |
| Average[c] | 1.25 | 1.29 | 1.40 | 1.47 | 1.40 | 1.44 | 1.30 | 1.31 | 1.21 | 1.27 | 1.25 | 1.06 |
| Number of prescriptions by syringe, *n* (%) | | | | | | | | | | | | |
| 0 | 6 (10) | 6 (10) | 8 (13) | 7 (12) | 9 (16) | 6 (11) | 13 (23) | 9 (16) | 11 (19) | 7 (13) | 8 (15) | 13 (24) |
| 1 | 10 (16) | 5 (8) | 3 (5) | 4 (7) | 2 (3) | 7 (12) | 3 (5) | 4 (7) | 5 (9) | 5 (9) | 4 (7) | 7 (13) |
| 2 | 14 (23) | 6 (10) | 10 (17) | 7 (12) | 9 (16) | 7 (12) | 7 (12) | 2 (4) | 12 (21) | 16 (30) | 13 (24) | 9 (17) |
| 3 | 8 (13) | 9 (15) | 11 (18) | 8 (14) | 8 (14) | 5 (9) | 5 (9) | 6 (11) | 7 (12) | 4 (7) | 11 (20) | 10 (19) |
| 4 | 23 (38) | 5 (8) | 28 (47) | 32 (55) | 29 (51) | 32 (56) | 29 (51) | 8 (14) | 22 (39) | 22 (41) | 18 (33) | 15 (28) |
| 5 | — | 5 (8) | — | — | — | — | — | 2 (4) | — | — | — | — |
| 6 | — | 24 (40) | — | — | — | — | — | 26 (46) | — | — | — | — |

[a] $P < .05$ when comparing the ESA dose distribution between AI and Dr group by Chi-Square and Fisher's Exact test.
[b] $P < .05$ when comparing the mean ESA dose between AI and Dr group by Mann-Whitney test.
[c] Mean ESA dose expressed by syringe given per participant per week.
Abbreviations: AI = artificial intelligence; Dr = doctor; ESA = erythropoietin stimulating agent.

improve the models. From the perspective of clinical outcomes, there is a need for a new index or measurement regarding therapeutic goals when testing ML/AI models on decision-making assistance, a crucial consideration for applying ML/AI to clinical research. We propose that this new index will primarily focus on the recommended ESA dosage, reflecting its direct impact on patient outcomes by maintaining Hb levels within the therapeutic range, rather than merely mimicking the prescription pattern.

Unlike the primary and secondary outcomes, the different total and mean ESA doses used between the AI and Dr groups imply that the AI models are not similar to the nephrologists when predicting the Hb response. The very high portion of extreme ESA dose in the AI group, either zero or ceiling, is quite the opposite of physicians' prescription patterns. Though trained with the dataset curated by nephrologists, based on clinical expertise and patient data, these models seem not to follow the guideline recommendations to titrate the ESA dose gradually as physicians did in the Dr group. A key reason why the model tends to recommend extreme ESA doses (either zero or ceiling) is its design, which predicts Hb changes based solely on ESA dosages, aiming to maintain Hb levels near 11 g/dL. However, this structure does not inherently regulate dose adjustments, making it more likely to reinforce extreme dosing patterns, particularly in a dataset where such doses are prevalent due to local prescription regulations. Furthermore, our evaluation metric, MAE, does not fully capture the clinical priority of maintaining Hb within the therapeutic range (10-12 g/dL) and may insufficiently penalize over- or under-dosing. Even with custom loss functions, determining appropriate error weightings and penalties remains challenging. The high prevalence of extreme dosages in our training data may have also influenced the model's ability to demonstrate non-inferiority, as such variability is not adequately accounted for by MAE alone. This issue may become more evident in clinical trials, where the true impact of dosage recommendations on patient outcomes will be directly assessed. In future work, we plan to implement a custom loss function that weights the dosage effect based on Hb response, ensuring the model better aligns with clinical goals while minimizing the risks associated with extreme dosing.

Our findings also highlight another important issue: the interpretability of models. Take the LSTM models as an example, as they have the most skewed ESA dose distribution among our models. ANNs are commonly used for ESA dose-response prediction because of their better simulation quality. However, ANNs and their successor, Deep Learning or Deep Neural Networks, have long been criticized for being "black-box," which means the algorithmic decision-making process (ie, prediction) cannot be easily comprehended without further interpretations.[39–41] In many situations, an ANN model that underperforms in some cases is selected simply for its low prediction error (eg, MAE). Furthermore, without explaining the model, one cannot know how the model arrived at such a prediction result, not to mention the feedback to improve the skewed distribution of ESA prescriptions.

In addition to being an RCT, the strength of our study lies in its emphasis on safety and the testing of multiple ML/AI models. Model safety is always a concern when applying AI to clinical research, especially with those uninterpretable models.[42,43] Unlike the pharmacologic clinical trial to have phases I to III to maintain study safety,[44] most AI studies do not have such regulations to follow, even after the publication of SPIRIT-AI and CONSORT-AI extension.[45,46] Introducing a second physician as a "safety guard" in our study design and blinding to the randomization lowered the potential risk caused by the prescriptions given by either AI or Dr group. Our study found that the AI group tended to taper the ESA dose too much while the Dr group kept the dose relatively high. Additionally, testing 4 models in the same trial enables the study group to collect and compare their performances efficiently, especially using non-inferiority to test the hypothesis.[47]

Unavoidably, there are several limitations to our study. First, the study period is too short to evaluate AI's impact on long-term anemia management in HD patients and possible lag effects of ESA on Hb. However, as a pilot study, we had to balance the model number and test time to select a feasible one for further study. Second, the ESA prescription principles of the Dr group might differ. Although 8 nephrologists were in charge of ESA prescription, the similar portion of Hb between 10 and 12 g/dL in either the Pretest or test period showed the prescription consistency in the Dr group. (Figure S4). Furthermore, we did not incorporate intravenous iron supplementation since the actual prescription may differ. Finally, the study demonstrates that the bagged REEM trees model is not inferior to the physician's decision to prescribe an ESA dose. However, the generalization of the results should be with caution. Model performance may be affected by different setting criteria, heterogeneity of training data, and applications to various real-world patients. More research may be needed to develop models with larger and more diverse databases and apply them to external patient populations.

## Conclusion

Maintaining Hb levels within a narrow target range is essential in the treatment of anemia in ESKD patients. To achieve this target range with minimal variability is time-consuming and challenging. AI integration in healthcare is expected to improve patient outcomes, efficiency, and access to personalized treatments and high-quality care.

Our results show that the REEM tree bagging model is not inferior to a physician's decision in prescribing ESA doses to maintain Hb level within 10-12 g/dL. We successfully demonstrated the non-inferiority of AI over physicians on ESA prescription to maintain Hb levels within the therapeutic range. Nonetheless, only 1 of 4 candidate ML/AI models passed the non-inferiority test, highlighting MAE/RMSE's limitations. In spite of the promise of AI-assisted anemia control, there is still a discrepancy between the technologies and the actual medical practice. A new parameter should be developed instead of relying on prediction error measurements alone.

## Acknowledgments

We acknowledge Bo-Wen Kuo and Fang-Yi Su from the Department of Information Management, National Sun Yat-sen University, for developing the models used in this study. This study also expresses gratitude to Dr Ken-Ning Hsu for his assistance with experimental statistics. We thank all the HD room staff in Kaohsiung Medical University Hospital for helping with the RCT conduction.

## Author contributions

Prof. Yi-Wen Chiu and Prof. Yihuang Kang had full access to all the data in the study. They took responsibility for the data's integrity and the data analysis's accuracy. Concept and design: Yi-Wen Chiu and Yihuang Kang; Acquisition, analysis, or interpretation of data: Lee-Moay Lim, Ming-Yen Lin, Chan Hsu, Chantung Ku, Yi-Pei Chen, Yi-Wen Chiu, and Yihuang Kang; Drafting of the manuscript: Yi-Wen Chiu, Lee-Moay Lim, Chantung Ku, and Yihuang Kang; Critical revision of the manuscript for important intellectual content: Lee-Moay Lim, Ming-Yen Lin, Chan Hsu, Chantung Ku, Yi-Pei Chen, Yi-Wen Chiu, and Yihuang Kang; Statistical analysis: Ming-Yen Lin, Chan Hsu, and Yi-Pei Chen; Obtained funding: Yi-Wen Chiu; Administrative, technical, or material support: Yi-Wen Chiu and Yihuang Kang; and Supervision: Yi-Wen Chiu and Yihuang Kang.

## Supplementary material

Supplementary material is available at *JAMIA Open* online.

## Funding

We thank the Ministry of Health and Welfare, Taiwan, and Kaohsiung Medical University Hospital, Kaohsiung, Taiwan (SI11002), for financial support. The National Health Research Institutes and Kaohsiung Medical University Hospital support M.Y.L. in performing this study through a research grant (grant numbers NHRI-EX113-11208PI and KMUH112-2R20).

## Conflicts of interest

The authors have no conflicts of interest to disclose.

## Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

## References

1. Singh AK, Szczech L, Tang KL, et al.; CHOIR Investigators. Correction of anemia with epoetin alfa in chronic kidney disease. *New Engl J Med*. 2006;355:2085-2098. https://doi.org/10.1056/NEJMoa065485
2. Eschbach JW, Abdulhadi MH, Browne JK, et al. Recombinant human erythropoietin in anemic patients with end-stage renal disease. Results of a phase III multicenter clinical trial. *Ann Intern Med*. 1989;111:992-1000. https://doi.org/10.7326/0003-4819-111-12-992
3. Besarab A, Bolton WK, Browne JK, et al. The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. *New Engl J Med*. 1998;339:584-590. https://doi.org/10.1056/NEJM199808273390903
4. Hanafusa N, Nakai S, Iseki K, Tsubakihara Y. Japanese society for dialysis therapy renal data registry-a window through which we can view the details of Japanese dialysis population. *Kidney Int Suppl (2011)*. 2015;5:15-22. https://doi.org/10.1038/kisup.2015.5
5. Ortiz A, Sanchez-Nino MD, Crespo-Barrio M, et al. The Spanish Society of Nephrology (SENEFRO) commentary to the Spain GBD 2016 report: keeping chronic kidney disease out of sight of health

authorities will only magnify the problem. *Nefrologia (Engl Ed)*. 2019;39:29-34. https://doi.org/10.1016/j.nefro.2018.09.002

6. Pfeffer MA, Burdmann EA, Chen CY, et al.; TREAT Investigators. A trial of darbepoetin alfa in type 2 diabetes and chronic kidney disease. *New Engl J Med*. 2009;361:2019-2032. https://doi.org/10.1056/NEJMoa0907845
7. Drueke TB, Locatelli F, Clyne N, et al.; CREATE Investigators. Normalization of hemoglobin level in patients with chronic kidney disease and anemia. *New Engl J Med*. 2006;355:2071-2084. https://doi.org/10.1056/NEJMoa062276
8. Phrommintikul A, Haas SJ, Elsik M, Krum H. Mortality and target haemoglobin concentrations in anaemic patients with chronic kidney disease treated with erythropoietin: a meta-analysis. *Lancet*. 2007;369:381-388. https://doi.org/10.1016/S0140-6736(07)60194-9
9. Zhao L, Hu C, Cheng J, Zhang P, Jiang H, Chen J. Haemoglobin variability and all-cause mortality in haemodialysis patients: a systematic review and meta-analysis. *Nephrology (Carlton)*. 2019;24:1265-1272. https://doi.org/10.1111/nep.13560
10. IV. NKF-K/DOQI Clinical practice guidelines for anemia of chronic kidney disease: update 2000. *Am J Kidney Dis*. 2001;37(1 Suppl 1):S182-S238. https://doi.org/10.1016/s0272-6386(01)70008-x
11. Brier ME, Gaweda AE, Dailey A, Aronoff GR, Jacobs AA. Randomized trial of model predictive control for improved anemia management. *Clin J Am Soc Nephrol*. 2010;5:814-820. https://doi.org/10.2215/CJN.07181009
12. Gaweda AE, Aronoff GR, Jacobs AA, Rai SN, Brier ME. Individualized anemia management reduces hemoglobin variability in hemodialysis patients. *J Am Soc Nephrol JASN*. 2014;25:159-166. https://doi.org/10.1681/asn.2013010089
13. Gaweda AE, Jacobs AA, Aronoff GR, Brier ME. Individualized anemia management in a dialysis facility—long-term utility as a single-center quality improvement experience. *Clin Nephrol*. 2018;90:276-285. https://doi.org/10.5414/CN109499
14. Gabutti L, Lötscher N, Bianda J, Marone C, Mombelli G, Burnier M. Would artificial neural networks implemented in clinical wards help nephrologists in predicting epoetin responsiveness? *BMC Nephrol*. 2006;7:13. https://doi.org/10.1186/1471-2369-7-13
15. Uehlinger DE, Gotch FA, Sheiner LB. A pharmacodynamic model of erythropoietin therapy for uremic anemia. *Clin Pharmacol Therap*. 1992;51:76-89. https://doi.org/10.1038/clpt.1992.10
16. Rogg S, Fuertinger DH, Volkwein S, Kappel F, Kotanko P. Optimal EPO dosing in hemodialysis patients using a non-linear model predictive control approach. *J Math Biol*. 2019;79:2281-2313. https://doi.org/10.1007/s00285-019-01429-1
17. Barbieri C, Molina M, Ponce P, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int*. 2016;90:422-429. https://doi.org/10.1016/j.kint.2016.03.036
18. Bucalo ML, Barbieri C, Roca S, et al. The anaemia control model: does it help nephrologists in therapeutic decision-making in the management of anaemia? *Nefrologia (Engl Ed)*. 2018;38:491-502. https://doi.org/10.1016/j.nefro.2018.03.004
19. Ohara T, Ikeda H, Sugitani Y, et al. Artificial intelligence supported anemia control system (AISACS) to prevent anemia in maintenance hemodialysis patients. *Int J Med Sci*. 2021;18:1831-1839. https://doi.org/10.7150/ijms.53298
20. Barbieri C, Bolzoni E, Mari F, et al. Performance of a predictive model for long-term hemoglobin response to darbepoetin and iron administration in a large cohort of hemodialysis patients. *PLoS One*. 2016;11:e0148938. https://doi.org/10.1371/journal.pone.0148938
21. Barbieri C, Mari F, Stopper A, et al. A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis. *Comput Biol Med*. 2015;61:56-61. https://doi.org/10.1016/j.compbiomed.2015.03.019
22. Brier ME, Gaweda AE. Artificial intelligence for optimal anemia management in end-stage renal disease. *Kidney Int*. 2016;90:259-261. https://doi.org/10.1016/j.kint.2016.05.018
23. Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. *Annu Rev Clin Psychol*. 2010;6:79-107. https://doi.org/10.1146/annurev.clinpsy.032408.153550
24. Shen Z, Liu J, He Y, et al. Towards Out-Of-Distribution Generalization: A Survey. 2023. arXiv (2108.13624 [cs.LG]). https://doi.org/10.48550/arXiv.2108.13624
25. Song P. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer; 2007.
26. Yun HR, Lee G, Jeon MJ, et al. Erythropoiesis stimulating agent recommendation model using recurrent neural networks for patient with kidney failure with replacement therapy. *Comput Biol Med*. 2021;137:104718. https://doi.org/10.1016/j.compbiomed.2021.104718
27. Brier ME, Gaweda AE, Aronoff GR. Personalized anemia management and precision medicine in ESA and iron pharmacology in end-stage kidney disease. *Semin Nephrol*. 2018;38:410-417. https://doi.org/10.1016/j.semnephrol.2018.05.010
28. Gaweda AE, Jacobs AA, Aronoff GR, Brier ME. Model predictive control of erythropoietin administration in the anemia of ESRD. *Am J Kidney Dis*. 2008;51:71-79. https://doi.org/10.1053/j.ajkd.2007.10.003
29. Sagi O, Rokach L. Ensemble learning: a survey. *WIREs Data Min Knowl Discov*. 2018;8:e1249. https://doi.org/10.1002/widm.1249
30. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. *Front Comput Sci*. 2020;14:241-258. https://doi.org/10.1007/s11704-019-8208-z
31. Breiman L. Bagging predictors. *Mach Learn*. 1996;24:123-140. https://doi.org/10.1007/BF00058655
32. Sela RJ, Simonoff JS. RE-EM trees: a data mining approach for longitudinal and clustered data. *Mach Learn*. 2012;86:169-207. https://doi.org/10.1007/s10994-011-5258-3
33. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444. https://doi.org/10.1038/nature14539
34. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
35. Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Mach Learn*. 2020;109:1997-2028. https://doi.org/10.1007/s10994-020-05910-7
36. Lobo B, Abdel-Rahman E, Brown D, Dunn L, Bowman B. A recurrent neural network approach to predicting hemoglobin trajectories in patients with end-stage renal disease. *Artif Intell Med*. 2020;104:101823. https://doi.org/10.1016/j.artmed.2020.101823
37. Liao S-C, Hung C-C, Lee C-T, et al. Switch from epoetin beta to darbepoetin alfa treatment of anemia in Taiwanese hemodialysis patients: dose equivalence by hemoglobin stratification. *Ther Apher Dial*. 2016;20:400-407. https://doi.org/10.1111/1744-9987.12401
38. Yu CS, Chang SS, Chang TH, et al. A COVID-19 pandemic artificial intelligence-based system with deep learning forecasting and automatic statistical data acquisition: development and implementation study. *J Med Internet Res*. 2021;23:e27806. https://doi.org/10.2196/27806
39. Molnar C. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Independently Published; 2022.
40. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052
41. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models Instead. *Nat Mach Intell*. 2019;1:206-215. https://doi.org/10.1038/s42256-019-0048-x
42. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open*. 2022;5:e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946
43. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large

health system: key elements and early successes. *Front Digit Health*. 2022;4:931439. https://doi.org/10.3389/fdgth.2022.931439

44. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. 1. Introduction to Clinical Trials. In: *Fundamentals of Clinical Trials*, 5 ed. Springer; 2007:4-9.
45. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ*. 2020;370:m3164. https://doi.org/10.1136/bmj.m3164
46. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2:e549-e560. https://doi.org/10.1016/s2589-7500(20)30219-3
47. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308:2594-2604. https://doi.org/10.1001/jama.2012.87802

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

JAMIA Open, 2025, 8, 1–10
https://doi.org/10.1093/jamiaopen/ooaf020
Research and Applications