Predicting Chronic Kidney Disease Progression from Stage III to Stage V using Language Models

Zainab Awan
Queen Mary University of London, UK
z.awan@qmul.ac.uk

Rafael Henkin
Queen Mary University of London, UK
r.henkin@qmul.ac.uk

Nick Reynolds
Newcastle University, UK
nick.reynolds@newcastle.ac.uk

Michael R. Barnes
Queen Mary University of London, UK
m.r.barnes@qmul.ac.uk

Abstract

Chronic Kidney Disease (CKD) is a global health challenge, affecting 5–10% of the population, with a significant burden on healthcare systems. Early prediction of CKD progression from stage III to stage V is crucial to enable timely interventions. Traditional predictive methods rely on biochemical markers and demographic factors, but are often limited by issues such as missing data and reliance on structured inputs. This study explores the potential of several encoder-based language models, to predict CKD progression using a cohort from the Clinical Practice Research Datalink (CPRD) GOLD database. We applied both Full Fine-Tuning (FFT) and Parameter-Efficient Fine-Tuning (PEFT) with LoRA to pre-trained models, comparing them against traditional machine learning algorithms such as Random Forest and XGBoost. Our results show that fine-tuned models, particularly dmis-lab/biobert-v1.1-FFT, outperform traditional models in predicting CKD progression, with an AUC of 0.7787, precision of 0.7261, and accuracy of 0.7045. Although LoRA-based models are more computationally efficient, they consistenly exhibit lower performance. These findings suggest that fine-tuned encoder models hold significant potential for improving CKD progression prediction. However, there is still room for further enhancement in their accuracy and applicability in clinical settings.

1 Introduction and Related Work

1.1 Introduction

Chronic Kidney Disease (CKD) is one of the leading causes of mortality worldwide, affecting approximately 5–10% of the global population (Eknoyan et al., 2004; Martínez-Castelao et al., 2014). The disease imposes a significant burden on healthcare systems, and early prediction of CKD progression is crucial for improving patient outcomes. CKD is classified into five stages: stage I, stage II, stage III, stage IV, and stage V —based on estimated glomerular filtration rate (eGFR) values: stage I (eGFR 90), stage II (60 eGFR 89), stage III (30 eGFR 59), stage IV (15 eGFR 29), and stage V (eGFR < 15). Accurate prediction of progression from stage III to stage V is critical to enable timely interventions that can help mitigate associated risks.

The United States Renal Data System (USRDS) report indicates that approximately 35.4% of CKD patients are referred to interdisciplinary programs later than recommended, likely due to insufficient risk profile classification (Isaza-Ruget et al., 2024; Mendelssohn et al., 2009). This delay can compromise the effectiveness of potential treatments, highlighting the need for more efficient methods of early detection and intervention.

Current predictive methods rely heavily on biochemical markers like urinary albumin/creatinine ratio (uACR), eGFR, and demographic factors such as age and sex. While these models are useful, they often suffer from limitations, such as missing data in biochemical measures, making imputation unreliable and potentially leading to biased predictions. However, it is important to note that bias is not exclusive to these methods—pre-trained language models and other machine learning approaches can also exhibit biases, depending on data distributions. Additionally, existing risk calculators are often constrained by structured data and require extensive manual feature engineering, which can limit their flexibility and adaptability.

The potential for language models to improve CKD progression prediction remains largely unexplored, especially in the context of large, complex datasets such as those from the Clinical Practice Research Datalink (CPRD). This study seeks to address this gap by applying state-of-the-art encoder models like BioBERT and ClinicalBERT to predict CKD progression, focusing on domain-specific fine-tuning to improve prediction accuracy. The

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 236–242

motivation for this approach stems from the recognition that language models pre-trained on medical texts can uncover subtle patterns in clinical data that traditional models may miss.

Our key contributions are as follows:

- Comparing domain-specific and general-domain BERT models for CKD progression prediction.
- Benchmarking BERT models against traditional machine learning approaches (XGBoost and Random Forest).
- Assessing Parameter-Efficient Fine-Tuning (PEFT) as a resource-efficient adaptation method.

1.2 Related Work

A significant body of research has focused on predicting CKD progression using machine learning models. For instance, a study by Isaza-Ruget et al. (2024) utilized logistic regression, random forests, and neural networks for CKD progression prediction. This study incorporated a variety of patient features, including demographics, lab results, and comorbidities, to build a robust risk prediction model. Despite promising results, traditional models like these are often constrained by the need for structured data and manual feature engineering, which can limit their scalability and accuracy when applied to diverse populations.

Similarly, Klinrisk's proprietary machine learning model (Tangri et al., 2024), validated in clinical trial populations such as CANVAS (Neal et al., 2013) and CREDENCE (Jardine et al., 2018), demonstrated improved prediction of CKD progression compared to the Kidney Disease Improving Global Outcomes (KDIGO) heatmaps and kidney failure risk equations (KFRE). These models rely on routinely collected laboratory data like eGFR and albuminuria. However, they still face challenges when dealing with unstructured clinical data or missing information, which transformer models could address more effectively. The study by Zhu et al. (2023), employs recurrent neural networks for CKD progression prediction. Their model achieved an AUROC of 0.957 with eGFR time-series data alone, improving to 0.967 with additional clinical variables.

In a similar vein, Reddy et al. (2024) developed explainable machine learning models, including decision trees and random forests, to predict CKD progression. Their models achieved high predictive accuracy (ROC-AUC: 0.94–0.98) using key variables like eGFR slope and recent eGFR.

Saito et al. (2024) applied time-series clustering and LightGBM to stratify patients based on eGFR trajectories, achieving a prediction accuracy of 0.675. According to Shapley values, the most predictive features included baseline eGFR, hemoglobin, and BMI, reinforcing the importance of these variables in forecasting renal function decline.

2 Methodology

This study aimed to predict the progression of chronic kidney disease (CKD) from stage III (CKD III) to stage V (CKD V) using a cohort of patients from the CPRD GOLD database. To address class imbalance, we employed age as a covariate in the propensity score matching process, ensuring comparability between patients with differing progression outcomes. For prediction, we utilised machine learning models, including traditional algorithms (Random Forest and XGBoost) and encoder-based language models. Our goal was to develop models for predicting CKD progression using both approaches.

We fine-tuned pre-trained models using Full Fine-Tuning (FFT) and Parameter-Efficient Fine-Tuning (PEFT) with LoRA, while also optimizing hyperparameters for Random Forest and XGBoost models. The following sections detail these approaches, including their implementation and evaluation.

2.1 Cohort Selection Criteria

We selected patients from the CPRD GOLD database who were registered in a GP practice between 01/01/2010 and 31/12/2020, aged 16 years or older, and had two or more long-term conditions (LTCs). We used READ v2 and ICD10 codes to identify individuals with CKD, specifically targeting stages III and V. A list of the relevant codes is available in the provided GitHub link. We excluded secondary care events that occurred after patients were transferred out of their GP practices, resulting in a distribution of 206,553 patients in class 0 (CKD3) and 4,606 patients in class 1 (CKD5). We then refined the cohort by removing patients from the negative class (CKD3) who had a median follow-up period of less than 6 years-2.4 months, excluding 93,926 patients. We also excluded 166

patients with terms related to preparatory care for dialysis, renal transplant planning, ligation of arteriovenous dialysis fistulas, acute hypercalcaemia of dialysis, or creation of graft fistulas for dialysis. The final cohort comprised 122,267 patients in class 0 (CKD3) and 4,606 patients in class 1 (CKD5).

2.2 Age-Matched Cohort

To reduce bias from confounding variables and address the extreme imbalance between the negative and positive classes in our dataset, we used the MatchIt R package with 1:1 nearest neighbor (NN) for propensity score matching (PSM). This approach matched patients from the CKD progression group with those from the non-progression group based on the key covariate: age. By minimising the confounding effect of age, which significantly influences CKD progression, we ensured a balanced and fair comparison between the two groups, despite the severe class imbalance. We chose age as the sole matching criterion because it is a critical risk factor for CKD progression. Differentiating between physiological and pathological kidney function decline becomes increasingly challenging with age (Noronha et al., 2022). Balancing the age distribution between the progression and non-progression cohorts was essential, given the strong link between ageing and renal function decline. After matching, the dataset included 4,596 instances in both the positive and negative classes.

2.3 Data Summary Statistics Table

Table 1 summarizes the key characteristics of the dataset, providing an overview of the variables and their distribution, which informs the subsequent analysis.

2.4 Full fine-tuning and Parameter Efficient fine-tuning (LoRA)

We framed CKD progression prediction as a sequence classification task, where each input sequence S represents a concatenation of patient-specific attributes and can be defined as in Equation 1. :

$$S = [E, C_1, P_1, C_2, P_2, \ldots, C_n, P_n], \quad (1)$$

where E denotes the patient's ethnicity, Ci represents the i-th LTC, and Pi denotes the i-th continuous prescription. A list of all possible LTCs can be found in the GitHub link: AI MULTIPLY GOLD Read Codes. A continuous prescription is defined as a group of consecutive prescriptions where each pair of prescriptions is at most 84 days (Guthrie et al., 2011) apart. This group of consecutive prescriptions must contain at least three prescriptions (Connor et al., 2024). We included continuous prescriptions in our analysis that are known to be associated with drug-induced renal injury and nephrotoxicity. A comprehensive list can be found in (Connor et al., 2024). The sequence length is variable and depends on the number of recorded conditions and prescriptions for each patient. Labels were assigned as y=1 for cases (progression) and y=0 for controls (non-progression). To reduce potential confounding, we introduced a 6-month buffer period before CKD stage III diagnosis, excluding clinical events that occurred within this window. A patient's CKD stage III diagnosis date might not reflect the exact onset of kidney dysfunction. Events occurring just before diagnosis might be influenced by external factors rather than true disease progression.

3 Experimental Setup

We evaluated several pre-trained encoder-based models, including UFNLP/gatortron-base (Yang et al., 2022), bert-base-uncased (Devlin, 2018), dmis-lab/biobert-v1.1 (Lee et al., 2020), microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext (Gu et al., 2021), allenai/scibert_scivocab_uncased (Beltagy et al., 2019), bionlp/bluebert_pubmed_mimic_uncased_L (Peng et al., 2019), and medicalai/ClinicalBERT (Huang et al., 2019). Model training was conducted using the Hugging Face Transformers library, with each model fine-tuned over three epochs. We tokenised the sequences to a maximum context length of 512, used a learning rate of 2e-5, and used AdamW optimization with weight decay of 0.001. Stepwise decay of the learning rate (gamma = 0.1) was applied, along with gradient clipping (max norm 1.0) to prevent exploding gradients. Early stopping was used to stop training when the validation error did not improve.

We compared LoRA (Low-Rank Adaptation) (Hu et al., 2021) and full fine-tuning (FFT) for CKD progression prediction, both of which used similar configurations (learning rate of 2e-5, 5 epochs, maximum sequence length of 512, and

Table 1: Summary Statistics of CKD Cohort

Variable	**Class 1 (N = 4,596)**	**Class 0 (N = 4,596)**
Age (mean ± SD)	66.34 ± 14.37	66.51 ± 13.84
Sex (Male / Female)	2,663 / 1,933	2,705 / 1,891
Ethnicity	-	-
White (%)	86.79	88.79
Black or Black British (%)	2.08	3.62
Asian or Asian British (%)	6.82	2.78
Mixed (%)	0.67	0.34
Unknown (%)	0.36	5.14
Chinese or Other Group (%)	1.71	0.84
Median progression time Stage V	6.24 years	NA

batch size of 8). In full fine-tuning (FFT), all model parameters are updated during training, which can be computationally expensive. In contrast, LoRA adapts the model weights using low-rank matrices with a reduced number of trainable parameters. Specifically, we apply a LoRA adaptation parameter r=16, which controls the rank of the matrices and significantly reduces the number of parameters being trained. This makes LoRA a more computationally efficient alternative to full fine-tuning, particularly for large pre-trained models.

Both methods were evaluated using stratified 5-Fold cross-validation, where each fold was split into training, validation, and test sets. The validation set comprised 10% of the training data, stratified by class labels. We reported the performance metrics (accuracy, F_1-score, precision, recall, and AUC) averaged across folds, using mean values.

We tokenised and encoded input sequences using each model's corresponding tokeniser, applying padding and truncation to ensure uniform input lengths. Training was performed with a batch size of 32 for FFT and 8 for LoRA, and Data-Parallel was used when multiple GPUs were available. The validation performance was assessed after each epoch, and the model with the lowest validation loss was selected for testing. The training process involved optimizing the models using the AdamW optimizer with weight decay and adjusting the learning rate using stepwise decay.

For each fold, the best model was evaluated on the corresponding test set. Predictions were made using softmax probabilities, which allowed us to compute additional metrics such as area under the receiver operating characteristic curve (AU-ROC), accuracy, precision and recall.

For the tabular models, we conducted a grid search to optimize hyperparameters for XGBoost (learning rate, max depth, and number of estimators) and Random Forest (number of estimators, max depth, and minimum samples per split). We employed 5-fold cross-validation, training the models on training subsets and evaluating them on validation subsets. We report averaged performance metrics: accuracy, F_1 score, precision, recall, and ROC AUC—across folds and record the best performing hyperparameter configurations for each metric.

4 Evaluation and Results

Figure 1 compares the performance of various models using five metrics: Accuracy, F_1, Precision, Recall, and AUC (Area Under the Curve). Models evaluated include different fine-tuning strategies FFT and LoRA in addition to RF and XGBoost.

The model dmis-lab/biobert-v1.1-FFT has the highest performance across most metrics, particularly AUC (0.7787), Precision (0.7261), Accuracy (0.7045) and F_1 scores (0.6890). Its recall is low (0.6622), meaning the model is highly selective in identifying progression but fails to detect many actual cases. In practice, this could mean missing patients whose disease progression might have slowed with earlier intervention. While some models like UFLNLP/gatortron-base-FFT perform well in accuracy (0.7034) and recall (0.7293), they slightly lag in precision (0.6477), which might not be ideal for our clinical applications. The contrasting performance of our fine-tuned models in precision and recall highlights the trade-off between these two metrics. A potential approach to mitigate this is employing a Mixture of Experts (MoE) architecture with a gating mechanism. Future work will explore MoE's effectiveness in optimizing both

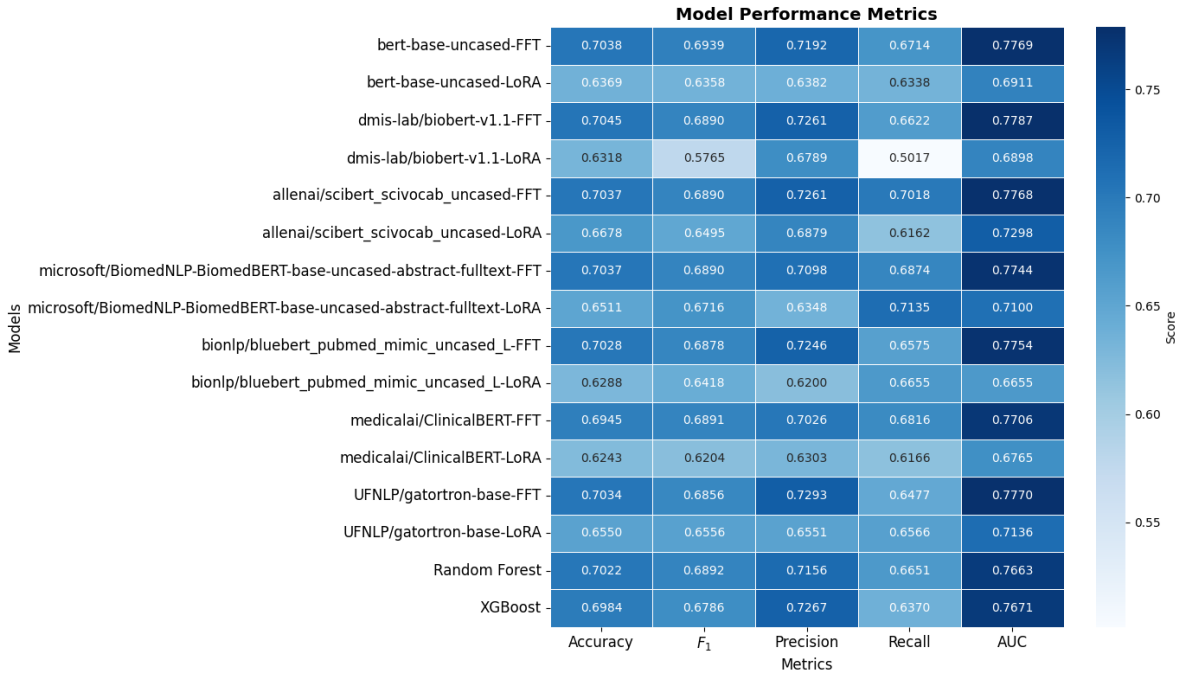

Figure 1: Heatmap of performances across various metrics.

precision and recall in CKD progression prediction.

Models fine-tuned using FFT generally outperform their LoRA counterparts across all metrics. This trend is consistent for models like bert-base-uncased, allenai/scibert_scivocab_uncased, and microsoft/BiomedNLP-BiomedBERT.

While traditional methods like Random Forest and XGBoost perform reasonably well (AUC of 0.7663 and 0.7671, respectively), they lag behind transformer-based models fine-tuned with FFT, particularly in metrics like Precision and Recall. Models pre-trained on biomedical data, such as dmis-lab/biobert-v1.1, microsoft/BiomedNLP-BiomedBERT, and bionlp/bluebert_pubmed_mimic_uncased, tend to perform better than general domain models like bert-base-uncased in terms of accuracy. While dmis-lab/biobert-v1.1-FFT achieved the best accuray, its recall (0.6622) remains a concern in clinical settings where minimizing false negatives is critical.

5 Discussion

In this study, we demonstrate the potential of encoder-based models for predicting CKD progression from stage III to stage V using LTCs, continuous prescriptions, and ethnicity from CPRD. To achieve this, we developed three types of models: full fine-tuning, parameter-efficient fine-tuning (PEFT) using Low-Rank Adaptation (LoRA), and tabular models, including Random Forest (RF) and XGBoost. Model names bearing the suffix *FFT* indicate that the models have been fully fine-tuned, whereas those with the suffix *LoRA* represent Low-Rank Adaptation fine-tuning, a method categorised under PEFT.

While our primary aim was to evaluate the potential of fine-tuned language models for predicting CKD progression, we also included tabular models in the study. This enabled us to compare the performance of advanced deep learning methods with traditional models like RF and XGBoost, which are often better suited to structured data. By incorporating both approaches, we provide a comprehensive assessment of the different modeling techniques for this task.

The results indicate that FFT consistently outperforms PEFT using LoRA across all evaluated metrics, particularly in recall and AUC suggesting that full adaptation of pre-trained models is necessary for tasks as complex as CKD progression. Among the fine-tuned models, those pre-trained on biomedical corpora, such as BioBERT, ClinicalBERT, and BlueBERT, demonstrate strong performance, with AUC values around 0.77. This reinforces the importance of domain-specific pre-training for clinical prediction tasks. Among all models tested, dmis-lab/biobert-v1.1-FFT achieved the highest

AUC (0.7787), Precision (0.7261), and Accuracy (0.7045), indicating its robustness in CKD progression prediction tasks. Its domain-specific pretraining on biomedical text, coupled with fully fine-tuned (FFT) models, has proven promising for the task of CKD progression prediction.

LoRA-based models exhibit lower performance, with AUC scores ranging between 0.6911 and 0.7298. While LoRA fine-tuning offers computational efficiency, its lower recall and precision suggest limitations in capturing subtle predictive patterns in the data. Notably, some LoRA models, such as BiomedBERT-LoRA, show comparatively better recall (0.7135) but at the expense of precision (0.6348), indicating a tendency towards higher false positives. While LoRA's efficiency may be compelling in resource-constrained scenarios, its limited recall capabilities could make it less suitable for critical, high-stakes clinical applications

The results highlight the importance of recall in high-stakes applications like CKD progression prediction, where minimizing false negatives is crucial for timely intervention. Among the models tested, bert-base-uncased-FFT achieves the highest recall, suggesting its potential for capturing at-risk patients. However, as a general-domain model fine-tuned on CPRD data, it lacks the medical domain specificity of models like BioBERT.

Interestingly, tabular models, including RF and XGBoost, perform competitively with language models. XGBoost and RF achieve AUC of 0.7671 and 0.7663, closely matching several fine-tuned language models.

6 Conclusion and Future Work

In conclusion, the study demonstrates that encoder models, particularly BioBERT FFT, significantly contribute to predicting the progression of CKD. Through the use of domain-specific pre-training and fine-tuning strategies, BioBERT surpasses traditional machine learning methods such as Random Forest and XGBoost. By identifying patterns in clinical data, BioBERT shows promise in predicting CKD progression with an accuracy of nearly 70%. While this isn't perfect, it points to the model's potential for advancing predictive analytics in kidney disease and could ultimately support better decision-making in both research and clinical settings.

Although the findings show potential, further improving the model's accuracy is essential for its practical application in medical settings. Therefore, future work will focus on extending this study to include prompt-based decoder models in few-shot and zero-shot settings with Chain-of-Thought reasoning, potentially incorporating domain knowledge. Additionally, we plan to evaluate these models against the standard kidney failure risk equation commonly used in general practice settings. Refining the predictions further by accounting for mortality as a competing risk will also be a key area of exploration.

7 Limitation

This study examines CKD progression over time, including patients who died during the observation period. While mortality may influence disease trajectories, our approach focuses on progression patterns independent of competing events. Future research could explore alternative modeling strategies that explicitly account for competing risks to provide a complementary perspective.

In addition to competing risks, an important limitation of this work is the lack of interpretability analysis. Techniques such as SHAP or LIME could offer insight into model decisions, and future work will explore these methods along with systematic error analysis. Further, as this study is limited to internal validation, future efforts will evaluate generalizability using an independent dataset. Lastly, while we frame CKD progression as a static classification problem, future research could incorporate time-series modeling or survival analysis to better capture disease dynamics.

Acknowledgments

This research utilised Queen Mary's DERI HPC cluster and Apocrita HPC facility, supported by QMUL Research-IT. It is funded by the National Institute for Health Research (NIHR) Newcastle Patient Safety Research Collaboration and NIHR 31672 AI-MULTIPLY.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Skylar Connor, Ting Li, Yanyan Qu, Ruth A Roberts, and Weida Tong. 2024. Generation of a drug-induced renal injury list to facilitate the development of new approach methodologies for nephrotoxicity. *Drug Discovery Today*, page 103938.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garabed Eknoyan, Norbert Lameire, Rashad Barsoum, Kai-Uwe Eckardt, Adeera Levin, Nathan Levin, Francesco Locatelli, Alison Macleod, Raymond Vanholder, Rowan Walker, et al. 2004. The burden of kidney disease: improving global outcomes. *Kidney international*, 66(4):1310–1314.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Bruce Guthrie, Colin McCowan, Peter Davey, Colin R Simpson, Tobias Dreischulte, and Karen Barnett. 2011. High risk prescribing in primary care patients particularly vulnerable to adverse drug events: cross sectional population database analysis in scottish general practice. *Bmj*, 342.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Mario A Isaza-Ruget, Nancy Yomayusa, Camilo A González, Fabio A de Oro V, Andrés Cely, Jossie Murcia, Abel Gonzalez-Velez, Adriana Robayo, Claudia C Colmenares-Mejía, Andrea Castillo, et al. 2024. Predicting chronic kidney disease progression with artificial intelligence. *BMC nephrology*, 25(1):148.

Meg J Jardine, Kenneth W Mahaffey, Bruce Neal, Rajiv Agarwal, George L Bakris, Barry M Brenner, Scott Bull, Christopher P Cannon, David M Charytan, Dick De Zeeuw, et al. 2018. The canagliflozin and renal endpoints in diabetes with established nephropathy clinical evaluation (credence) study rationale, design, and baseline characteristics. *American journal of nephrology*, 46(6):462–472.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Alberto Martínez-Castelao, José L Górriz-Teruel, Jordi Bover-Sanjuán, Julián Segura-de la Morena, Jesús Cebollada, Javier Escalada, Enric Esmatjes, Lorenzo Fácila, Javier Gamarra, et al. 2014. Consensus document for the detection and management of chronic kidney disease. *Nefrología (English Edition)*, 34(2):243–262.

David C Mendelssohn, Christine Malmberg, and Bassem Hamandi. 2009. An integrated review of" unplanned" dialysis initiation: reframing the terminology to" suboptimal" initiation. *BMC nephrology*, 10:1–8.

Bruce Neal, Vlado Perkovic, Dick de Zeeuw, Kenneth W Mahaffey, Greg Fulcher, Peter Stein, Mehul Desai, Wayne Shaw, Joel Jiang, Frank Vercruysse, et al. 2013. Rationale, design, and baseline characteristics of the canagliflozin cardiovascular assessment study (canvas)—a randomized placebo-controlled trial. *American heart journal*, 166(2):217–223.

Irene L Noronha, Guilherme P Santa-Catharina, Lucia Andrade, Venceslau A Coelho, Wilson Jacob-Filho, and Rosilene M Elias. 2022. Glomerular filtration in the aging population. *Frontiers in Medicine*, 9:769329.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Sandeep Reddy, Supriya Roy, Kay Weng Choy, Sourav Sharma, Karen M Dwyer, Chaitanya Manapragada, Zane Miller, Joy Cheon, and Bahareh Nakisa. 2024. Predicting chronic kidney disease progression using small pathology datasets and explainable machine learning models. *Computer Methods and Programs in Biomedicine Update*, 6:100160.

Hirotaka Saito, Hiroki Yoshimura, Kenichi Tanaka, Hiroshi Kimura, Kimio Watanabe, Masaharu Tsubokura, Hiroki Ejiri, Tianchen Zhao, Akihiko Ozaki, Sakumi Kazama, et al. 2024. Predicting ckd progression using time-series clustering and light gradient boosting machines. *Scientific Reports*, 14(1):1723.

Navdeep Tangri, Thomas W Ferguson, Ryan J Bamforth, Silvia J Leon, Clare Arnott, Kenneth W Mahaffey, Sradha Kotwal, Hiddo JL Heerspink, Vlado Perkovic, Robert A Fletcher, et al. 2024. Machine learning for prediction of chronic kidney disease progression: Validation of the klinrisk model in the canvas program and credence trial. *Diabetes, Obesity and Metabolism*.

X Yang, A Chen, N PourNejatian, HC Shin, KE Smith, C Parisien, C Compas, C Martin, AB Costa, MG Flores, et al. 2022. A large language model for electronic health records. npj digital medicine, 5 (1), 1–9.

Yitan Zhu, Dehua Bi, Milda Saunders, and Yuan Ji. 2023. Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. *Scientific Reports*, 13(1):22091.