Kidney Diseases

Research Article

Kidney Dis 2026;12:18–28
DOI: 10.1159/000547604

Received: December 20, 2024
Accepted: June 25, 2025
Published online: October 26, 2025

Predicting Chronic Kidney Disease in Type 2 Diabetes Using Natural Language Processing on Healthcare Data

Juan F. Navarro-González[a, b, c, d] Leopoldo Pérez de Isla[e] Gloria Cánovas Molina[f] Miguel Ángel Brito-Sanfiel[g] David Emilio Barajas Galindo[h] Luís Ángel Cuellar Olmedo[i] Dídac Mauricio[j] Santiago Tofé Povedano[k] José Antonio Balsa Barro[l] Matilde Rubio Almanza[m, n] José Juan Aparicio Sánchez[o] Miren Sequera Mutiozabal[o] Belén Pimentel[o] Ana Pérez Domínguez[o] Carlos Arias-Cabrales[p] Víctor Fanjul[p] Antonio Jesús Blanco-Carrasco[q] Juan Francisco Merino Torres[m, n]

[a]Unidad de Investigación y Servicio de Nefrología, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain; [b]RICORS2040, Instituto de Salud Carlos III, Madrid, Spain; [c]Instituto de Tecnologías Biomédicas, Universidad de La Laguna, Tenerife, Spain; [d]Facultad de Ciencias de la Salud, Universidad Fernando Pessoa Canarias, Las Palmas de Gran Canaria, Spain; [e]Servicio de Cardiología, Hospital Clínico San Carlos, Madrid, Spain; [f]Servicio de Endocrinología y Nutrición, Hospital Universitario de Fuenlabrada, Madrid, Spain; [g]Servicio de Endocrinología y Nutrición, Hospital Universitario Puerta de Hierro, Madrid, Spain; [h]Servicio de Endocrinología y Nutrición, Hospital Universitario de León, León, Spain; [i]Servicio de Endocrinología y Nutrición, Hospital Universitario Río Hortega, Valladolid, Spain; [j]Servicio de Endocrinología y Nutrición, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain; [k]Servicio de Endocrinología y Nutrición, Hospital Universitari Son Espases, Mallorca, Spain; [l]Servicio de Endocrinología y Nutrición, Hospital Universitario Infanta Sofía, Madrid, Spain; [m]Servicio de Endocrinología y Nutrición, Departamento de Medicina, Hospital Universitari i Politècnic La Fe, Valencia, Spain; [n]Instituto Investigación Sanitaria La Fe, Universidad de Valencia, Valencia, Spain; [o]Departamento Médico Cardiovascular, Renal y Metabolismo, AstraZeneca España, Madrid, Spain; [p]SAVANA Research Group, S.A., Madrid, Spain; [q]Servicio de Endocrinología y Nutrición, Hospital Clínic de Barcelona, Barcelona, Spain

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

Keywords
Chronic kidney disease · Machine learning · Natural language processing · Predictive model · Type 2 diabetes mellitus · Real-world data · Electronic health records

Antonio Jesús Blanco-Carrasco and Juan Francisco Merino Torres contributed equally to this work.

karger@karger.com
www.karger.com/kdd

© 2025 The Author(s).
Published by S. Karger AG, Basel

This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY) (https://karger.com/Services/OpenAccessLicense). Usage, derivative works and distribution are permitted provided that proper credit is given to the author and the original publisher.

Correspondence to:
Juan F. Navarro-González, jnavgon@gobiernodecanarias.org

Abstract

Introduction: Persons with type 2 diabetes mellitus (T2DM) attending hospitals frequently experience major complications. We assessed the potential use of unstructured free-text data extracted from electronic health records (EHRs) using natural language processing (NLP) and machine learning (ML) to develop a predictive model for chronic kidney disease (CKD) in T2DM. ***Methods:*** This multicenter retrospective study included data from eight Spanish hospitals (2013–2018), extracted using NLP and ML techniques (EHRead®) based on SNOMED CT terminology. From a cohort of individuals with T2DM, we identified those with and without CKD at inclusion. Among individuals without CKD, we trained and validated a 2-year predictive model for CKD development. The model showing the best balance between performance and clinical interpretability was selected for integration into a web-based tool to support early detection and risk stratification. ***Results:*** Of 588,786 individuals with T2DM, 316,597 were included for model development (training: 291,429 [92.1%]; validation: 25,168 [7.9%]; CKD incidence: 15.4% and 18.4%, respectively). A high proportion of missing data was observed in key clinical variables. Among models evaluated, logistic regression achieved the best performance (receiver operating characteristic area under the curve 0.72) using 27 predictors. Both a reduced 10-predictor model and a clinically refined 8-predictor model showed comparable performance to the full model in training and validation cohorts. The clinically refined model was selected for implementation in the web-based tool. ***Conclusion:*** Unstructured EHR data enabled the development of a predictive model for 2-year CKD risk in persons with T2DM. Improving EHR data completeness remains essential to enhance future predictive modeling.

© 2025 The Author(s).
Published by S. Karger AG, Basel

Introduction

Type 2 diabetes mellitus (T2DM) accounts for approximately 90% of all diabetes diagnoses and is a significant global health concern [1, 2]. Among the microvascular complications associated with T2DM, chronic kidney disease (CKD) is particularly prominent, with a prevalence of approximately 50% according to recent data from a Canadian cohort of individuals with T2DM [1, 2]. As the rate of diabetes continues to rise globally, a corresponding increase in CKD associated with T2DM is also projected [3–6]. The development and progression of CKD increase cardiovascular risk, mortality, and the need for medical interventions such as dialysis or transplantation [7]. This results in a higher healthcare burden, including hospitalizations, emergency room visits, and outpatient visits [8–10].

Although several risk factors and aspects of CKD progression can potentially be modified through lifestyle changes and therapeutic interventions, particularly when implemented early [11–13], detecting CKD in its initial stages remains challenging due to its usually asymptomatic nature [14]. Furthermore, conventional biomarkers such as albuminuria and estimated glomerular filtration rate (GFR) are limited at detecting renal histological changes before they become advanced [15, 16]. Given these challenges, electronic health records (EHRs) have become a valuable resource for CKD research as they capture rich longitudinal clinical data including comorbidities, medications, and subtle clinical signs that may precede overt kidney damage [17]. Then, EHRs can help identify early risk patterns and support earlier detection of CKD in routine clinical settings. However, much of the clinical information remains embedded in unstructured free-text narratives [18, 19]. Machine learning (ML) methods, particularly natural language processing (NLP), are essential for transforming free-text clinical data into structured insights, enabling improved risk prediction, guiding clinical decisions, and potentially preventing disease progression [19, 20].

Recent studies have demonstrated the potential of leveraging EHR data for CKD detection, prediction, and management. Systematic reviews highlight that artificial intelligence and ML models trained on EHR data can effectively predict CKD progression, with some achieving high performance in terms of accuracy, sensitivity, and specificity [21]. However, methodological challenges remain, including data quality issues, and missing values [19]. Moreover, while traditional predictive models have largely relied on structured EHR data [22], they struggle with real-world applicability and generalization; emerging evidence suggests that incorporating unstructured text from clinical notes using NLP can enhance detection and prognostic accuracy [17]. Additionally, real-world applications of NLP-based tools have been piloted in hospital settings, showing promise in automating CKD detection and stratifying patient risk profiles [17]. Therefore, there is a growing need for robust and clinically interpretable models that can be seamlessly integrated into routine care [23].

The present work is part of the larger Diabetic@ study which aimed to evaluate the use of unstructured data from EHRs that were extracted and analyzed using NLP and ML to characterize the prevalence and clinical characteristics of DM in the hospital setting [23]. In this

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

particular study, the objective was to assess the potential use of unstructured data from EHRs to train and validate a model to predict CKD development within 2 years in persons with T2DM.

Methods

Study Design and Study Population

This study was conducted within the Diabetic@ study cohort, a multicenter, retrospective, and observational study based on RWD by reusing clinical information captured from the EHRs of persons with DM [23]. The Diabetic@ study included clinical information from January 1, 2013, to December 31, 2018, from eight Spanish tertiary referral hospitals in five regions of the Spanish National Healthcare Network (online suppl. Table 1; for all online suppl. material, see https://doi.org/10.1159/000547604).

For the present analysis, we selected adults with T2DM who attended the participating hospitals during the study period, encompassing both inpatients (hospitalized) and outpatients (ambulatory) individuals. On the inclusion date – defined as the earliest date on which a diagnosis of T2DM was recorded – patients were classified according to the presence of CKD into prevalent CKD (pCKD) and non-pCKD groups. This overall population was used to conduct a descriptive analysis of baseline characteristics and complications during follow-up.

Patients without CKD at inclusion – as CKD was the outcome to be predicted– were used to train and validate a predictive model for CKD development over a 2-year period. Within this subgroup, patients were classified as having incident CKD (iCKD) if they developed CKD during this 2-year period or as non-iCKD if they remained free of CKD throughout the same interval. Patients with less than 2 years of follow-up were excluded only if they did not develop CKD during this period of time. This approach was designed to ensure that all patients classified as "non-events" truly had sufficient follow-up time to confirm the absence of CKD, thereby minimizing the risk of outcome misclassification due to incomplete observation windows.

Study Variables and Data Extraction

Study variables were defined and curated by medical experts and included patient demographics, comorbidities, complications, laboratory parameters, and pharmacological treatments. We conducted a retrospective cross-sectional analysis at the inclusion date and during the follow-up period, respectively – defined as the time from inclusion to the last EHR available within the study period. To reconstruct patient history, all information before the inclusion date was analyzed. For variables analyzed around specific time points, the closest value to the time point within reference time windows was taken. Reference time windows accounted for the variability in healthcare management between patients, specialists, and hospitals, maximizing data retrieval from EHRs. Time window ranges for each variable or group of variables are detailed in the table footnotes.

CKD was defined based on clinical criteria when the following terms or clinical concepts appeared in free text: CKD, nephrotic syndrome, impaired renal function, urine albumin ≥30 mg/24 h, urine albumin-to-creatinine ratio ≥30 mg/g, urine proteins ≥150 mg/dL, creatinine >1.2 mg/dL, GFR <60 mL/min/1.73 m^2. We extracted clinical data recorded as unstructured free-text using the EHRead® technology and organized the data using the SNOMED CT terminology, following previously described methods [24]. This data-driven methodology relies on NLP and ML to generate a synthetic database that contains any detection of medical concepts (and associated metadata) in the study population [25, 26]. EHRead® performance was validated as previously described [27]. Further details on the EHRead® technology, including data source, data processing, extraction and its reading performance for this study are provided in the Supplementary Methods, online supplemental Table 2 as well as in a previous publication of the Diabetic@ Study [23].

Descriptive Analysis

Categorical variables were presented as frequencies, while numerical variables were summarized as mean, standard deviation, median, and first and third quartiles (Q1, Q3). The number of available values was also provided for numeric variables. Missing free-text information from categorical variables was interpreted as no occurrence or absence of the characteristic (i.e., true zero values). For predictive model development, missing data from numerical variables was imputed using random forest multiple imputation, when needed. For complications, cumulative incidence was estimated by the Kaplan-Meier method that accounted for censoring. Laboratory parameters evolution was assessed with generalized additive models. Kaplan-Meier curves and generalized additive model outputs were plotted with 95% confidence intervals (CIs). All the results were stratified by the presence or absence of CKD at the inclusion date. Data analysis was performed using "R" software (version 4.0.2) and Python (version 3.7.12).

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

Predictive Model Training and Validation

The model was trained on data from six hospitals and validated using data from two additional hospitals and focused on predicting the development of iCKD within 2 years after the inclusion date. The division into development and validation datasets was based solely on geographic location, with no clinical or demographic variables influencing the assignment. This strategy was employed to evaluate model generalization for independent populations.

Forty-two potential predictors, including demographics, comorbidities, treatments, and laboratory parameters, were identified by medical experts based on clinical criteria and a literature review (online suppl. Table 3). Variables with over 50% missing data were excluded. This threshold was chosen to ensure a minimum level of data completeness that would support reliable estimation and avoid excessive reliance on imputation.

Four ML models (decision tree, random forest, extreme gradient boosting [XGBoost], and logistic regression [LR]) were trained using all available predictors, with default hyperparameters and a fixed random seed. Class imbalance was addressed through class-weight regularization. Model performance was assessed using stratified 10-fold cross-validation.

To obtain models with a reduced set of predictors (i.e., reduced models), feature selection was performed using the maximum relevance minimum redundancy method [28]. These were then reviewed and refined by a clinical panel from the Spanish scientific societies of endocrinology and nephrology. As part of this process, a series of workshops were held with physicians involved in the management of CKD, during which the models were evaluated, adjusted, and finalized based on clinical judgment and consensus. The resulting models – clinically interpretable and optimized for practical use – were subsequently validated on the remaining two hospitals using the area under the curve (AUC) receiver operating characteristic (ROC) and precision-recall curves, with CIs estimated through bootstrapping. To support clinical decision-making, we developed a user-friendly web application (Predict 2 Prevent, P2P) that allows researchers to input patient characteristics and receive CKD risk predictions based on the model with the best performance and interpretability identified in this study. Additional details about the model training and validation process are provided in the online supplementary Material.

Results

Study Population

A total of 2,580,778 patients attended the participating hospitals during the study period. Of these, 588,756 patients had a T2DM diagnosis, 148,229 (25.2%) of whom were classified as pCKD, and 440,527 (74.8%) as non-pCKD (Fig. 1).

After applying exclusion criteria for the predictive model, 316,597 (53.8%) patients were included in its development. The training set had 291,429 (92.1%) patients, of whom 44,757 (15.4%) had CKD reported during the subsequent 2-year follow-up period (iCKD). The validation set included 25,168 (7.9%) patients, with 4,642 (18.4%) developing CKD and included as iCKD (Fig. 1).

Descriptive Analysis

The baseline characteristics, treatment patterns, follow-up duration, and complication rates for the overall T2DM population (including pCKD and non-pCKD) are described in the supplementary results (online suppl. Table S4 and Fig. 2, 3). The main characteristics and treatments of the predictive population are summarized in Table 1. Notably, the median (Q1, Q3) age in the training set was 67 (53, 79) years for iCKD individuals and 52 (38, 66) years for non-iCKD individuals. Among those who developed iCKD, 50.4% were female. Ischemic heart disease and chronic heart failure were reported in 21.2% and 18.7% of iCKD individuals, and in 11.3% and 8.1% of non-iCKD individuals, respectively. Finally, the use of antihypertensive treatment and statins was more frequent among iCKD individuals (44.7% and 25.7%, respectively). These findings were consistent in the validation set.

Predictive Model Training and Validation

Four "full" ML models were trained using 27 of the original 42 predictors; 15 laboratory parameters with >50% missing values were discarded. Table 2 shows the performance metrics of the ML prediction models including full, reduced and refined models. Among the full models, XGBoost and LR performed best in the cross validation, with mean (95% CI) AUC ROC scores of 0.72 (0.71–0.74) and 0.72 (0.70–0.74), respectively (Table 2). Despite similar performance, the LR model was chosen for better interpretability. The reduced LR model with 10 predictors (potassium-sparing agents, diabetic retinopathy, loop diuretics, chronic heart failure, peripheral vascular disease, smoking, atrial fibrillation, ischemic heart disease, antihypertensives, and age at index)

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

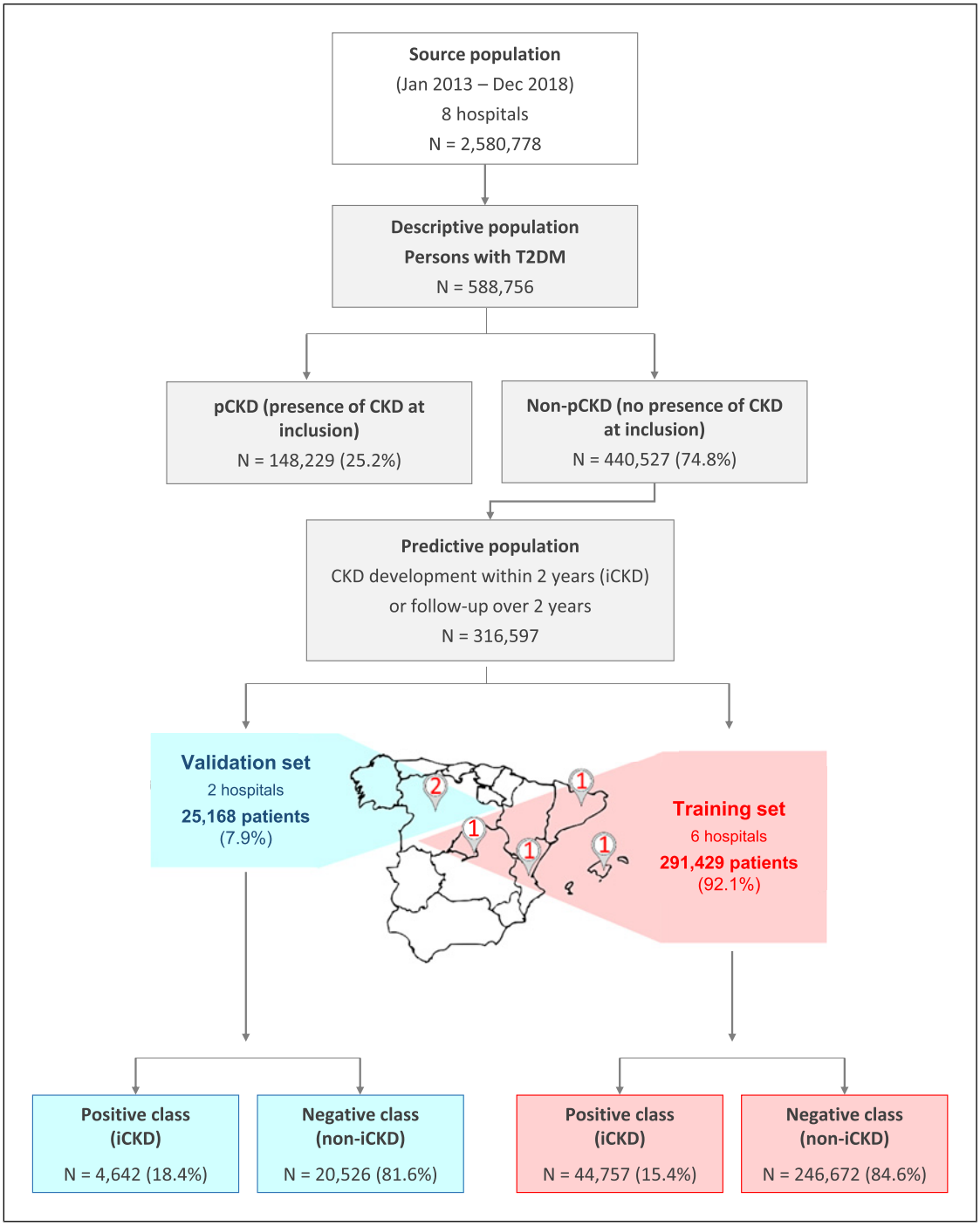

Fig. 1. Patients included in the descriptive analysis (descriptive population, stratified by pCKD and non-pCKD at inclusion) and predictive analysis (predictive population, focused on individuals with T2DM, without CKD at inclusion, excluding those with less than 2 years of follow-up without evidence of developing CKD). The development of CKD in the predictive population was termed incident CKD (iCKD), and those patients who did not develop CKD during the follow-up were designated as non-iCKD. Participating hospitals are indicated by numbers on the map according to their inclusion in either the training (1) or validation (2) set: 1, Hospital Son Espases (Balearic Islands), Hospital Universitari i Politècnic La Fe (Valencia), Hospital Universitario de Fuenlabrada, Hospital Universitario Puerta de Hierro and Hospital Universitario Infanta Sofia (Madrid), and Hospital de la Santa Creu i Sant Pau (Catalonia); 2, Hospital Universitario de León and Hospital Universitario Río Hortega de Valladolid (Castilla y León). *Only patients who developed CKD within 2 years or had a follow-up period longer than 2 years were included. CKD, chronic kidney disease; T2DM, type 2 diabetes mellitus.

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

Table 1. Demographic characteristics, comorbidities, treatments, and clinical parameters of the population used for the development of predictive models

	Training set			Validation set		
	iCKD (*N* = 44,757)	Non-iCKD (*N* = 246,672)	Overall (*N* = 291,429)	iCKD (*N* = 4,642)	Non-iCKD (*N* = 20,526)	Overall (*N* = 25,168)
Demographic characteristics						
Age at index, years, median (Q1, Q3)	67 (53, 79)	52 (38, 66)	55 (40, 69)	74 (63, 82)	64 (52, 75)	66 (54, 77)
Female sex, *n* (%)	22,572 (50.4)	132,964 (53.9)	155,536 (53.4)	2,038 (43.9)	9,831 (47.9)	11,869 (47.2)
Smoking (current/former), *n* (%)	13,082 (29.2)	55,253 (22.4)	68,335 (23.4)	1,137 (24.5)	4,296 (20.9)	5,433 (21.6)
Comorbidities, *n* (%)						
Ischemic heart disease	9,498 (21.2)	27,787 (11.3)	37,285 (12.8)	959 (20.7)	2,676 (13.0)	3,635 (14.4)
Chronic heart failure	8,357 (18.7)	20,014 (8.1)	28,371 (9.7)	972 (20.9)	2,130 (10.4)	3,102 (12.3)
Peripheral vascular disease	6,755 (15.1)	16,215 (6.6)	22,970 (7.9)	707 (15.2)	1,644 (8.0)	2,351 (9.3)
Atrial fibrillation	6,116 (13.7)	13,232 (5.4)	19,348 (6.6)	801 (17.3)	1,805 (8.8)	2,606 (10.4)
Obesity	4,324 (9.7)	13,897 (5.6)	18,221 (6.3)	598 (12.9)	2,014 (9.8)	2,612 (10.4)
Cerebrovascular disease	3,994 (8.9)	10,049 (4.1)	14,043 (4.8)	394 (8.5)	1,118 (5.4)	1,512 (6.0)
Peripheral arterial disease	1,509 (3.4)	3,525 (1.4)	5,034 (1.7)	207 (4.5)	375 (1.8)	582 (2.3)
Diabetic neuropathy	542 (1.2)	2,097 (0.9)	2,639 (0.9)	47 (1.0)	253 (1.2)	300 (1.2)
Diabetic retinopathy	1,164 (2.6)	3,373 (1.4)	4,537 (1.6)	198 (4.3)	353 (1.7)	551 (2.2)
Foot amputation	111 (0.2)	241 (0.1)	352 (0.1)	50 (1.1)	46 (0.2)	96 (0.4)
Pharmacological treatments, *n* (%)						
Any antihypertensive treatment	20,000 (44.7)	57,726 (23.4)	77,726 (26.7)	2,436 (52.5)	6,811 (33.2)	9,247 (36.7)
Statins	11,486 (25.7)	36,205 (14.7)	47,691 (16.4)	1,265 (27.3)	4,233 (20.6)	5,498 (21.8)
ACE inhibitors	8,428 (18.8)	26,044 (10.6)	34,472 (11.8)	925 (19.9)	2,771 (13.5)	3,696 (14.7)
Angiotensin II receptor antagonists	6,653 (14.9)	18,260 (7.4)	24,913 (8.5)	897 (19.3)	2,216 (10.8)	3,113 (12.4)
Potassium-sparing agents	1,940 (4.3)	3,236 (1.3)	5,176 (1.8)	334 (7.2)	480 (2.3)	814 (3.2)
Antiplatelet agents	9,378 (21.0)	26,692 (10.8)	36,070 (12.4)	1,136 (24.5)	3,639 (17.7)	4,775 (19.0)
Beta-blocking agents	7,250 (16.2)	18,521 (7.5)	25,771 (8.8)	982 (21.2)	2,723 (13.3)	3,705 (14.7)
Loop diuretics	5,819 (13)	8,282 (3.4)	14,101 (4.8)	1,014 (21.8)	1,578 (7.7)	2,592 (10.3)
Calcium channel blockers	4,426 (9.9)	10,918 (4.4)	15,344 (5.3)	405 (8.7)	1,001 (4.9)	1,406 (5.6)
Anticoagulants	4,279 (9.6)	8,437 (3.4)	12,716 (4.4)	709 (15.3)	1,350 (6.6)	2,059 (8.2)
Low-ceiling diuretics (thiazides)	1,421 (3.2)	3,524 (1.4)	4,945 (1.7)	70 (1.5)	217 (1.1)	287 (1.1)
Unspecified antihypertensives	1,387 (3.1)	2,883 (1.2)	4,270 (1.5)	125 (2.7)	279 (1.4)	404 (1.6)
Fibrates	1,019 (2.3)	3,488 (1.4)	4,507 (1.5)	128 (2.8)	406 (2.0)	534 (2.1)
Low-ceiling diuretics (other)	716 (1.6)	2,259 (0.9)	2,975 (1.0)	102 (2.2)	302 (1.5)	404 (1.6)
Ezetimibe	580 (1.3)	1,738 (0.7)	2,318 (0.8)	55 (1.2)	203 (1.0)	258 (1)
Bile-acid sequestrants	83 (0.2)	211 (0.1)	294 (0.1)	2 (0.0)	24 (0.1)	26 (0.1)
Clinical parameters						
HbA1c, %						
Available, *n* (%)	2,716 (6.1)	6,308 (2.6)	9,024 (3.1)	875 (18.8)	2,494 (12.2)	3,369 (13.4)
Median (Q1, Q3)	6.3 (5.7–7.4)	6 (5.5–6.9)	6.1 (5.5–7.0)	6.9 (6.1–8.1)	6.6 (5.9–7.8)	6.7 (6.0–7.9)
GFR, mL/min/1.73 m^2						
Available, *n* (%)	15,313 (34.2)	68,330 (27.7)	83,643 (28.7)	2,459 (53)	7,865 (38.3)	10,324 (41)
Median (Q1, Q3)	84.4 (72.2–100.1)	91.8 (79.4–106.9)	90.6 (78–105.8)	73.6 (60.1–90.8)	80.4 (60–100.9)	78.4 (60–98.5)

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

Table 1 (continued)

	Training set			Validation set		
	iCKD (*N* = 44,757)	Non-iCKD (*N* = 246,672)	Overall (*N* = 291,429)	iCKD (*N* = 4,642)	Non-iCKD (*N* = 20,526)	Overall (*N* = 25,168)
Proteins in urine, mg/dL						
Available, *n* (%)	1,549 (3.5)	7,315 (3)	8,864 (3)	259 (5.6)	580 (2.8)	839 (3.3)
Median (Q1, Q3)	0.5 (0.1–75)	0.1 (0.1–0.9)	0.2 (0.1–2.5)	25 (25–75)	25 (25–53.8)	25 (25–75)

For comorbidities and T2DM-related complications, the presence of each feature was analyzed at index considering all prior information in the EHRs. For pharmacological treatments, only information from the last year before index was used. CKD, chronic kidney disease; ACE, angiotensin-converting enzyme; eGFR, estimated glomerular filtration rate; HbA1c, glycated hemoglobin; T2DM, type 2 diabetes mellitus.

Table 2. Performance of the ML prediction models

Model, mean (95% CI)	ROC-AUC	Accuracy	Precision	Recall	F1-score	F2-score
A. Cross-validation metrics of the regularized models						
Full models						
XGB classifier	0.72 (0.71–0.74)	0.65 (0.56–0.70)	0.27 (0.25–0.29)	0.68 (0.61–0.75)	0.38 (0.36–0.40)	0.51 (0.49–0.53)
LR	0.72 (0.70–0.74)	0.67 (0.62–0.72)	0.28 (0.26–0.29)	0.65 (0.59–0.72)	0.38 (0.36–0.40)	0.50 (0.48–0.53)
Random forest classifier	0.57 (0.56–0.58)	0.76 (0.73–0.78)	0.24 (0.23–0.26)	0.28 (0.25–0.30)	0.26 (0.25–0.27)	0.27 (0.25–0.95)
Decision tree classifier	0.53 (0.52–0.53)	0.72 (0.70–0.75)	0.24 (0.23–0.26)	0.36 (0.33–0.39)	0.28 (0.27–0.30)	0.32 (0.30–0.34)
Reduced and refined models						
LR reduced model	0.72 (0.70–0.74)	0.67 (0.61–0.72)	0.27 (0.25–0.29)	0.65 (0.58–0.72)	0.38 (0.36–0.40)	0.50 (0.47–0.53)
LR refined model	0.72 (0.69–0.73)	0.66 (0.60–0.71)	0.27 (0.25–0.29)	0.66 (0.59–0.73)	0.38 (0.36–0.39)	0.50 (0.47–0.53)
B. Validation metrics of the LR models						
Full model	0.70 (0.70–0.71)	0.52 (0.52–0.53)	0.25 (0.25–0.26)	0.80 (0.79–0.81)	0.38 (0.38–0.39)	0.56 (0.55–0.57)
Reduced model	0.70 (0.69–0.71)	0.53 (0.53–0.54)	0.25 (0.25–0.26)	0.79 (0.78–0.80)	0.39 (0.38–0.39)	0.56 (0.55–0.56)
Refined model	0.69 (0.69–0.70)	0.52 (0.51–0.53	0.25 (0.245–0.25)	0.80 (0.79–0.81)	0.38 (0.37–0.38)	0.55 (0.55–0.56)

CI, confidence interval; ROC-AUC, receiver operating characteristic area under the curve; XGB, extreme gradient boosting.

achieved an AUC-ROC of 0.72 (0.70–0.74) (Table 2) with variable importance shown in online supplemental Table 5. Following clinical panel review, the predictors potassium-sparing agents, diabetic retinopathy, and peripheral vascular disease were removed, while cerebrovascular disease and obesity were added. The final LR model included eight predictors shown in Table 3 with their respective variable importances, yielding an AUC-ROC of 0.72 (0.69–0.73) (Table 2 and online suppl. Fig. S1).

The full, reduced, and refined LR models were tested on the validation set, showing a mean (95% CI) AUC

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

Table 3. Variable importance in the refined LR model

Feature	Coefficient	Odds ratio
Obesity	0.408	1.504
Chronic heart failure	0.467	1.596
Smoking (current/former)	0.314	1.369
Atrial fibrillation	0.330	1.390
Ischemic heart disease	0.246	1.279
Antihypertensives	0.302	1.353
Cerebrovascular disease	0.230	1.259
Age at index, years	0.031	1.032

ROC and recall scores of 0.70 (0.70–0.71) and 0.80 (0.79–0.81), 0.70 (0.69–0.71) and 0.79 (0.78–0.80), and 0.69 (0.69–0.70) and 0.80 (0.79–0.81), respectively. All models showed similar accuracy and precision (Table 2). The P2P application was based on the final LR model for risk prediction as it demonstrated similar performance to XGBoost while offering greater interpretability.

Discussion

Comprehensive clinical characterization, risk assessment, and early detection of CKD in individuals with T2DM in the hospital setting are essential for optimizing therapeutic strategies and reducing renal disease progression [29]. In this study, we used NLP and ML to extract unstructured clinical data from EHRs in a large Spanish cohort. These data allowed us to develop and externally validate a data-driven, clinically refined ML model for CKD prediction in this population. However, we observed a significant underreporting of key variables in the EHRs free text, which hindered a complete population description and affected model performance.

The descriptive analysis of both the overall study population and the predictive subpopulation showed that patients with T2DM managed in hospital settings exhibited complex clinical profiles with a high burden of cardiovascular comorbidities, in line with previous studies ([5, 6, 30–32]). Our results showed that approximately one-quarter of individuals with T2DM had prevalent CKD, while 15–18% of those without CKD at baseline developed CKD over a median follow-up of 2 years, consistent with previously reported prevalence and incidence rates [29–32]. Both pCKD and iCKD [29–32] were associated with older age and with more frequent documentation of comorbidities (especially cardiovascular), pharmacological treatments, and laboratory parameters in clinical records. Additionally, cardiovascular complications were ~10% more frequent in the pCKD group. These findings reflect the burden of complications across the cardio-renal-metabolic spectrum in persons with T2DM and support the need for integrated and early preventive strategies [33–35]. In this context, risk stratification through predictive models is essential to enable early detection and management of CKD in T2DM.

Our ML-based predictive model for CKD combined automated selection procedures with clinical insights. This approach allowed us to refine a wide range of predictors described in the literature [36, 37], resulting in an easily interpretable and potentially applicable model with reliable performance. Several predictive models for CKD, based on clinical trials or registries [38–41], have shown AUC-ROC values of 0.68–0.84, similar to our study. However, many focus on 5-year predictions and are limited by population heterogeneity and varied predictor definitions. In contrast, our model focuses on a 2-year prediction horizon, defined by a multidisciplinary clinical panel as an optimal window for early risk detection and clinical actionability. This shorter timeframe enables timely interventions while ensuring sufficient data completeness and minimizing bias related to follow-up variability.

To our knowledge, only one prior study combined NLP and ML for CKD prediction in persons with T2DM, achieving an AUC-ROC of 0.74, but was limited to a single center and lacked independent validation [42]. Our model offers interpretability and has been validated on an independent dataset including widely recognized predictors [4, 38, 43–45]. In this regard, it may support early identification of individuals with T2DM at high risk of developing CKD, even in the absence of structured biomarkers. This risk stratification could help guide clinical decision-making, such as prioritizing patients for further testing (e.g., estimated GFR, albuminuria), intensifying monitoring, or implementing preventive measures (e.g., optimization of blood pressure or glycemic control, use of reno-protective therapies). Additionally, the model's reliance on routinely available unstructured EHR data suggests that it could be integrated into hospital information systems to generate risk alerts, potentially improving long-term outcomes by enabling earlier intervention. To facilitate clinical use, we developed the web-based tool named P2P. Among the candidate models, we selected LR over XGBoost due to their comparable performance and the greater

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

interpretability of LR. We also chose to implement the reduced and clinically refined model as it requires fewer input variables, making it more practical and easier to apply in routine clinical settings.

The present study has several strengths. The main strength of this multicenter study is the analysis of a vast amount of EHRs from eight Spanish hospitals spanning a study period of 6 years. Furthermore, the methodology used to develop and validate the predictive models enabled the creation of a clinically applicable model, further supported by the development of a ready-to-use web-based tool. While our study has its strengths, we also recognize its limitations, mainly tied to data source and extraction challenges. Some important variables to characterize individuals with T2DM are more frequently reported as structured information within the EHRs, rather than in free-text notes, which were the focus of our extraction approach. Consequently, we encountered incomplete documentation of critical variables such as laboratory results and concurrent treatments. This limitation may introduce information bias, hindering a detailed clinical characterization of the study population and potentially affecting the performance and interpretability of our predictive model. In addition, as our study focused on a population composed exclusively of hospitalized individuals with high comorbidity burden and complexity, selection bias may arise and the generalizability of the results to the wider population is likely to be limited. As with any retrospective analysis of RWD, this study may also be subject to confounding bias. Although we applied clinically guided feature selection and adjusted for key covariates, when possible, residual bias cannot be fully ruled out and is acknowledged as a study limitation. Regarding the predictive models, antidiabetic treatments and clinically relevant variables, such as the time since diagnosis of T2DM, were excluded from the list of predictors due to limitations in capturing them accurately in EHR free text. Despite their high recall (sensitivity) – indicating successful identification of most true iCKD cases – they also had low precision. This means that a significant proportion of the positive predictions made by the models could be false positives. Consequently, these models may be best suited to identify persons with T2DM who have a potentially high risk of developing CKD but should not be used as the only tool for therapeutic decision making.

In summary, leveraging unstructured EHR data extracted with NLP enabled us to generate a large real-world database of patients with T2DM managed in the hospital setting. This rich dataset allowed not only a comprehensive characterization of this population – providing valuable insights into their clinical profiles – but also the development of a predictive model for 2-year CKD risk, which integrates ML and expert knowledge and shows high sensitivity for early detection. To facilitate clinical translation, we implemented this model into a web-based tool (P2P), designed to support risk stratification and assist healthcare professionals in guiding early testing, monitoring, and preventive interventions. Furthermore, this approach opens new opportunities to expand our understanding of the cardio-renal-metabolic spectrum and to develop additional predictive tools addressing other relevant outcomes in this high-risk population. Although clinical information from free-text narratives is valuable, there are still some limitations to understanding the whole complexity of the T2DM population-based exclusively on this data source. Improving data recording practices in hospital settings and combining this information with structured data is crucial to maximize NLP utility and ensure a more accurate analysis. Further studies are needed to validate the clinical efficacy of our predictive model and could explore extended time horizons as data completeness allows.

Acknowledgment

Savana Research Group members include Sebastian Menke, Ignacio Salcedo, Natalia Polo, Judith Marín-Corral, David Casadevall, and Miren Taberna.

Statement of Ethics

This study was classified as a non-post-authorization study by the Spanish Agency for Medicines and Health Products (AEMPS) and was approved by the Institutional Review Board of the reference center, Hospital Universitario La Fe in Madrid, Spain (Approval No. CEIm-F-PE-01–14 Ed.01-ACTA 475–12/02/2020). All methods and analyses were performed in compliance with local legal and regulatory requirements, as well as generally accepted research practices as described in the latest version of the Declaration of Helsinki and Good Pharmacoepidemiology Practices. Data were analyzed from anonymized EHRs, which were aggregated in an irreversibly dissociated manner. Therefore, individual patient consent was not required in the study. As this was a retrospective study based on RWD, data collection and adjudication were blinded for all variables. All variables were automatically extracted from EHRs, reflecting routine clinical practice, and therefore, the outcome assessment was based on clinical decisions made by each physician, with no influence from the predictors. Results

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

are reported according to the transparent reporting of a multivariate prediction model for individual prognosis or diagnosis (TRIPOD) [46].

Conflict of Interest Statement

José Juan Aparicio Sánchez, Miren Sequera Mutiozabal, Belén Pimentel, and Ana Pérez Domínguez are AstraZeneca employees. Carlos Arias-Cabrales and Víctor Fanjul are Savana Research S.L employees. Antonio Jesús Blanco-Carrasco, Leopoldo Pérez de Isla, Juan F. Navarro-González, and Juan Francisco Merino Torres are members of the External Scientific Committee of the Diabetic@ study that has been sponsored and funded by AstraZeneca. Leopoldo Pérez de Isla, Antonio Jesús Blanco-Carrasco, and Juan Francisco Merino Torres declare no conflict of interest. Juan F. Navarro-González acts as an external advisor and expert speaker for AbbVie, Amgen, AstraZeneca, Bayer, Boehringer Ingelheim, Esteve, Genzyme, Lilly, MSD, Novartis, NovoNordisk, Servier, Shire, and Vifor Pharma.

Funding Sources

The sponsorship for this study was provided by AstraZeneca. Aid RD21/0005/0013 funded by the Carlos III Health Institute. Recovery, Transformation and Resilience Plan. European Union – NextGenerationEU; and RD24/0004/0022 funded by the Carlos III Health Institute and co-financed by the European Union.

Author Contributions

Inception of the study idea and design of the study protocol: Ana Pérez Domínguez, José Juan Aparicio Sánchez, Miren Sequera Mutiozabal, Antonio Jesús Blanco-Carrasco, Juan Francisco Merino Torres, Leopoldo Pérez de Isla, and Juan F. Navarro-González. Scientific analysis of the study results: José Juan Aparicio Sánchez, Miren Sequera Mutiozabal, Antonio Jesús Blanco-Carrasco, Juan F. Merino Torres, Carlos Arias-Cabrales, Víctor Fanjul, Leopoldo Pérez de Isla, and Juan F. Navarro-González. Scientific and clinical advice on the manuscript content and messages to communicate, and critical reading and reviewing of the manuscript: all authors. Clinical data suppliers: Juan Francisco Merino Torres, Leopoldo Pérez de Isla, Juan F. Navarro-González, Gloria Cánovas Molina, Miguel Ángel Brito-Sanfiel, David Emilio Barajas Galindo, Luís Ángel Cuellar Olmedo, Dídac Mauricio, Santiago Tofé Povedano, José Antonio Balsa Barro, and Matilde Rubio Almanza. Writing of the manuscript: Carlos Arias-Cabrales, Víctor Fanjul, José Juan Aparicio Sánchez, Belén Pimentel, Antonio Jesús Blanco-Carrasco, Juan Francisco Merino Torres, Leopoldo Pérez de Isla, and Juan F. Navarro-González. NLP-ML technique implementation: Savana Research group.

Data Availability Statement

The data underlying this article were provided to the authors by permission of the affiliate institutions. Data could be shared on reasonable request to the corresponding author, after permission of the institutions involved.

References

1 WHO. Diabetes. WHO [World Health Organization)].
2 Standl E, Khunti K, Hansen TB, Schnell O. The global epidemics of diabetes in the 21st century: current situation and perspectives. Eur J Prev Cardiol. 2019;26(2_Suppl l):7–14. https://doi.org/10.1177/2047487319881021
3 Kainz A, Hronsky M, Stel VS, Jager KJ, Geroldinger A, Dunkler D, et al. Prediction of prevalence of chronic kidney disease in diabetic patients in countries of the European union up to 2025. Nephrol Dial Transpl. 2015;30(Suppl 4):iv113–8. https://doi.org/10.1093/ndt/gfv073
4 Wu B, Bell K, Stanford A, Kern DM, Tunceli O, Vupputuri S, et al. Understanding CKD among patients with T2DM: prevalence, temporal trends, and treatment patterns-NHANES 2007-2012. BMJ Open Diab Res Care. 2016;4(1):e000154. https://doi.org/10.1136/bmjdrc-2015-000154
5 Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. Lancet. 2018;392(10159):2052–90. https://doi.org/10.1016/S0140-6736(18)31694-5
6 Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. Diabetologia. 2019;62(1):3–16. https://doi.org/10.1007/s00125-018-4711-2
7 Chu L, Fuller M, Jervis K, Ciaccia A, Abitbol A. Prevalence of chronic kidney disease in type 2 diabetes: the Canadian REgistry of Chronic Kidney Disease In Diabetes Outcomes (CREDO) Study. Clin Ther. 2021;43(9):1558–73. https://doi.org/10.1016/j.clinthera.2021.07.015
8 McQueen RB, Farahbakhshian S, Bell KF, Nair KV, Saseen JJ. Economic burden of comorbid chronic kidney disease and diabetes. J Med Econ. 2017;20(6):585–91. https://doi.org/10.1080/13696998.2017.1288127
9 Usó-Talamantes R, González-de-Julián S, Díaz-Carnicero J, Saurí-Ferrer I, Trillo-Mata JL, Carrasco-Pérez M, et al. Cost of type 2 diabetes patients with chronic kidney disease based on real-world data: an observational population-based study in Spain. Int J Environ Res Public Health. 2021;18(18):9853. https://doi.org/10.3390/ijerph18189853
10 Mullins CD, Pantalone KM, Betts KA, Song J, Wu A, Chen Y, et al. CKD progression and economic burden in individuals with CKD associated with type 2 diabetes. Kidney Med. 2022;4(11):100532. https://doi.org/10.1016/j.xkme.2022.100532
11 Wong MG, Perkovic V, Chalmers J, Woodward M, Li Q, Cooper ME, et al. Long-term benefits of intensive glucose control for preventing end-stage kidney disease: advance-on. Diabetes Care. 2016;39(5):694–700. https://doi.org/10.2337/dc15-2322
12 MacIsaac RJ, Jerums G, Ekinci EI. Effects of glycaemic management on diabetic kidney disease. World J Diabetes. 2017;8(5):172–86. https://doi.org/10.4239/wjd.v8.i5.172
13 Zelniker TA, Wiviott SD, Raz I, Im K, Goodrich EL, Bonaca MP, et al. SGLT2 inhibitors for primary and secondary prevention of cardiovascular and renal outcomes in type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. Lancet. 2019;393(10166):31–9. https://doi.org/10.1016/S0140-6736(18)32590-X

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026

14 Hamza W, Burton JO. Chronic kidney disease awareness and updates on the management of diabetic kidney disease. Pract Diabetes. 2023 2023;40(3):16–20. https://doi.org/10.1002/pdi.2454
15 Lin CH, Chang YC, Chuang LM. Early detection of diabetic kidney disease: present limitations and future perspectives. World J Diabetes. 2016;7(14):290–301. https://doi.org/10.4239/wjd.v7.i14.290
16 Datar M, Ramakrishnan S, Montgomery E, Coca SG, Vassalotti JA, Goss T. A qualitative study documenting unmet needs in the management of diabetic kidney disease (DKD) in the primary care setting. BMC Public Health. 2021;21(1):930. https://doi.org/10.1186/s12889-021-10959-7
17 Farrell D, Chan L. Application of natural language processing in nephrology research. CJASN. 2023;18(6):806–8. https://doi.org/10.2215/CJN.0000000000000118
18 Roberts A. Language, structure, and reuse in the electronic health record. AMA J Ethics. 2017;19(3):281–8. https://doi.org/10.1001/journalofethics.2017.19.3.stas1-1703
19 Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7(2):e12239. https://doi.org/10.2196/12239
20 Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. J Biomed Inform. 2018;88:11–9. https://doi.org/10.1016/j.jbi.2018.10.005
21 Khalid F, Alsadoun L, Khilji F, Mushtaq M, Eze-Odurukwe A, Mushtaq MM, et al. Predicting the progression of chronic kidney disease: a systematic review of artificial intelligence and machine learning approaches. Cureus. 2024;16(5):e60145. https://doi.org/10.7759/cureus.60145
22 Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med. 2019;25(1):57–9. https://doi.org/10.1038/s41591-018-0239-8
23 Merino-Torres J, Blanco-Carrasco A, Rubio Almanza M, Canovas Molina G, Brito-Sanfiel M, Barajas Galindo D, et al. Characterizing the clinical profile and prevalence of people with diabetes attended in the hospital setting by using unstructured healthcare data and natural language processing: the diabetic@ study Diabetes Research and Clinical Practice. 2025. In Press.
24 Izquierdo JL, Morena D, González Y, Paredero JM, Pérez B, Graziani D, et al. Clinical management of COPD in a real-world setting. A big data analysis. Arch Bronconeumol. 2021;57(2):94–100. https://doi.org/10.1016/j.arbres.2019.12.025
25 Benson T. Principles of health interoperability HL7 and SNOMED. London, UK: Springer; 2012.
26 Espinosa-Anke L, Tello J, Pardo A, Medrano I, Ureña A, Salcedo I, et al. Savana: a global information extraction and terminology expansion framework in the medical domain. Proces Leng Nat. 2016;57:23–30.
27 Canales L, Menke S, Marchesseau S, D'Agostino A, Del Rio-Bermudez C, Taberna M, et al. Assessing the performance of clinical natural Language processing systems: development of an evaluation methodology. JMIR Med Inform. 2021;9(7):e20492. https://doi.org/10.2196/20492
28 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. https://doi.org/10.2307/2529310
29 Calvo-Hueros JI, Martín-Hidalgo-Barquero MV, Morales-Gabardino JA, Buitrago F. Chronic kidney disease prevalence and cardiovascular risk in a cohort of patients with type 2 diabetes followed for 10 years in Badajoz (Spain). An observational study. Prim Care Diabetes. 2021;15(2):391–6. https://doi.org/10.1016/j.pcd.2020.11.015
30 Koye DN, Shaw JE, Reid CM, Atkins RC, Reutens AT, Magliano DJ. Incidence of chronic kidney disease among people with diabetes: a systematic review of observational studies. Diabet Med. 2017;34(7):887–901. https://doi.org/10.1111/dme.13324
31 Koye DN, Magliano DJ, Nelson RG, Pavkov ME. The global epidemiology of diabetes and kidney disease. Adv Chronic Kidney Dis. 2018;25(2):121–32. https://doi.org/10.1053/j.ackd.2017.10.011
32 Bramlage P, Lanzinger S, Hess E, Fahrner S, Heyer CHJ, Friebe M, et al. Renal function deterioration in adult patients with type-2 diabetes. BMC Nephrol. 2020;21(1):312. https://doi.org/10.1186/s12882-020-01952-0
33 Cosentino F, Grant PJ, Aboyans V, Bailey CJ, Ceriello A, Delgado V, et al. 2019 ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. Eur Heart J. 2020;41(2):255–323. https://doi.org/10.1093/eurheartj/ehz486
34 Zhang R, Mamza JB, Morris T, Godfrey G, Asselbergs FW, Denaxas S, et al. Lifetime risk of cardiovascular-renal disease in type 2 diabetes: a population-based study in 473,399 individuals. BMC Med. 2022;20(1):63. https://doi.org/10.1186/s12916-022-02234-2
35 Vora J, Cherney D, Kosiborod MN, Spaak J, Kanumilli N, Khunti K, et al. Interrelationships between cardiovascular, renal and metabolic diseases: underlying evidence and implications for integrated interdisciplinary care and management. Diabetes Obes Metab. 2024;26(5):1567–81. https://doi.org/10.1111/dom.15485
36 Rossing P. Risk factors symptoms biomarkers and stages of chronic kidney disease. Compendia. 2021 2021;2021(1):8–12. https://doi.org/10.2337/db20211-8
37 Roy S, Schweiker-Kahn O, Jafry B, Masel-Miller R, Raju RS, O'Neill LMO, et al. Risk factors and comorbidities associated with diabetic kidney disease. J Prim Care Community Health. 2021;12:21501327211048556. https://doi.org/10.1177/21501327211048556
38 Nelson RG, Grams ME, Ballew SH, Sang Y, Azizi F, Chadban SJ, et al. Development of risk prediction equations for incident chronic kidney disease. JAMA. 2019;322(21):2104–14. https://doi.org/10.1001/jama.2019.17379
39 Jiang W, Wang J, Shen X, Lu W, Wang Y, Li W, et al. Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. Diabetes Care. 2020;43(4):925–33. https://doi.org/10.2337/dc19-1897
40 Lin CC, Niu MJ, Li CI, Liu CS, Lin CH, Yang SY, et al. Development and validation of a risk prediction model for chronic kidney disease among individuals with type 2 diabetes. Sci Rep. 2022;12(1):4794. https://doi.org/10.1038/s41598-022-08284-z
41 Kress S, Bramlage P, Holl RW, Möller CD, Mühldorfer S, Reindel J, et al. Validation of a risk prediction model for early chronic kidney disease in patients with type 2 diabetes: data from the German/Austrian Diabetes prospective Follow-up registry. Diabetes Obes Metab. 2023;25(3):776–84. https://doi.org/10.1111/dom.14925
42 Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep. 2019;9(1):11862. https://doi.org/10.1038/s41598-019-48263-5
43 Athyros VG, Doumas M, Imprialos KP, Stavropoulos K, Georgianou E, Katsimardou A, et al. Diabetes and lipid metabolism. Hormones (Basel). 2018;17(1):61–7. https://doi.org/10.1007/s42000-018-0014-8
44 Pugh D, Gallacher PJ, Dhaun N. Management of hypertension in chronic kidney disease. Drugs. 2019;79(4):365–79. https://doi.org/10.1007/s40265-019-1064-1
45 Chen Y, Dabbas W, Gangemi A, Benedetti E, Lash J, Finn PW, et al. Obesity management and chronic kidney disease. Semin Nephrol. 2021;41(4):392–402. https://doi.org/10.1016/j.semnephrol.2021.06.010
46 Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73. https://doi.org/10.7326/M14-0698

Downloaded from http://karger.com/kdd/article-pdf/12/1/18/4457926/000547604.pdf by Universidad de Valencia user on 09 March 2026