Development of an ensemble learning algorithm to detect patients at high risk of Chronic Kidney Disease using readily available clinical features

Authors
1.Julian Martinez, 2. Alejandra Perez, 3. Jose Zea, 4. Isabella Llano, 5. Natalia Castaño-Villegas, 6. Diego Caro, 7. Jose Javier Arango

Affiliations
1, 2, 3, 4, 5. Arkangel AI, 6. Astrazeneca Colombia, 7. Universidad de Caldas, Quindio, Colombia

Background

Chronic kidney disease (CKD) is a progressive loss of kidney function that affects over 10% of the adult population. Due to its characteristic late onset of visible symptoms, CKD has high mortality rates (1). CKD diagnosis usually relies on laboratory testing, mainly to estimate glomerular filtration rate (eGFR) (2,3).
Machine learning (ML) algorithms have shown promise in identifying at-risk patients from large volumes of data, facilitating timely intervention, and improving patient outcomes (4). Algorithms for CKD early diagnosis have been developed on datasets including rigurous information including data from urine and blood samples obtaining very promising results (5-8). Nevertheless, these particular models rely upon a battery of specialised laboratory tests, often costly and commonly incomplete or missing in clinical charts, especially in low and middle-income countries where resources are more limited. Models that were developed using simpler clinical and laboratory parameters, easily obtained at the point of care are limited, either by small datasets (6) or by focusing on populations with specific risk factors like Type 2 Diabetes (T2D) (9).

Objective

Considering the limitations of obtaining highly detailed data from all at-risk patients and the need for a simpler screening processes, in this study we propose the development of a machine learning algorithm capable of identifying patients at high risk of CKD using low resource clinical variables. We also propose separate approaches for diabetics and non-diabetics.

Methodology

We used data from 3 different databases, 2 from health institutions of the caribbean region of Colombia and 1 from Peru to build a multicentric dataset. As predictors we used the features depicted in Figure 1, model output was the presence or absence of CKD risk. We excluded diabetic patients without a date of diabetes diagnosis and those without creatinine or eGFR values. For training, risk identification was positive for patients with eGFR < 60 mL/min 1.73 m2 or those already officially diagnosed. We then separated diabetic (T2D) and non diabetic patients (NT2D). This resulted in two datasets 19194 T2D and 169842 NT2D.
We used the Arkangel AI web app to train, evaluate and rank several models using both machine learning and deep learning architectures. We then chose the ones with the best performance and built ensemble models for T2D and NT2D patients. These ensembles take outputs from two models and average them according to modifiable weights to give a final decision. We also calculated the Shapely Additive Predictive (SHAP) values to understand the influence of each of the variables on model decision and ensure clinical coherence.

Discussion

Due to the nature of CKD and the purpose of the algorithms we prioritized sensitivity for better screening and patient identification. We achieved this by exploring ensemble models which was succesful in increasing sensitivity by 10% for T2D patients. Specificity and AUC decreased by 21% and 6% respectively. This underlies the importance of considering CKD prevalence in the population. A population with a low pre-test (i.e. known) prevalence of CKD might not be the best to apply this screening tool. On the contrary, if the baseline prevalence is high or it is a targeted population for high-risk of CKD (diabetic or hypertensive patients, among others) the ensemble model presented here will accurately discriminate 91% of the diseased patients.

Conclusions

Scarcity of information and low quality data threaten the effectiveness of other more sophisticated algorithms. Our models achieved high performance, comparable to that of other complex models using fewer readily available clinical features. It also presents one of the first approximations to CKD risk identification in machine learning using Latin-American data and considering diabetic and non-diabetic populations separately. We present a promising starting point for the continued validation and improvement of a low-cost predictive model for chronic kidney disease.

Results

The best performing model for each population was selected. For T2D it was a Random Forest Classifier while for NT2D it was a deep fully connected neural network. For T2D patients the best results were obtained on an ensemble model using the best T2D and the best NT2D model with weight distribution 2T2D : 1NT2D. For NT2D patients best results were obtained using only the NT2D model. For NT2D patients the most influential variables were age, hypertension and sex while for T2D they were age, BMI and diabetes duration.

Table 1. Performance metrics for model testing.

Metric	Model for Diabetics			Model for non-diabetics
	T2D Model	NT2D Model	Ensemble T2D	NT2D Model
Sensitivity	0.815	0.975	0.910	0.925
Specificity	0.596	0.114	0.390	0.972
Precision	0.729	0.595	0.666	0.932
Accuracy	0.721	0.606	0.687	0.958
F1 Score	0.770	0.739	0.769	0.928
AUC	0.706	0.544	0.650	0.948

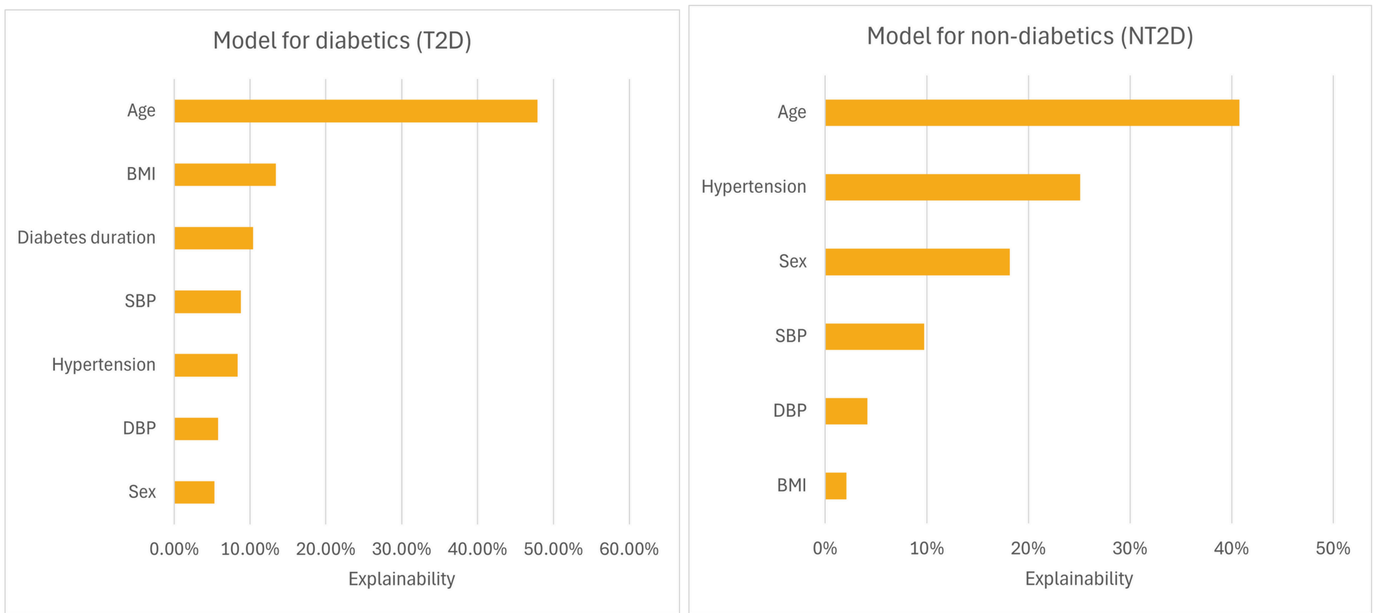

Figure 1. SHAP values for each predictive feature in the diabetic model (left) and non diabetic model (right)

DISCLAIMER: This publication was the product of an unrestricted grant funded by Asrrazeneca

References: (1) Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. The Lancet. 2021 Aug;398(10302):786–802. **(2)** Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl. 2022 Apr 1;12(1):7–11. **(3)** Evans M, Lewis RD, Morgan AR, Whyte MB, Hanif W, Bain SC, et al. A Narrative Review of Chronic Kidney Disease in Clinical Practice: Current Challenges and Future Perspectives. Adv Ther. 2022 Jan 1;39(1):33–43. **(4)** Wu CC, Islam MM, Poly TN, Weng YC. Artificial Intelligence in Kidney Disease: A Comprehensive Study and Directions for Future Research. Diagnostics. 2024 Jan;14(4):397. **(5)** L. Rubini PS. Chronic Kidney Disease [Internet]. UCI Machine Learning Repository; 2015 [cited 2024 May 22]. Available from: https://archive.ics.uci.edu/dataset/336 **(6)** Rashed-Al-Mahfuz Md, Haque A, Azad A, Alyami SA, Quinn JMW, Moni MA. Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening. IEEE J Transl Eng Health Med. 2021;9:1–11. **(7)** Swain D, Mehta U, Bhatt A, Patel H, Patel K, Mehta D, et al. A Robust Chronic Kidney Disease Classifier Using Machine Learning. Electronics. 2023 Jan;12(1):212. **(8)** Chittora P, Chaurasia S, Chakrabarti P, Kumawat G, Chakrabarti T, Leonowicz Z, et al. Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access. 2021;9:17312–34. **(9)** Sammut-Powell C, Sisk R, Budd J, Patel N, Edge M, Cameron R. Development of minimal resource pre-screening tools for chronic kidney disease in people with type 2 diabetes. Future Heal J. 2022 Nov 1;9(3):305–9.