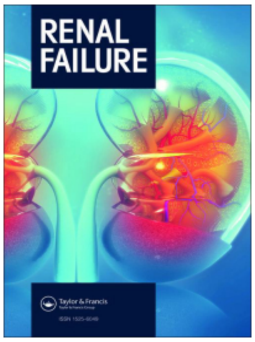

Renal Failure

ISSN: 0886-022X (Print) 1525-6049 (Online) Journal homepage: www.tandfonline.com/journals/irnf20

Natural language processing for kidney ultrasound analysis: correlating imaging reports with chronic kidney disease diagnosis

Chenlu Wang , Ritwik Banerjee , Harry Kuperstein , Hamza Malick , Ruqiyya Bano , Robin L. Cunningham , Hira Tahir , Priyal Sakhuja , Janos Hajagos & Farrukh M. Koraishy

To cite this article: Chenlu Wang , Ritwik Banerjee , Harry Kuperstein , Hamza Malick , Ruqiyya Bano , Robin L. Cunningham , Hira Tahir , Priyal Sakhuja , Janos Hajagos & Farrukh M. Koraishy (2025) Natural language processing for kidney ultrasound analysis: correlating imaging reports with chronic kidney disease diagnosis, Renal Failure, 47:1, 2539938, DOI: 10.1080/0886022X.2025.2539938

To link to this article: https://doi.org/10.1080/0886022X.2025.2539938

Published online: 04 Aug 2025.

RENAL FAILURE
2025, VOL. 47, NO. 1, 2539938
https://doi.org/10.1080/0886022X.2025.2539938

Taylor & Francis
Taylor & Francis Group

CLINICAL STUDY

Natural language processing for kidney ultrasound analysis: correlating imaging reports with chronic kidney disease diagnosis

Chenlu Wang[a], Ritwik Banerjee[a], Harry Kuperstein[b], Hamza Malick[b], Ruqiyya Bano[c], Robin L. Cunningham[d], Hira Tahir[c], Priyal Sakhuja[c], Janos Hajagos[e] and Farrukh M. Koraishy[c]

[a]Department of Computer Science, Stony Brook University, NY, USA; [b]Renaissance School of Medicine, Stony Brook University, NY, USA; [c]Department of Medicine, Stony Brook University, NY, USA; [d]Department of Radiology, Stony Brook University, NY, USA; [e]Department of Biomedical Informatics, Stony Brook University, NY, USA

ABSTRACT

Introduction: Natural language processing (NLP) has been used to analyze unstructured imaging report data, yet its application in identifying chronic kidney disease (CKD) features from kidney ultrasound reports remains unexplored.

Methods: In a single-center pilot study, we analyzed 1,068 kidney ultrasound reports using NLP techniques. To identify kidney echogenicity as either “normal” or “increased,” we used two methods: one that looks at individual words and another that analyzes full sentences. Kidney length was identified as “small” if its length was below the 10th percentile. Nephrologists reviewed 100 randomly selected reports to create the reference standard (ground truth) for initial model training followed by model validation on an independent set of 100 reports.

Results: The word-level NLP model outperformed the sentence-level approach in classifying increased echogenicity (accuracy: 0.96 vs. 0.89 for the left kidney; 0.97 vs. 0.92 for the right kidney). This model was then applied to the full dataset to assess associations with CKD. Multivariable logistic regression identified bilaterally increased echogenicity as the strongest predictor of CKD (odds ratio [OR] = 7.642, 95% confidence interval [CI]: 4.887–11.949; $p < 0.0001$), followed by bilaterally small kidneys (OR = 4.981 [1.522, 16.300]; $p = 0.008$). Among individuals without CKD, those with bilaterally increased echogenicity had significantly lower kidney function than those with normal echogenicity.

Conclusions: State-of-the-art NLP models can accurately extract CKD-related features from ultrasound reports, with the potential of providing a scalable tool for early detection and risk stratification. Future research should focus on validating these models across different healthcare systems.

ARTICLE HISTORY
Received 21 March 2025
Revised 3 July 2025
Accepted 22 July 2025

KEYWORDS
Natural language processing; deep syntactic structures; language embeddings; kidney ultrasound reports; chronic kidney disease

Introduction

Chronic kidney disease (CKD) affects approximately 15% of adults in the United States and is associated with poor health outcomes, increased mortality risk, and substantial healthcare costs [1–3]. Kidney ultrasound is a key imaging modality in the diagnosis and management of CKD [4]. Notably, reduced kidney size and increased cortical echogenicity are key sonographic indicators of CKD [5]. Echogenicity refers to the appearance or texture of the kidney in ultrasound images, specifically how the renal tissue reflects ultrasound waves. Increased echogenicity, often indicative of fibrosis or scarring, can signal kidney damage or worsening CKD [5]. Additionally, small kidneys are defined as those falling below the 10th percentile in size compared to population-based reference values, typically indicating advanced stages of CKD where kidney atrophy is present [4]. A previous study demonstrated that patients with both increased echogenicity and reduced kidney size exhibited the worst prognostic features on kidney biopsy, highlighting the clinical importance of these ultrasound findings [6].

CONTACT Farrukh M. Koraishy Farrukh.Koraishy@stonybrookmedicine.edu Professor of Nephrology/Medicine, Stony Brook University Hospital; Health Science Center, 16, Room 080. 101 Nicolls Road, Stony Brook, NY11716-8166, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0886022X.2025.2539938.

Despite the widespread use of kidney ultrasound reports in patient care, their integration into medical research remains limited. A primary challenge lies in the unstructured nature of free-text reports, which traditional statistical methods cannot efficiently analyze. Natural language processing (NLP) offers a powerful solution by enabling the scalable and accurate extraction of key information from electronic health records (EHRs), facilitating correlations with clinical outcomes [7,8].

CKD is often under-documented in EHRs, and NLP has shown promise in enhancing CKD detection through the analysis of clinical notes [9,10]. Simple NLP approaches, such as word count-based methods, have been employed to improve CKD identification [9], and to predict disease progression, including progression to kidney failure [10,11]. Furthermore, deep learning models have demonstrated high accuracy in classifying CKD directly from ultrasound images [12]. However, despite these advancements, the application of advanced NLP techniques to analyze kidney ultrasound reports remains unexplored. Unlike image visualization, which requires radiologic expertise, written radiology reports are routinely used for medical decision-making by non-radiology clinicians, making their automated analysis particularly valuable.

In this study, we developed an advanced NLP model to identify key CKD indicators—specifically echogenicity and kidney length—from kidney ultrasound reports and assessed their correlation with CKD diagnosis in a large dataset. Our approach leveraged deep syntactic structures [13] (which capture grammatical relationships to derive core sentence meaning) and language embeddings [14] (which represent words as numerical vectors to capture semantic relationships). We hypothesized that our advanced NLP model would exhibit a strong correlation with CKD diagnosis, thereby bridging a critical gap in current knowledge and providing a scalable tool for future research.

Materials and methods

Data source, ethics approval, privacy protection and data security

This study was approved by the Stony Brook University Hospital (SBUH) Institutional Review Board (IRB # IRB2024-00014). Informed consent was waived as all data were deidentified and analyzed anonymously. Initially, study investigators conducting the formal analysis had access to protected health information (PHI) during data extraction from the SBUH electronic health record (EHR) system. These files were securely stored on a HIPAA-compliant server. Following deidentification, all subsequent analyses were conducted exclusively on deidentified patient data. Other authors had access only to deidentified data and summary results during research meetings.

Patient population and data selection

Kidney ultrasound reports were extracted from the SBUH EHR system for hospitalized patients (2020–2024). Each patient was included only once, with the ultrasound report closest to the clinical visit (when diagnosis codes were entered into the EHR) selected for analysis. The initial extraction yielded 1,167 unique reports. Patients with end-stage kidney disease (ESKD) were excluded, as these conditions are associated with abnormal ultrasound morphology (e.g., multiple cysts, shrunken kidneys). Reports containing only addenda were also removed, resulting in a final dataset of 1,068 reports (Supplementary Figure 1). All our reports contained complete information on kidney length and echogenicity. Reports with incomplete data were excluded. Other ultrasound features of kidney disease like cortical thickness and corticomedullary differentiation were documented very infrequently and were not the focus of this study.

Study variables

Basic demographic information (sex, age, race, and ethnicity) was obtained from the SBUH EHR. Comorbid conditions, including CKD, diabetes mellitus (DM), hypertension (HTN), heart failure (HF), coronary artery disease (CAD), chronic obstructive pulmonary disease (COPD), asthma, COVID-19, acute kidney injury (AKI), and cancer, were identified using International Classification of Diseases, Tenth

Revision (ICD-10) codes. Body mass index (BMI) was calculated using recorded height and weight data. Baseline estimated glomerular filtration rate (eGFR) was computed using the 2021 CKD-EPI equation [15].

NLP model development for kidney echogenicity

Sentence extraction and preprocessing

To isolate echogenicity-related sentences, we utilized the Natural Language Toolkit (NLTK) [16] sentence segmentation function. A comprehensive list of echogenicity-related terms was compiled, including 'echogenic,' 'hyperechoic,' 'echotexture,' and 'isoechoic', along with exclusion terms such as 'cyst,' 'mass,' 'lesion,' and 'cancer' to eliminate irrelevant sentences. Sentences were classified based on explicit references to 'right' or 'left' kidneys. If echogenicity was described in reference to 'kidneys' or 'both kidneys', the sentence was assigned to both kidneys.

Word-level NLP method (Figure 1)

A pretrained clinical word embeddings model (clinical-embeddings-100d-w2v-cr) [17] was used to identify 30 key terms related to echogenicity. Following expert validation by nephrologists and radiologists, these terms were categorized as indicative of 'normal,' 'increased,' or 'other' echogenicity.

To refine classification, we employed syntactic parse tree features [18], which represent the hierarchical structure of sentences. This approach enabled us to analyze grammatical relationships, distinguishing negations and semantic context. Sentences were classified into three categories:

1. **Normal** echogenicity – Sentences containing terms like 'normal,' 'unremarkable,' and 'isoechoic.'
2. **Increased** echogenicity – Sentences containing 'echogenic,' 'hyperechoic,' 'hyperechogenicity,' and related terms.
3. **Other** – Sentences describing decreased echogenicity (e.g., 'hypoechoic,' 'anechoic') or ambiguous findings.

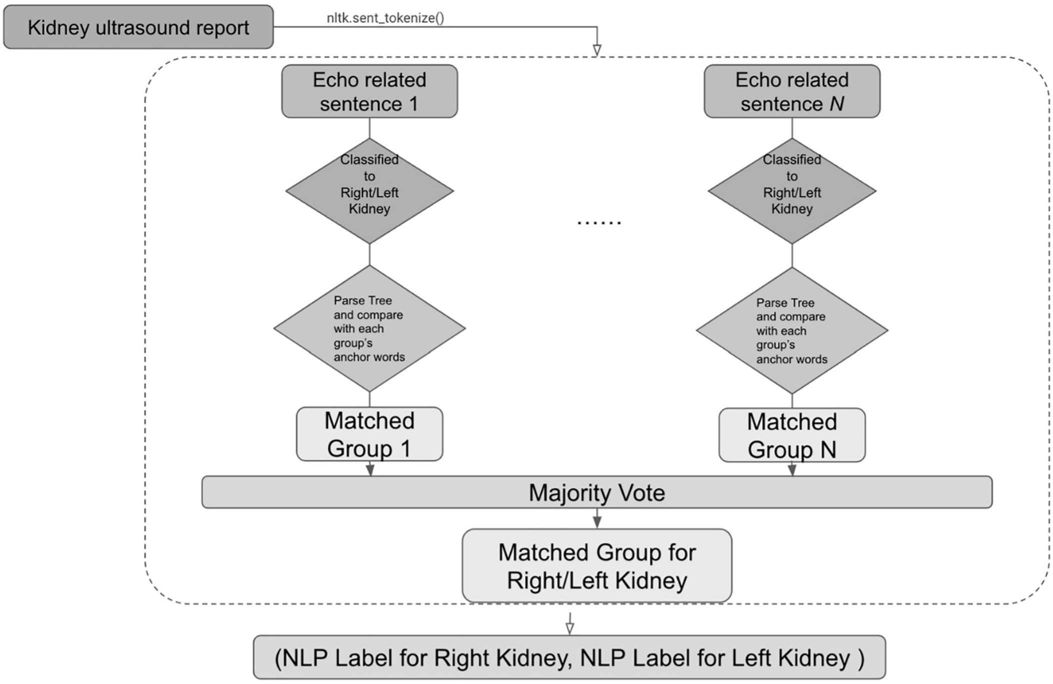
Kidney ultrasound report
nltk.sent_tokenize()
Echo related sentence 1
Echo related sentence *N*
Classified to Right/Left Kidney
Classified to Right/Left Kidney
......
Parse Tree and compare with each group's anchor words
Parse Tree and compare with each group's anchor words
Matched Group 1
Matched Group N
Majority Vote
Matched Group for Right/Left Kidney
(NLP Label for Right Kidney, NLP Label for Left Kidney)

Figure 1. the figure presents the workflow of our word-level NLP method. Initially, the report was segmented into echo-related sentences, which were then individually classified as pertaining either to the right or left kidney. These sentences underwent further parsing to compare specific anchor words. A majority vote was used to decide the final label for each kidney. The outcome was an NLP-generated label for each kidney.

Final classifications were determined using a majority voting method between two independent investigators, with a third reviewer resolving discrepancies when necessary.

Sentence-level NLP method (Figure 2)

We applied BioBERT [19], a pretrained biomedical sentence embedding model, based on the Transformer architecture [20]. BioBERT-generated high-dimensional sentence embeddings, which were compared using cosine similarity to predefined 'anchor sentences' validated by nephrology experts. Each sentence was assigned to the most similar anchor category (**normal**, **increased** echogenicity, or **other**), followed by a majority voting approach for final classification.

Manual physician annotation for ground-truth labels

To establish ground-truth labels, two independent nephrologists annotated 100 randomly selected ultrasound reports, classifying them as:

- Both kidneys normal
- One kidney echogenic
- Both kidneys echogenic
- Other (ambiguous or absent echogenicity data)

The two nephrologists reached 100% consensus, confirming the high reliability of the classification method.

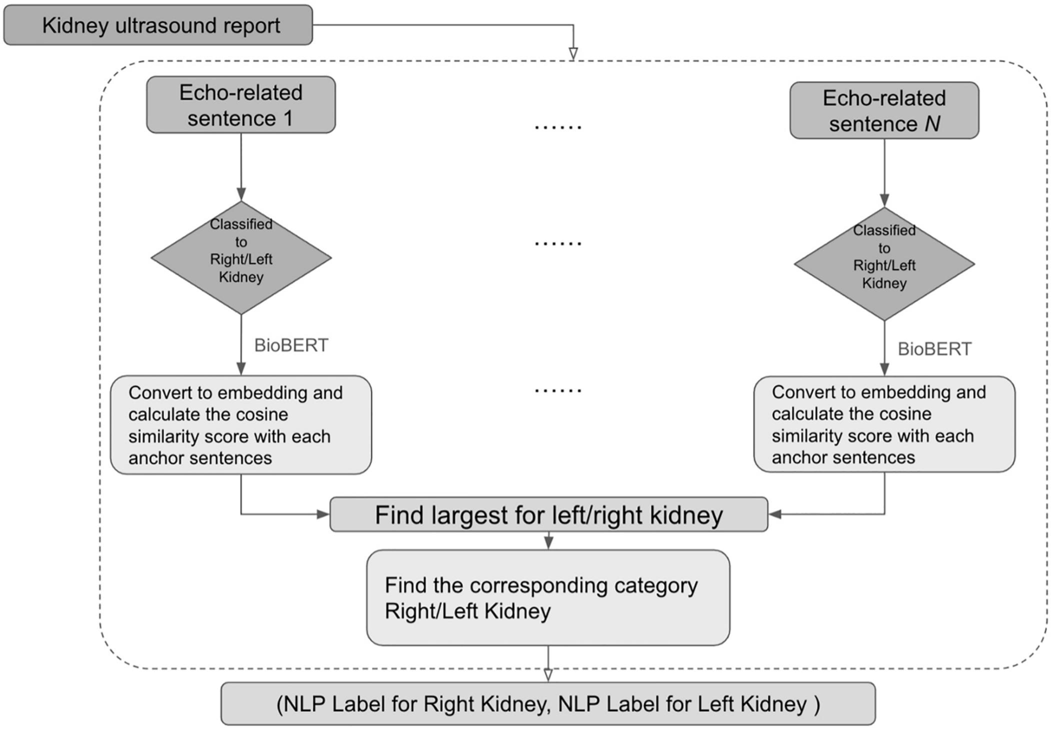
Kidney ultrasound report
Echo-related sentence 1
......
Echo-related sentence *N*
Classified to Right/Left Kidney
......
Classified to Right/Left Kidney
BioBERT
BioBERT
Convert to embedding and calculate the cosine similarity score with each anchor sentences
......
Convert to embedding and calculate the cosine similarity score with each anchor sentences
Find largest for left/right kidney
Find the corresponding category Right/Left Kidney
(NLP Label for Right Kidney, NLP Label for Left Kidney)

Figure 2. The figure illustrates the process of our sentence-level NLP method. Each echo-related sentence from the reports was classified as pertaining to either the right or left kidney. BioBERT was then utilized to convert the classified text into embeddings, with the cosine similarity score being calculated against predefined anchor sentences. The sentence with the highest cosine similarity score for each kidney was identified, determining the most relevant category for the right or left kidney, respectively. The final output was an NLP-generated label indicating the echogenicity condition of each kidney.

NLP model development for kidney length (Figure 3)

We extracted kidney length from ultrasound reports using deep syntactic analysis with a constituency parser based on ELMo (Embeddings from Language Models) [21]. Unlike rule-based methods, ELMo's context-aware embeddings allowed robust extraction of kidney dimensions despite variations in reporting styles.

Key steps included:

- Sentence selection – Identifying sentences mentioning "kidney" and "cm" (centimeters).
- Noise filtering – Excluding irrelevant terms such as "cyst," "foci," and "mass."
- Tree structure parsing – Analyzing syntactic relationships to extract the kidney length (e.g., "right kidney: 9.8 cm").
- Sex-based classification – Categorizing kidney length based on sex-specific reference percentiles.

Statistical analysis

Categorical variables were reported as frequency and percentages, while continuous variables were summarized using mean and standard deviation.

Univariate analyses were conducted using Chi-square test (categorical variables) and Independent samples t-test (continuous variables)

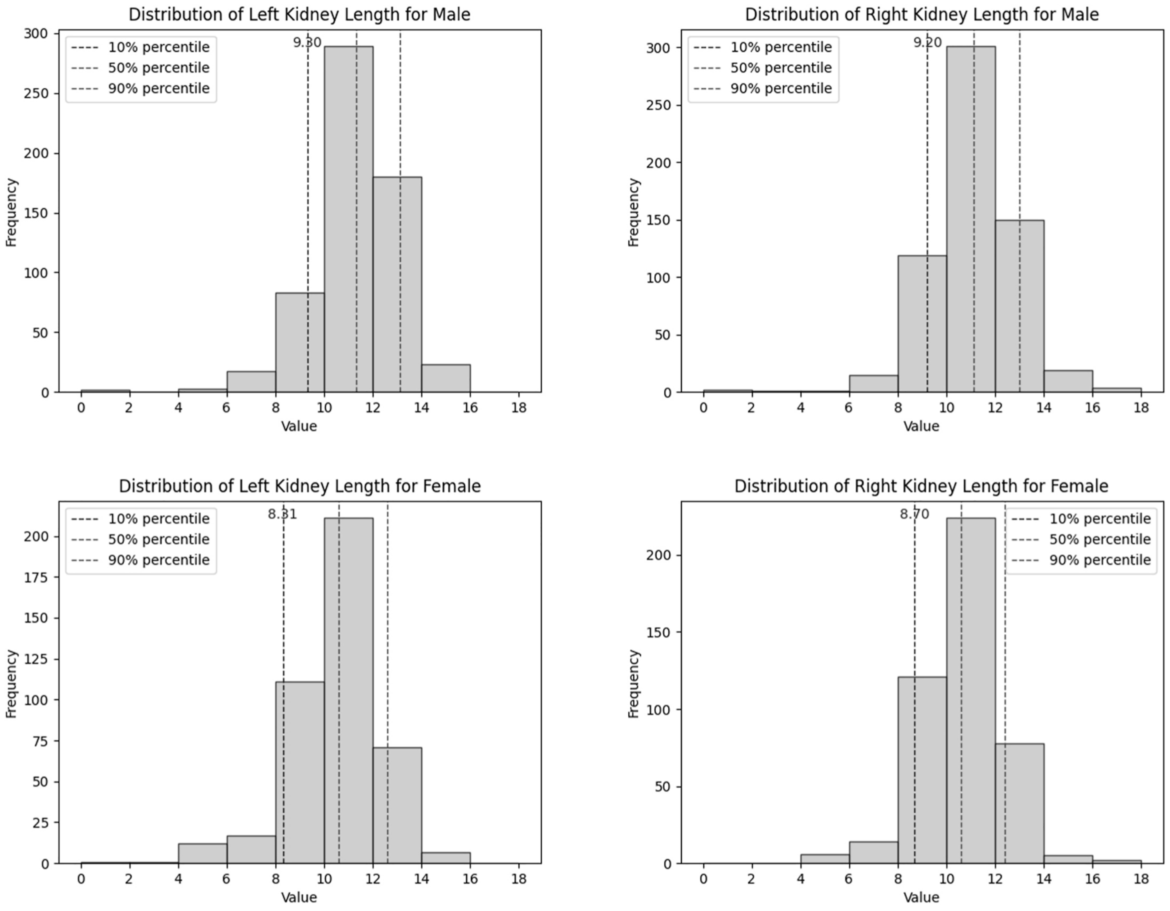

Figure 3. The figure represents the distribution of kidney lengths for male and female patients, with separate histograms for the left and right kidneys. For each subgroup, the 10th, 50th, and 90th percentiles are marked, providing statistical reference points to understand the typical and outlier measurements within the respective populations. The histograms offer a visual comparison between the distributions of kidney length across both sexes, allowing for the identification of potential differences in kidney dimensions between male and female patients.

Multivariable logistic regression was performed for key associations, incorporating variables significant in univariate analyses. Logistic regression was conducted using the statsmodels package [22], with statistical significance set at $p < 0.05$ (Wald test).

All statistical analyses were performed using Python, employing the scipy.stats module [23] and pandas [24].

Results

Comparison of word-level and sentence-level NLP methods for kidney echogenicity labeling (initial model training)

We analyzed 1,068 unique kidney ultrasound reports (Supplementary Figure 1). Two NLP methods—word-level (Figure 1) and sentence-level (Figure 2)—were evaluated using a physician-annotated dataset of 100 reports (ground-truth labels). The word-level approach consistently outperformed the sentence-level method across all key performance metrics (Supplementary Figure 2).

For the right kidney, the word-level method achieved an accuracy of 0.96, compared with 0.89 for the sentence-level method. Similarly, for the left kidney, the word-level method demonstrated an accuracy of 0.97, surpassing the sentence-level method, which scored 0.92.

Given its superior performance, the word-level NLP model was selected for subsequent clinical correlation analyses.

Validation of the word-level NLP model in an independent set of 100 kidney ultrasound reports

In an independent set, kidney ultrasound reports also obtained from our SBUH EHR from 100 patients outside of the study cohort for validation of the NLP before testing for clinical associations. The model demonstrated similar accuracy, precision, recall, and F1-scores compared to those observed with the original set used for ground truth labeling (Table 1).

Assessment of the word-level NLP model based on individual radiologists

Thirteen different radiologists had authored at least two of the 100 reports used for model development. We observed similar levels of accuracy, precision, recall, and F1-scores across reports from different radiologists (Supplementary Table 1).

Kidney length labeling using NLP

Analysis of kidney length distributions revealed significant sex-based differences (Figure 3). Quartile markers (10th, 50th, and 90th percentiles) provided insights into data dispersion and central tendency. Among females, the median kidney length was approximately 10.6 cm for both kidneys. In males, kidney sizes were generally larger, with median lengths of 11.3 cm for the left kidney and 11.1 cm for the right kidney, aligning with previously reported sex-based differences in kidney dimensions [25].

Table 1. Kidney echogenicity classification performance.

100 Kidney US Reports (labeling in the original cohort)						
Side	Overall Accuracy	Class	Precision	Recall	F1-score	Number
Right kidney	0.9375	Increased echogenicity	0.903	0.903	0.903	31
		Normal	0.971	0.957	0.964	69
Left kidney	0.9297	Increased echogenicity	0.848	0.966	0.903	29
		Normal	1.0	0.917	0.957	71
100 Independent Kidney US Reports (independent validation)						
Side ($n = 100$)	Overall Accuracy	Class	Precision	Recall	F1-score	Number
Right kidney	0.94	Increased echogenicity	0.885	1.0	0.939	26
		Normal	0.959	0.986	0.973	74
Left kidney	0.92	Increased echogenicity	0.88	0.917	0.898	25
		Normal	0.933	0.986	0.959	75

Kidneys measuring below the 10th percentile—8.5 cm in females and 9.0 cm in males—were classified as 'small,' consistent with prior definitions of small kidneys [4].

Outlier values (length < 4 cm) were excluded from analysis.

Association between CKD and kidney ultrasound features

Patients with CKD were more likely to have echogenic and small kidneys (Table 2). In multivariable logistic regression (LR) analysis, adjusting for significant features from univariate analysis, the presence of bilaterally echogenic kidneys emerged as the strongest predictor of CKD (OR = 7.642 [4.887, 11.949]; $p < 0.0001$) (Figure 4). The second strongest predictor was bilaterally small kidneys (OR = 4.981 [1.522, 16.300]; $p = 0.0079$). Other factors associated with CKD were male sex, older age, DM, HF and AKI (Figure 4).

Additionally, among patients without a documented CKD diagnosis, those with an eGFR < 60 mL/min/1.73m^2 were more likely to have bilaterally echogenic kidneys (Supplementary Table 2).

Association between kidney echogenicity and clinical features

Compared to those with bilaterally normal kidneys, patients with bilaterally echogenic kidneys were more likely to have a documented CKD diagnosis (Table 3). To explore the correlation between echogenicity and the degree of CKD, we stratified CKD severity based on GFR categories. Increased echogenicity

Table 2. Univariate analysis of patients with and without CKD.

	Total (*N* = 1068)		No CKD (*N* = 564 (52.80%)		CKD (*N* = 504) (47.19%)		
Variables	(Mean/N)	(Std/%)	(Mean/N)	(Std/%)	(Mean/N)	(Std/%)	*p*-value
Demographics							
Sex (*N*, %)							
Female	450	42.13	279	49.29	171	33.92	**5.7E-77**
Race (*N*, %)							
White	796	74.53	415	73.58	381	75.59	0.3800
Non-White	86	8.05	40	7.09	46	9.12	0.3800
Unknown	186	17.41	109	19.32	77	15.27	0.0967
Ethnicity (*N*, %)							
Non-Hispanic	812	76.02	433	76.77	379	75.19	0.6462
Hispanic	83	7.77	47	8.33	36	7.14	0.6462
Unknown	173	16.19	84	14.89	89	17.65	0.2537
Age (Mean, SD)	67.69	19.03	61.91	20.87	74.17	14.19	**4.0E-27**
Co-morbid conditions (*N*, %)							
DM	410	38.38	160	28.36	250	49.60	**1.6E-12**
HF	363	33.98	111	19.68	252	50.00	**3.1E-25**
COPD	178	16.66	72	12.76	106	21.03	**0.0004**
HTN	475	44.47	291	51.59	184	36.51	**1E-06**
CAD	397	37.17	142	25.17	255	50.60	**1.6E-17**
Cancer	202	18.91	106	18.79	96	19.05	0.9783
Asthma	62	5.80	39	6.91	23	4.56	0.1312
COVID-19	156	14.60	75	13.29	81	16.07	0.2323
BMI	28.88	9.68	28.59	8.36	29.20	10.94	0.2954
AKI	611	57.20	213	37.76	398	78.97	**1.1E-41**
Baseline eGFR	63.93	32.81	83.55	29.11	42.86	21.57	**7.8E-112**
Ultrasound Findings							
Kidney Echogenicity							
Both kidney normal (*N*, %)	657	61.51	416	73.75	241	47.81	**5.5E-25**
Both kidneys echogenic (*N*, %)	228	21.34	52	9.21	176	34.92	**5.5E-25**
One kidney echogenic (*N*, %)	15	1.40	7	1.24	8	1.58	**5.5E-25**
Others (*N*, %)	168	15.73	89	15.78	79	15.67	1.0000
Kidney Length							
Both kidneys normal length (*N*, %)	826	77.34	463	82.09	363	72.02	**0.0001**
Both kidneys small (*N*, %)	21	1.96	6	1.06	15	2.97	**0.0001**
One kidney small (*N*, %)	88	8.23	32	5.67	56	11.11	**0.0001**
Others (*N*, %)	133	12.45	63	11.17	70	13.88	0.2111

Abbreviations: DM = diabetes mellitus, HF = heart failure, CKD = chronic kidney disease, COPD = chronic obstructive pulmonary disease, HTN = hypertension, CAD = coronary artery disease, Coronavirus disease 2019 (COVID-19), acute Kidney injury (AKI), BMI = Body Mass Index, eGFR = estimated glomerular filtration rate (mL/min/1.73m^2). Baseline eGFR was calculated using the 2021 CKD-EPI equation. Bold values indicate statistically significant results ($p < 0.05$).

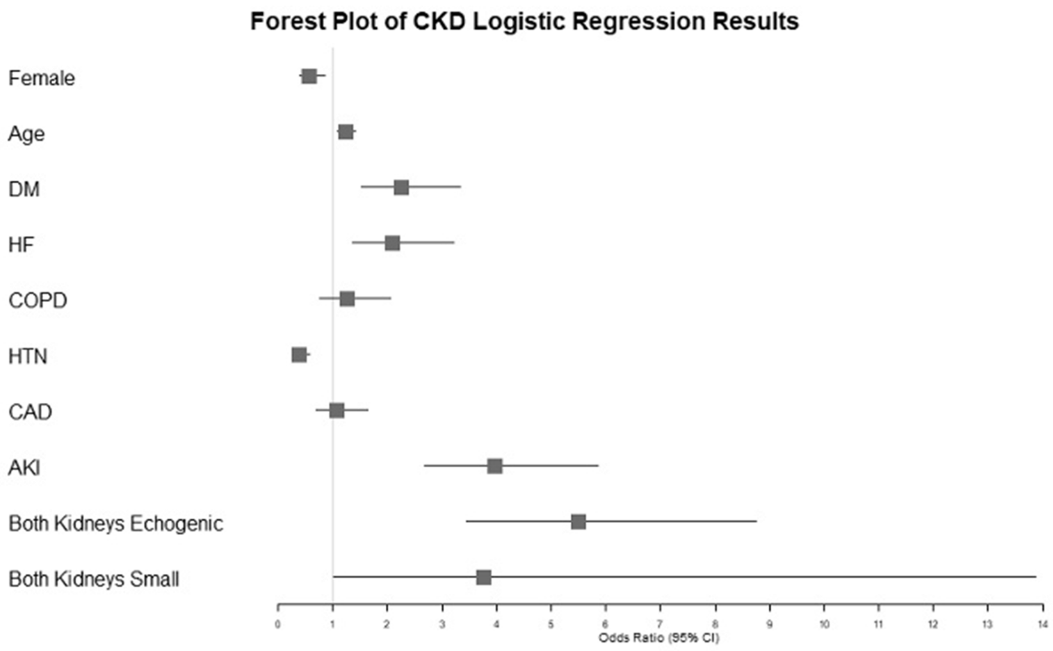

Figure 4. Forest plot showing the associations of clinical features with CKD in a multivariable analysis. Features that were significantly associated with CKD in the univariable analysis were selected. Age reported in units of 10 years increase.

was more commonly associated with advanced CKD (stages 4 and 5), while patients with early stage 3 CKD were more likely to have kidneys with normal echogenicity (Table 2). Low GFR, specially < 60 mL/min/1.73m^2) was strongly correlated with bilateral kidney echogenicity (Figure 5). In multivariable LR analysis, CKD diagnosis was the strongest predictor of bilaterally echogenic kidneys (OR = 4.913 [3.146, 7.672]; $p < 0.0001$) (Supplementary Figure 3). Other factors associated with echogenic kidneys were older age, lower eGFR and COVID-19.

Furthermore, among patients bilaterally echogenic kidneys, those without a CKD diagnosis had a lower mean eGFR (68.74 mL/min/1.73m^2) (Supplementary Table 3), compared to the overall non-CKD group (83.55 mL/min/1.73m^2) (Table 1).

Discussion

To our knowledge, this is the first study to develop an advanced NLP model specifically for extracting kidney ultrasound features associated with CKD, utilizing clinical embeddings and deep syntactic features. Our model demonstrated high accuracy and precision in identifying two key CKD indicators: increased echogenicity and small kidney length. Furthermore, it showed a strong correlation with clinical CKD diagnoses. Notably, the word-level NLP method outperformed the sentence-level approach in classifying increased echogenicity, suggesting that a more granular text analysis enhances classification performance. The model also effectively categorized small kidney length, with bilaterally echogenic kidneys emerging as the strongest predictor of CKD, followed by bilaterally small kidneys. Patients with advanced CKD (GFR < 30 mL/min/1.73m^2 were more likely to have echogenic kidneys. Even among patients without a documented CKD diagnosis, the NLP model identified increased echogenicity in those with eGFR < 60 mL/min/1.73m^2, and these patients exhibited lower mean eGFR.

While computed tomography and magnetic resonance imaging offer superior resolution, ultrasound remains the preferred imaging modality for CKD due to its safety, affordability, and widespread availability. In CKD, cortical echogenicity increases due to fibrosis, a key pathological feature [4,26]. In advanced disease, severe fibrosis leads to kidney atrophy, resulting in reduced kidney length on ultrasound [4,27,28]. Notably, patients with both echogenic and small kidneys have been reported to have the worst histopathological findings on kidney biopsy [6]. Our study aligns with these observations, demonstrating that

Table 3. Univariate analysis of patients with and without bilateral echogenic kidneys.

	Both Kidneys Normal (N=657) (61.51%)		Both Kidneys Echogenic (N=228) (21.34%)		
Variables	(Mean/Number)	(Std Dev/ %)	(Mean/Number)	(Std Dev/ %)	*p*-value
Sex					
Female	268	40.79	96	42.10	0.7877
Race					
White	490	74.58	166	72.80	0.2809
Non-White	49	7.45	23	10.08	0.2809
Unknown	118	17.96	39	17.10	0.8488
Ethnicity					
Non-Hispanic	503	76.56	175	76.75	0.5519
Hispanic	54	8.21	15	6.57	0.5519
Unknown	100	15.22	38	16.66	0.6798
Age	64.95	19.51	74.67	14.94	**1.4E-11**
Co-morbid conditions					
DM	252	38.36	100	43.86	0.166
HF	209	31.81	94	41.23	**0.012**
CKD	241	36.68	176	77.19	**<0.001**
COPD	110	16.74	41	17.98	0.744
HTN	294	44.75	105	46.05	0.792
CAD	223	33.94	110	48.25	**<0.001**
Cancer	120	18.26	45	19.74	0.694
Asthma	43	6.54	7	3.07	0.073
COVID-19	83	12.63	44	19.30	**0.018**
BMI	29.28	10.63	27.47	7.23	0.544
AKI	341	51.90	166	72.81	**<0.001**
Baseline eGFR	71.34	32.84	44.97	25.22	**<0.001**
CKD Stage					
3a	103	41.5	42	25.8	**0.0014**
3b	78	31.5	50	30.7	**0.0014**
4	50	20.2	49	30.1	**0.0014**
5	17	6.9	22	13.5	**0.0014**
Ultrasound Findings					
Kidney Length					
Both kidneys normal length	536	81.56	168	73.68	**0.007**
Both kidneys small	15	2.28	3	1.31	**0.007**
One kidney small	41	6.24	28	12.28	**0.007**
Others	65	9.89	29	12.71	0.285

Abbreviations: DM = diabetes mellitus, HF = heart failure, CKD = chronic kidney disease, COPD = chronic obstructive pulmonary disease, HTN = hypertension, CAD = coronary artery disease, Coronavirus disease 2019 (COVID-19), acute Kidney injury (AKI), BMI = Body Mass Index, eGFR = estimated glomerular filtration rate (mL/min/1.73m^2). Baseline eGFR calculated using the 2021 CKD-EPI equation. Bold values indicate statistically significant results ($p < 0.05$).

increased echogenicity—an earlier feature of CKD—was more prevalent than small kidney length, which typically manifests in later stages. CKD diagnosis is often underreported in billing codes, leading to underdiagnosis in both clinical practice and research29. Importantly, our NLP model identified increased echogenicity in patients with low eGFR, even in the absence of a documented CKD diagnosis, highlighting its potential role in improving disease detection.

The application of NLP in nephrology research is emerging [7], with early studies leveraging rule-based methods to extract CKD-related information from clinical text [9]. However, these approaches often rely on simplistic keyword-based algorithms with limited capacity to interpret the complex language of nephrology. Our study represents a significant advancement by incorporating domain-specific clinical embeddings [17] and deep syntactic features from parse trees. Unlike generic NLP tools, this approach is well-suited for analyzing nephrology-specific terminology, which includes rare and morphologically complex terms [7,29]. Interestingly, our word-level approach outperformed BioBERT embeddings, likely due to the model's ability to capture specialized nephrology and radiology terms that are infrequent in general medical texts. The word-level model also benefited from its ability to focus on precise clinical terminology, such as 'hyperechoic' or 'isoechoic,' which are directly associated with ultrasound findings in CKD. Additionally, integrating syntactic structures allowed our model to capture complex relationships between terms that may be spatially distant within a sentence. For example, when a report described 'bilaterally hyperechoic kidneys with reduced size,' the model could still correctly interpret the combination of both echogenicity and kidney length, even when they appeared in different parts of the

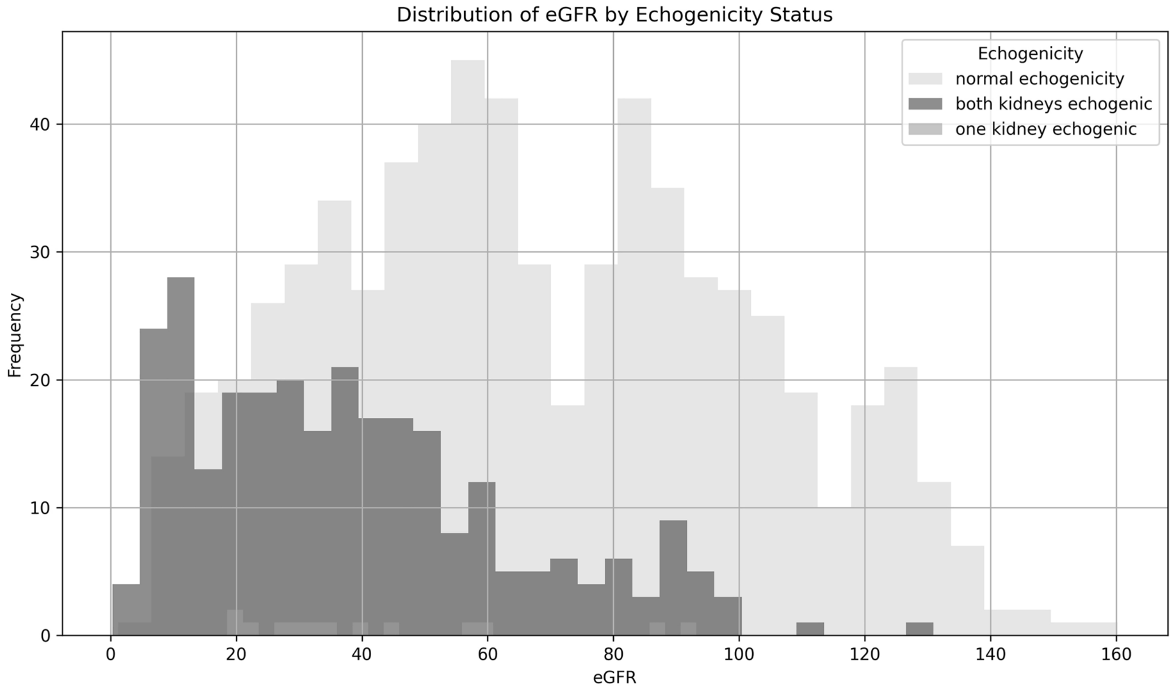

Figure 5. A histogram showing the distribution of eGFR by echogenicity status. Y-axis shows the frequency of reports. X-axis shows eGFR (mL/min/1.73m^2).

sentence. We used Transformer-based clinical embeddings—specifically, contextual representations from ClinicalBERT—in combination with syntactic parse trees to capture both semantic and structural information. For example, in the sentence 'Renal cortical echogenicity is within normal limits bilaterally,' the rule engine detects the pattern renal … echogenicity + within normal limits + bilaterally. The term 'bilaterally' triggers two parallel extractions—one for the left kidney and one for the right—both labeled 'normal echogenicity.' Another example is 'Mild increase in echogenicity.' Here, the phrase increase in echogenicity matches a rule that maps any modifier of increase (e.g., mild, moderate) to the normalized attribute 'increased echogenicity,' so the finding is categorized under increased echogenicity. These findings support the effectiveness of context-sensitive supervised learning systems in clinical NLP [30,31].

Although artificial intelligence has demonstrated performance comparable to radiologists in detecting ultrasound features of human disease [32], radiologist-read reports remain a cornerstone of clinical decision-making. Our NLP model has several potential *clinical applications*. First, since it relies on routinely available kidney ultrasound reports within the electronic health record, it can be seamlessly integrated as a clinical decision-support tool, pending validation in diverse healthcare settings. For instance, clinicians could use this model to assist in the early identification of CKD, particularly in patients who have risk factors but lack a formal CKD diagnosis. If the NLP model detects signs like bilateral increased echogenicity, which is indicative of early kidney damage or fibrosis, it could prompt further testing, such as laboratory work to measure eGFR or a referral to a nephrologist. Additionally, when the model detects small kidneys, it could help identify patients who need closer monitoring or may be eligible for a kidney transplant. The model's simplicity—it requires only two key features (echogenicity and kidney length)—enhances its reproducibility in validation studies. Third, we demonstrated that our NLP approach improves the detection of low kidney function, even in patients lacking a formal CKD diagnosis, underscoring its potential role in early disease identification. The word-level approach, in particular, allows for more granular and efficient analysis of key terms, contributing to better model performance in clinical settings.

While our study offers valuable insights, it has several limitations. First, this was an early-stage pilot study conducted within a single healthcare system, which limits generalizability. To address this, in future studies, we plan external validation using two approaches: (1) integrating semantic embeddings from modern language models to better capture the meaning of clinical language, and (2) applying the model to datasets from other health systems to test performance across different settings.

Kidney ultrasound reports typically categorize echogenicity as either 'normal' or 'increased,' without specifying degrees. Although adding descriptors like 'mild,' 'moderate,' or 'severe' could enhance clinical nuance, such gradations are uncommon and susceptible to interobserver variation. To ensure wide applicability, our NLP model focused on features consistently reported—primarily kidney length and echogenicity. While echogenicity may worsen with CKD progression, we lacked serial imaging data to examine this trend longitudinally, which will be the focus of future research. Other relevant markers, such as cortical thickness or corticomedullary differentiation, were excluded due to infrequent documentation. Vascular indices from Doppler studies were also not included. These gaps highlight the potential value of updating radiology templates to include standardized descriptors for features like cortical thinning. Model performance depends heavily on the quality and consistency of input data. Variability in radiologists' reporting—such as differences in terminology, structure, and detail—can limit accuracy. Adopting standardized templates would reduce this variability, support uniform language use, and improve NLP performance. However, implementing such standards across diverse healthcare environments may be challenging. Partnerships with radiologists and health IT teams will be essential to test the model in systems using standardized reporting formats.

Potential errors in our word-based NLP models should be mentioned. For instance, in one report several earlier sentences explicitly mention the 'right kidney,' but a subsequent sentence states only 'Renal cortical echogenicity is normal' without specifying laterality. Because our current context-propagation logic is insufficient, the system mistakenly interprets this sentence as applying to both kidneys rather than just the right one, leading to label confusion. Highlighting such 'missing laterality' errors will help us identify and prioritize improvements in cross-sentence context handling.

Our model was intentionally designed to avoid opaque, black-box methods in favor of interpretable features. While this increases transparency, it can also make the model more vulnerable to learning site-specific language patterns. Although advanced language models offer strong predictive power, their lack of interpretability can limit clinical adoption. Clinicians must understand a model's reasoning to trust its output. Future efforts will include developing tools that explain model decisions—such as SHAP values or attention maps—and comparing our current interpretable model to more complex alternatives. Enhancing report standardization and interpretability will be key to broader application, particularly in capturing underreported but clinically important features like cortical thinning [33].

In summary, we demonstrate that state-of-the-art NLP models can effectively extract CKD-related features from kidney ultrasound reports, with potential implications for improving disease detection and diagnosis. Future research should focus on validating these models across different healthcare systems, exploring their integration into clinical workflows, and addressing the limitations of data variability and model interpretability. Standardizing clinical reports could enhance NLP model performance and further advance the potential for automated CKD detection, especially in settings where timely diagnosis is critical.

Acknowledgments

Chenlu Wang was involved in data cleaning, conducted all the NLP analyses and contributed to writing the original draft. Ritwik Banerjee supervised the NLP design and analyses and contributed to writing the revised draft. Harry Kuperstein, Hamza Malick, Hira Tahir, and Priyal Sakhuja were involved in literature review and writing the original draft. Ruqiyya Bano contributed to statistical analyses and to writing the revised draft. Robin Cunningham was involved in guidance on the interpretation of kidney ultrasound reports and contributed to writing the revised draft. Janos Hajagos was involved in data procurement and contributed to writing the revised draft. Farrukh Koraishy was involved in study design, statistical analyses, writing the original and revised draft versions and supervision of the project.

Statement of ethics

This study protocol was reviewed and approved by the Stony Brook University Hospital (SBUH) Institutional Review Board (IRB # IRB2024-00014). Consent was not obtained (waived) since the data were deidentified and analyzed anonymously. The SBUH IRB granted an exemption from requiring written informed consent, Initially, the authors involved in the formal analysis had access to protected health information (PHI) when it was extracted from the SBUH EHR system. These files were stored on a secure, HIPAA-compliant server. PHI was promptly removed, and all subsequent analyses were conducted on de-identified patient data. All other authors had access only to de-identified data and summary results during research meetings.

Author contributions

CRediT: **Chenlu Wang**: Data curation, Formal analysis, Investigation, Software, Visualization, Writing – original draft; **Ritwik Banerjee**: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Writing – review & editing; **Harry Kuperstein**: Resources, Writing – review & editing; **Hamza Malick**: Resources, Writing – review & editing; **Ruqiyya Bano**: Formal analysis, Software, Validation, Writing – review & editing; **Robin L. Cunningham**: Investigation, Resources, Writing – review & editing; **Hira Tahir**: Resources, Writing – review & editing; **Priyal Sakhuja**: Resources, Writing – review & editing; **Janos Hajagos**: Resources, Writing – review & editing; **Farrukh M. Koraishy**: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

None.

Data availability statement

The deidentified data used in this study is with the corresponding author and will be available on request to editors, reviewers and readers *via* a data sharing agreement.

References

[1] Kalantar-Zadeh K, Rhee CM, Chou J, et al. The obesity paradox in kidney disease: how to reconcile it with obesity management. Kidney Int Rep. 2017;2(2):271–281. doi: 10.1016/j.ekir.2017.01.009.

[2] (NIDDK) NIoDaDaKD. USRDS 2022 Annual Data Report. https://usrds-adr.niddk.nih.gov/2022. Published 2022. Accessed.

[3] Kalantar-Zadeh K, Jafar TH, Nitsch D, et al. Chronic kidney disease. Lancet. 2021;398(10302):786–802. doi: 10.1016/S0140-6736(21)00519-5.

[4] Ahmed S, Bughio S, Hassan M, et al. Role of ultrasound in the diagnosis of chronic kidney disease and its correlation with serum creatinine level. Cureus. 2019;11(3):e4241. doi: 10.7759/cureus.4241.

[5] Petrucci I, Clementi A, Sessa C, et al. Ultrasound and color Doppler applications in chronic kidney disease. J Nephrol. 2018;31(6):863–879. doi: 10.1007/s40620-018-0531-1.

[6] Moghazi S, Jones E, Schroepple J, et al. Correlation of renal histopathology with sonographic findings. Kidney Int. 2005;67(4):1515–1520. doi: 10.1111/j.1523-1755.2005.00230.x.

[7] Van Vleck TT, Farrell D, Chan L. Natural language processing in nephrology. Adv Chronic Kidney Dis. 2022;29(5): 465–471. doi: 10.1053/j.ackd.2022.07.001.

[8] Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23(5):1007–1015. doi: 10.1093/jamia/ocv180.

[9] Chase HS, Radhakrishnan J, Shirazian S, et al. Under-documentation of chronic kidney disease in the electronic health record in outpatients. J Am Med Inform Assoc. 2010;17(5):588–594. doi: 10.1136/jamia.2009.001396.

[10] Singh K, Betensky RA, Wright A, et al. A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol. 2016;11(12):2150–2158. doi: 10.2215/CJN.02420316.

[11] Makino M, Yoshimoto R, Ono M, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep. 2019;9(1):11862. doi: 10.1038/s41598-019-48263-5.

[12] Kuo CC, Chang CM, Liu KT, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. NPJ Digit Med. 2019;2(1):29. doi: 10.1038/s41746-019-0104-2.

[13] Shabbir S. Exploring Language Syntax and Structure in Deep NLP.https://medium.com/@datascientist_SheezaShabbir/exploring-language-syntax-and-structure-in-deep-nlp-bbcc9e4474cb. Published 2023. Accessed; 2025.

[14] Turing. A Guide on Word Embeddings in NLP. https://www.turing.com/kb/guide-on-word-embeddings-in-nlp. Published 2025. Accessed 2025.

[15] Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med. 2021;385(19):1737–1749. doi: 10.1056/NEJMoa2102953.

[16] Python Text Processing with NLTK 2.0 Cookbook [computer program]. Birmingham, UK: Packt Publishing Ltd; 2010.

[17] Flamholz ZN, Crane-Droesch A, Ungar LH, et al. Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information. J Biomed Inform. 2022;125:103971. doi: 10.1016/j.jbi.2021.103971.

[18] Ágel V. Dependency and valency. Berlin: De Gruyter; 2003.

[19] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240. doi: 10.1093/bioinformatics/btz682.

[20] Vaswani A, Shazeer NM, Parmar N Attention is All you Need. Paper presented at: neural Information Processing Systems, et al. 2017.

[21] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations; 2018.

[22] Seabold S, Perktold J. Statsmodels: econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference. 2010;2010.

[23] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–272. doi: 10.1038/s41592-019-0686-2.

[24] McKinney W. Python for Data Analysis: data Wrangling with Pandas, NumPy, and IPython. Sebastopol, CA: O'Reilly Media, Incorporated; 2017.

[25] Kalucki SA, Lardi C, Garessus J, et al. Reference values and sex differences in absolute and relative kidney size. A Swiss autopsy study. BMC Nephrol. 2020;21(1):289. doi: 10.1186/s12882-020-01946-y.

[26] Panizo S, Martínez-Arias L, Alonso-Montes C, et al. Fibrosis in chronic kidney disease: pathogenesis and consequences. Int J Mol Sci. 2021;22(1):408. doi: 10.3390/ijms22010408.

[27] Jovanović D, Gasic B, Pavlovic S, et al. Correlation of kidney size with kidney function and anthropometric parameters in healthy subjects and patients with chronic kidney diseases. Ren Fail. 2013;35(6):896–900. doi: 10.3109/0886022X.2013.794683.

[28] Zhang WX, Zhang ZM, Cao BS, et al. Sonographic measurement of renal size in patients undergoing chronic hemodialysis: correlation with residual renal function. Exp Ther Med. 2014;7(5):1259–1264. doi: 10.3892/etm.2014.1560.

[29] Farrell D, Chan L. Application of natural language processing in nephrology research. Clin J Am Soc Nephrol. 2023;18(6):806–808. doi: 10.2215/CJN.0000000000000118.

[30] Ventrella P, Delgrossi G, Ferrario G, et al. Supervised machine learning for the assessment of Chronic Kidney Disease advancement. Comput Methods Programs Biomed. 2021;209:106329. doi: 10.1016/j.cmpb.2021.106329.

[31] Ortiz A, Portoles J, Pino-Pino MD, et al. Clinical characteristics and management of patients with secondary hyperparathyroidism undergoing hemodialysis: a feasibility analysis of electronic health records using natural language processing. Kidney Dis (Basel). 2023;9(3):187–196. doi: 10.1159/000528784.

[32] Potipimpanon P, Charakorn N, Hirunwiwatkul P. A comparison of artificial intelligence versus radiologists in the diagnosis of thyroid nodules using ultrasonography: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol. 2022;279(11):5363–5373. doi: 10.1007/s00405-022-07436-1.

[33] Yamashita SR, von Atzingen AC, Iared W, et al. Value of renal cortical thickness as a predictor of renal function impairment in chronic renal disease patients. Radiol Bras. 2015;48(1):12–16. doi: 10.1590/0100-3984.2014.0008.