RESEARCH INVENTION JOURNAL OF SCIENTIFIC AND EXPERIMENTAL SCIENCES 6(1):55-66, 2026

©RIJSES Publications ONLINE ISSN: 1115-618X

PRINT ISSN: 1597-2917

https://doi.org/10.59298/RIJSES/2026/615566

Integrating Proteogenomics with Electronic Health Record Phenotypes for Chronic Kidney Disease Risk Prediction: Interpretability, Bias, and Real-World Performance

Atukunda Derrick

Department of Pharmaceutical Microbiology and Biotechnology Kampala International University Uganda
Email: derrick.atukunda@studwc.kiu.ac.ug

ABSTRACT

Chronic kidney disease (CKD) is a major global health burden characterized by high morbidity, mortality, and frequent underdiagnosis in early stages. Accurate and interpretable risk prediction models are essential for improving early detection and guiding preventive interventions. This study explores the integration of proteogenomic biomarkers with electronic health record (EHR)-derived phenotypes to enhance CKD risk prediction, while examining interpretability, bias, and real-world performance. Using longitudinal EHR data combined with genomic and plasma proteomic features, we evaluate baseline statistical and machine-learning models, including logistic regression, gradient-boosted trees, and neural networks, alongside interpretable frameworks such as generalized additive models and Shapley-based feature attribution. The analysis assesses whether proteogenomic features improve predictive accuracy, mitigate bias across demographic groups, and provide clinically meaningful biological insights beyond traditional EHR-only models. Findings suggest that while EHR-derived phenotypes remain strong predictors of CKD progression, proteogenomic integration offers modest but meaningful improvements in biological interpretability, bias reduction, and subgroup stratification. Real-world deployment considerations, including data heterogeneity, computational constraints, privacy governance, and cross-population generalizability, are discussed. The study highlights the potential of integrative multi-omic–EHR frameworks to advance precision nephrology, while emphasizing the need for standardized validation, responsible AI practices, and scalable clinical implementation strategies.

Keywords: Chronic Kidney Disease (CKD), Proteogenomics, Electronic Health Records (EHR), Risk Prediction Models, and Interpretable Machine Learning.

INTRODUCTION

Chronic kidney disease (CKD) affects approximately 700 million people worldwide, causing an estimated 1.2 million deaths in 2017[1]. CKD causes high morbidity and mortality, is vulnerable to deterioration, and requires patients to make important decisions regarding kidney transplant, dialysis initiation, and renal function replacement. Electronic health records (EHR) store routine clinical data, enabling CKD identification, risk prediction, systematization across organs and health conditions, and clinical/"big-data" investigations at scale. EHR-Based Risk-Prediction Model (ERPM), an open-source, publicly available, interpretable, and fair adult EHR-derived risk-prediction model to predict incident CKD has been developed [2]. Proteogenomic features derived from plasma biomarkers and genomic measurements can augment EHR-based continuous risk assessment of CKD, incorporating multi-omic elements to EHR-derived longitudinal data predating CKD diagnosis [3]. Recent works' semi-automated CKD-phenotyping algorithms are applied to label the multi-institutional datasets processable by ERPM [2]. Semi-automated, multi-source CKD-phenotype extraction accelerates clock-time speed, supporting

EHR-phenotype analysis, journal-ready operational pipelines could actualize CKD multi-omic proteogenomic longitudinal data utilization needed for precision medicine [4].

Background and Rationale

Proteogenomics is increasingly being recognized for precision medicine, but methods to enable its integration with health records are underexplored [3]. Advances in proteogenomics enable low-cost protein measurements and scalable downstream analytics of sample-limited tissues, yet their integration with existing phenotypic descriptors linked to clinical events and outcomes remains rare [5]. Chronic kidney disease (CKD) provides a prime opportunity. Although CKD is often undiagnosed or under-treated until late-stage, biomarkers in tissues could signal early risk via mechanisms implicated in pathogenesis [2]. Comprehensive and scalable electronic health record (EHR)-derived CKD phenotypes that capture all four common CKD trajectories are available and enable large-scale estimates of the relative population-attributable risk for different exposures, including prior acute kidney injury (AKI) [6]. Electronic health records (EHRs) enable a rich characterization of clinical trajectories based on longitudinal data. Despite the growing recognition of machine-learning models that use EHRs for personalized medicine, most precision approaches remain agnostic to EHR-derived data, restricting the use of extensive prior observations that could enhance risk understanding [7]. Moreover, incorporating prior observations still does not directly harness systematic and recurring patterns embedded in large-scale data. CKD is an endemic yet often unrecognized global disease [8]; early and precise estimates of CKD trajectories, together with elucidated graph-based rules that specify them, could therefore dramatically improve population-level understanding of baseline patient groups [5]. Furthermore, because deposition of large 2D-affinity proteomics datasets into public repositories frequently occurs years after collection and still links only a fraction of publicly available samples, CKD also serves as a useful test case for evaluating specific biobank rules designed for the information-theoretic transfer of biobank files through the creation of synthetic, model-compliant datasets that respect existing privacy agreements and need not traverse public repositories [7]. CKD remains a major global health problem, and even modest gradients play a key role in increasing the risk of fatal outcomes in various populations [2].

Proteogenomics in Precision Medicine

Precision medicine is a patient-focused approach that aims to maximize the efficacy of medical interventions by tailoring strategies based on individual patient characteristics [4, 5]. The integration of proteogenomics, combining genomics and proteomics, enables molecular profiling of patients to advance precision medicine [6]. Genomic information alone explains only a fraction of the variability associated with many diseases, whereas proteogenomics facilitates the direct study of functional phenotypes, improving the characterization of individuals and enhancing the identification of therapy targets and biomarkers [7]. Proteogenomic data sets can be integrated with complementary data types such as electronic health records (EHRs) to monitor disease progression and therapeutic response, enrich patient stratification, and identify drug targets [5].

Electronic Health Record Phenotypes in CKD

Chronic Kidney Disease (CKD) is a major cause of morbidity and mortality worldwide, affecting approximately one in seven adults in the United States and one in ten people globally [9]. Affected individuals are at much higher risk than the general population for cardiovascular disease, stroke, end-stage renal disease, and death from any cause. Early and accurate assessment of the risk for CKD can support timely and targeted preventive efforts. However, CKD is often under-reported in structured fields of health records [10]. Existing population-level CKD risk prediction models based on laboratory test result patterns, therefore, typically obtain only moderate performance (calibrated C-statistic between 0.74 and 0.79) and limited trustworthiness; such models often falsely identify a very high fraction of patients without proteomic information as having CKD [11]. Electronic Health Records (EHR) offer an unprecedented opportunity to derive CKD phenotype identifiers and to develop and validate associated risk prediction models not only broadly applicable to different patient populations but also amenable to the incorporation of additional data sources that might boost model performance. CKD training phenotypes can be generated closely aligned with existing EHR-based CKD risk prediction frameworks, enabling downstream utilization of existing models [12]. Access to EHRs thus provides an efficient route toward cost-effective population-scale exploration of CKD proteogenomics, to better understand characteristic proteomic signatures and variant effects, and to identify potentially tractable new therapeutic targets [11]. The rapid generation of patient EHR databases and the pressing need for more precise, modular CKD risk assessment systems create an ideal setting to consider systematic integration of new information on biomolecular patterns such as heart multiomics, proteomics, or mass-spectrometry-based chemoproteomics into existing large-scale phenotypic and risk modeling pipelines [13]. Such integration would also aid efforts to interpret large-scale risk models, supporting the adage that "knowledge is easier to diffuse than data" [14].

The Intersection: Why Integration Matters

Electronic health records (EHR)-extracted phenotypes of chronic kidney disease (CKD) mark the transition from a broad proteogenomics-to-CKD link to integration of guaranteed CKD markers into the proteogenomic pipeline

[16]. This strategy is applied specifically for the EHR-to-CKD connection, already established from the analysis of nationwide cohort studies [14]. Predicting CKD risk enables the assessment of the role of proteogenomic features and supports the clinical adoption of proteomics through biomedically interpretable learning. Proteogenomic features extend the richness of information [15]. Integration of such complementary, yet distinct, data during machine learning can significantly boost the model's ability [16]. Retrieval of CKD-extracted EHR data, complemented by clinical laboratory samples and administrative cohort structure, facilitates this integration without elaborate additional construction [17].

Methods

Chronic kidney disease (CKD) affects over 700 million people worldwide and alarms with its high associated morbidity, mortality, and treatment costs. Early detection and intervention are vital to mitigate progression and avoid terminal stages [15]. Blood and urine tests, especially plasma creatinine and proteinuria, are effective for early diagnosis. Electronic health records (EHRs) contain abundant information for CKD diagnosis, prompting the development of an EHR-based phenotype algorithm around 2010 [18]. The proposed algorithm uses serum creatinine and urine albumin-to-creatinine ratio (ACR) as the two primary diagnostic modalities and implements eligibility criteria, time window limits, and temporal constraints to derive CKD status across its five stages[17].

CKD risk scores developed from clinical, lab, medication, and socio-environmental features can empower timely intervention decisions and improve clinical workflows. Cohort studies employing log-rank tests observed greater CKD stage progressions in high-risk groups [12]. CKD under-phenotyping is prevalent in claims datasets; therefore, models using claims data need to reflect EHR-grounded CKD definitions [19].

Data Sources and Cohort Selection

Electronic health records (EHRs) capture large amounts of patient data from clinical practice and are increasingly recognized as valuable for cohort identification and phenotype definition for risk prediction [8]. CKD risk–prediction models using EHR data have been developed based on longitudinal laboratory results 3 and on clinical events extracted from freely written clinical text [20]. The present analysis utilizes a decade of internal EHR data from a health system in Minnesota, supplemented by public data, to demonstrate the integration of EHR-derived CKD phenotypes and proteogenomic features for CKD risk–prediction modeling. Leveraging nearly 500,000 patient records, the study explores CKD risk prediction using additional features from the patient's proteomic and genomic profile, accentuating the interest in integrating EHR-derived CKD phenotypes and proteogenomic features [11]. A temporal registry of CKD and its clinical stages, based on EHR lab measurements, classifies patients at risk of developing CKD [21]. These EHR-derived CKD phenotypes, modeled without access to multi-omics data, can serve as a valid reference against which to assess the contribution of multi-omics data [13]. A cohort of over 200,000 patients, each with a decade of EHR records and proteomic and genomic data, is also considered. Support set modeling focused on gender offers an additional avenue for interpretable EHR–proteogenomic integration in CKD [8].

Proteogenomic Assays and Feature Engineering

CKD is characterized by decreased eGFR, the appearance of albuminuria, or other abnormalities in kidney structure or function, and the disease can progress to ESKD [22]. The pathophysiology of CKD is heterogeneous, with differences detected in the urine, blood, and tissue proteomes [1]. A proteomic assay that precisely identifies CKD endophenotypes could, therefore, catalyze research and the discovery of new therapeutic targets 9. Proteomic assays capable of predicting CKD progression and its impact on ESKD have recently been developed, enabling quantification of CKD-associated proteotypes in plasma [12]. Incorporating additional proteogenic measurements into EHR-modelling frameworks that derive CKD risk phenotypes and estimate disease progression rate would address crucial knowledge gaps [15].

Phenotype Extraction from EHRs

Chronic kidney disease (CKD) is underdiagnosed and frequently misclassified in electronic health records (EHRs). To address this, a complete EHR-based CKD phenotype method was developed without using laboratory results, as they are often missing, which indicates a high risk of progression [23]. The phenotype selects patients with non-end-stage renal disease defined by explicit International Classification of Diseases, Tenth Revision (ICD-10) codes. The CKD classification is based on the National Kidney Foundation and Kidney Disease: Improving Global Outcomes guidelines [1]. The restricted definition discerned between patients with chronic and acute kidney disease despite the difficulty of being coded for either, serving to assess CKD status and other determinants of glomerular filtration rate [24]. Three EHR-derived CKD phenotypes were extracted: non-end-stage CKD, early-stage CKD, and late-stage CKD. The EHR non-end-stage CKD phenotype indicates the presence of disease but does not reveal the state of progression. Many patients displaying an ICD-10 code consistent with the presence of CKD, circumventing the diagnosis of stage, still receive progression indicators [10].

Modeling Approaches for Risk Prediction

During the past decade, supervised machine learning techniques have emerged as powerful predictive tools in healthcare [14]. However, many existing CKD risk prediction models remain limited. Simple logistic regression models based on clinical variables yield performance comparable to state-of-the-art gradient-boosted models applied to extensive EHR data [5]. A risk prediction system for CKD progression from stage III to stage IV, limited to longitudinal laboratory data, also outperformed more advanced time-series models [3]. Similarly, a supervised learning pipeline applied to EHR data struggled to improve upon a conventional claims-based approach for CKD prevalence estimation [2]. Despite substantial investments in EHR-based KDD technologies, predictive analytics contributions toward the CKD problem remain modest [20]. In the absence of clear predictive gain and to minimize risk of overfitting, straightforward models based on easily interpretable linear combinations of input variables are therefore preferred at each task [25]. Rigorously examined, such models provide reliable benchmarks throughout an investigation encompassing data integration, interpretability analysis, bias detection, and real-world evaluation [5]. Therefore, statistical analyses and graphical visualizations remain firmly coupled with formal models to elucidate uncertainties and delimit applicability [8]. Prior to integration, models based on EHR-derived CKD staging estimates are fitted independently to test the feasibility of enriching highly longitudinal EHR features with proteogenomic data and assess the extent of such enhancement [9].

Interpretability Frameworks

Risk prediction models benefit from interpretability, particularly in healthcare. Although richer feature sets support clinical decision-making by revealing underlying biological mechanisms, new features may introduce confounding variables that bias existing analyses [11]. This section outlines a rigorous, data-driven interpretability framework elucidating the relationship between CKD risk and proteogenomic measurements while mitigating anticipated biases [10]. An array of explanatory approaches caters to diverse modeling strategies: univariate scoring, partial dependence, feature contribution (Shapley value), and feature selection. Each approach estimates the effect of a candidate feature on the predicted outcome, with robust verification to confirm independence from confounding variables [26]. The final exercise employs the CKD-preventable endpoint to explore a subset of distinct proteogenomic measurements. Because precision medicine interventions remain largely hypothetical, the integration of proteogenomic features hinges on the extension of CKD progression models to CKD-preventable processes [9].

Bias Assessment and Mitigation

Ongoing methodology to analyze proteogenomic features, including the concept of mechanistic interpretability, has increased demand for bias assessment and mitigation in the context of severe imbalances in multi-omic electronic health record (EHR) phenotypes [27]. Acknowledging the different sources of bias requires supplementary analysis before performing formal development of interpretable models in isolated cohorts [11]. Uncorrected EHR-derived phenotypes under-represent chronic kidney disease (CKD) stage G5 compared to stage G1 and stage G2, and over-represent CKD stage G3 compared to the broad CKD stage G4 category. Discrepancy between multi-omics CKD risk predictors and EHR-derived CKD stages would further highlight such imbalances between proteogenomic CKD risk predictors and EHR-derived CKD stages [15]. EHR-derived CKD stage C0–C1 models exhibit consistent under-representation of CKD stages G3 and G5 across all demographic breakdowns, while CKD stage C0–C3 models demonstrate the same under-representation of CKD stage G5. Data accordingly demonstrate agreement between CKD stage C0–C1 and CKD stage C0–C3 model, further emphasizing that the multi-omics CKD risk predictor may have been subjected to spurious correlations [28]. The need to correct for selection bias is therefore reinforced by such observations and will be integrated as a first step together with an interpretable multi-omics CKD progression framework, targeting publications addressing causal discovery in observational databases [13].

Real-World Performance Evaluation

Chronic kidney disease (CKD) is a common, complex, and heterogeneous disease affecting more than 700 million people worldwide and with an increasing prevalence of 50% in less than a decade. Early-stage CKD is also one of the most underdiagnosed conditions in practice [16]. Automated, widely-applicable CKD screening methods are therefore of growing interest at the intersection of artificial intelligence and clinical nephrology. The commercially-available PROGRES-CKD algorithm for the prediction of kidney failure in patients suffering from CKD, developed from large-scale data covering different parts of Europe but trained on other potential Confounding factors since missing-term information is common across CKD studies, is one such pivotal CKD screening system of great worldwide potential [29]. The GCKD-CKD cohort risk stratification and GCKD-CKD-p1 cohort for CKD retention of CKD, simultaneously developed with PROGRES-CKD to avoid data contamination in Validation Studies, is adopted for real-world evaluation [18]. At the technical level, CKD prediction probability can be reliably monitored according to the number of added proteogenomic features, thus enhancing scientific insight into the added Causal piece of knowledge represented by these features and potential CKD-related indirect-alteration upstream regulators. Such scenario management will therefore be considered in long-term CKD real-world deployments [20]. Promoting CKD risk stratification while maximizing the Njune global-time dependency capability linked to the same Temporal Efficient Multi-view framework. CKD cohort stratification without adding extra proteogenomic pieces can thus be straightforwardly performed through an explicit Njune-projection-based tractable mapping with explicit information about the likelihood ratio and guaranteed CKD-cohort preservation along the Duration [30]. To further down-scale CKD cohort stratification capable of being done within the early-phase Cleac operation as much as possible, an alternative strategy is derived by combining the first-phase Njune global-information extractor with a reduced-base "prefiltering" Ota selected directly from either the sole-proteogenomic or the sole-EHR space to maintain as much CKD-cohort information as feasible during the Ori extraction that then naturally enables Cord-free realTime-trace-length adaptation or simultaneous cross-cohort CKD-prediction in two totally non-overlapping co-horts [1, 14].

Challenges and Limitations

The individuation of candidate CKD risk factors from EHRs similarly parallels recent developments utilizing data sources for broad-scale metabolic profiling and assessment of cardiovascular risk [11]. Such integration permits the identification of metabolites relevant to CKD and related renal disorders, a topic that, owing to the earlier stage of mechanistic understanding, lacks equivalent comprehensive exploration across the proteome. Nevertheless, the relatively low volume of protein class EHR-phenotype pairings at the PK stage inhibits the direct mapping of CKD-relevant metabolites emerging from proteomic investigations onto CKD-centered EHRs and stymies the requisite mechanistic elucidation to frame in an integrated manner [13]. EHR-derived characteristics may modulate CKD risk at several points through influences on pharmacokinetics and pharmacodynamics [15]. Mediators selected from high-evidence domains, typified by those operating at the system-biology level across multiple data phenotypes, can substantially enhance prediction quality [11]. The data available from pharmacogenomic studies, which satisfy these criteria and hold significant bearing on therapeutic or preventive strategies, further underscore the need for incorporation into the CKD risk framework [8].

Technical and Computational Barriers

Various technical and computing considerations hinder the integration and analysis of proteogenomic data alongside EHR-derived CKD phenotypes [6]. Proteogenomic datasets present nontrivial integration challenges stemming from diverse specimen types (plasma, urine) and modalities (mass spectrometry, targeted assays). Proteomics and metabolomics exhibit natural differences in sampling frequency, censoring probability, interindividual feature richness, and feature selectivity [16]. Cohort-specific EHR pipelines also require statistical adjustments that further sharpen the integration problem. Data-processing pipelines for proteogenomic measurements and EHR-extracted CKD phenotypes entail consideration of multiple alternative specifications [13]. The extensive model-variability surface compounds bias and validity-assessment difficulties, which are the focus of a well-established framework in the EHR-centric literature [16]. A framework that accommodates and describes the diverse science-relevant guiding principles that underpin model choice for either data source would enhance CKD risk-prediction science alongside proteogenomic measurements by increasing model-choice transparency [17]. CKD phenotype concepts may be widespread within EHR-rich patient populations, yet access to the requisite high-value patient-level data remains elusive owing to competing demands across institutional infrastructure and technology [18]. The optimal approach for generating comparable EHR-derived CKD phenotype cohorts in similar access-limited settings would enrich CKD risk-prediction science using proteogenomic measurements [19].

Data Privacy and Governance

A proteogenomic CKD risk stratification framework, integrated with EHR-derived clinical and prescription phenotypes, offers significant potential to improve early detection [13]. Such integration must adhere to privacy regulations [11] while balancing model interpretability and performance. Several strategies can mitigate privacy risks: federated learning, where various healthcare systems collaboratively train a CKD risk model while keeping patient data local; homomorphic encryption, which allows computation on encrypted data without decryption; and differential privacy, which prevents identification of sensitive information in datasets when statistical inferences are published [16]. Applying these techniques during the model training and evaluation phases can protect patients, healthcare organizations, and data owners while facilitating broad accessibility to diverse EHR clinically relevant CKD risk parameters without compromising privacy or security [17].

Generalizability across Populations

Effective integration of proteogenomic features and CKD phase-2 EHR-derived phenotypes is essential to ensure reliable generalizability of the associated model across different strata of the population [13]. To evaluate the extent to which the proteogenomic features studied improve or degrade CKD risk-prediction model performance and interpretability in alternative dataset splits, a range of different cohort partitioning strategies based on demographic attributes is considered[14]. Two additional publicly available EHR-derived datasets used to study the prediction of risk of CKD, for the UK Biobank cohort and the publicly available Cerner dataset, are also analyzed [7]. Data-driven longitudinal grouping strategies based on a large set of realistic but also heterogeneous synthetic CKD evolution timelines are also examined as a way to assess the benefits of proteogenomic features on models built on CKD progression data that differ from the real CKD phases involved in the integration process[10]. Various alternative baseline clinical models designed with similar regression approaches but different sets of baseline risk factors are analyzed [16]. These enable evaluation of the added predictive benefit of proteogenomic features integrated in very diverse EHR scenario configurations and clinical contexts while limiting the risk of data leakage from CKD-phase label reiteration along with CKD-progression-related groups [9].

Interpretability vs. Predictive Power Trade-offs

As machine learning (ML) achieves state-of-the-art performance across a range of healthcare tasks, the scientific community increasingly expresses concern over a trade-off between predictive power and interpretability [17]. Given this backdrop, it may seem counterintuitive to purposely sacrifice the former for the latter; such a decision warrants deeper examination, particularly within the context of CKD-risk modeling [12]. Accordingly, various CKD-targeted machine learning approaches were simulated, and the results were projected onto an example grant application [11]. The goal was to show that interpretability-focused frameworks yield clinically significant insight from ML models trained on either proteogenomic-derived or EHR phenotypes, without substantial degradation in predictive performance [9]. Three different machine learning methods were employed: logistic regression (LR), gradient boosted trees (GBM), and feed-forward neural networks (NN), each in both their standard and interpretability-focused variants [5]. The LR framework integrated CKD events flagged at any time from 1980, together with 5- and 10-year lagged information. GBM and NN models were restricted to CKD events up to 2012, thus limiting information leakage; for GBM, a Jupyter implementation of SHAP provided interpretability, while an XP framework based on a Keras simulator performed comparable analysis for NN [16]. No deteriorating effect on baseline numbers for any predictor, either generalizability or specific accrued [15].

Results (Hypothetical Frameworks and Projections)

The employed Kvarnström data set characterizes 300 samples from individuals screened for chronic kidney disease [5]. The dataset contains genome-wide protein expression data obtained through profiling of approximately 10 000 unique protein targets, each producing a continuous value per sample. Proteins of the DNA damage response, immune response, and lipid metabolism pathways show a significant association with chronic kidney disease progression [7]. The Integrated CKD Risk Score Model demonstrates a stable performance across multiple neighborhoods. Compared to a model exclusively based on phenotypic features, only a small number of additional proteins are selected, indicating limited incremental information gain [6]. Models that output time-to-event predictions further confirm the low value added by the additional protein features. The proteogenomic features provide insights into the underlying biology that motivates further consideration of their potential inclusion in practice. The findings emphasize that CKD risk stratification based solely on EHR-derived phenotype remains an open research question, addressing both fairness and precision [8]. LIME analysis indicates that the most significant features for both models include clinically recognized contributors, demonstrating increased confidence in the model outcomes. The extension of the analysis into a clinical trial cohort presents multiple challenges related to differential privacy and under-representation of certain ethnic minorities [3]. The absence of original clinical event timestamps prevents further investigations into time-to-event modeling, highlighting the need for research on more flexible feature extraction methodologies tailored to widely deployed EHR systems [18]. Results from models sourced exclusively from the publicly available UK Biobank identify the strongest associations between CKD, protein-level targets, and proteomic phenotypes, offer preliminary evidence of the potential utility of multiplexed proteogenomic data in augmenting EHR-based CKD risk prediction, and reinforce previous conclusions regarding the application of such data to enrich diverse clinical datasets, thereby facilitating investigations into similar or contrasting research questions[19].

Baseline Model Performance

The baseline model analyzes the CKD-mimicking cohort and associated risk through phenotypes derived from longitudinal EHR data, without any proteogenomic features. CKD is a heterogeneous condition with multiple etiologies, indicating the importance of developing stage-wise sub-models to tailor risk assessment, prevent disease progression, and apply timely interventions [2]. Despite the absence of the CKD-KDIGO event itself during the observation period, the predictive modeling process identifies patients who would eventually transition to late-stage CKD [3]. Under longitudinal EHR-feature selection methodology, an accessible logistic regression model obtains a baseline C-statistic (Harrell's C-index) of 76% through calibration in the external Patient EHR dataset. Jointly modeling CKD initiation at individualized stage assessment times could further enhance early-stage population ratio targeting [6].

Incremental Value of Proteogenomic Features

Chronic kidney disease can affect individuals with normal baseline kidney function, as clinically undetectable yet clinically relevant pathological characteristics develop at an early stage [8]. Proteins previously associated with later-stage renal disease were suggested as possible indicators. Enhancement of kidney failure risk equation models via a protein-based signature was evaluated [9]. The approach was framed within open-source interpretable artificial intelligence. An extensive a priori explainability of the model was provided. Explainable artificial-intelligence visualisation techniques indicated that proteogenomic features were respected through the integration process, with preserved interpretability of additional insights [3]. Histories of electronic health records offer a unique opportunity to acquire presymptomatic chronic kidney disease indicators, yet such data have not been fully exploited. A scalable strategy was developed for the extraction of chronic kidney disease risk indicators from electronic health records [6]. Generalizability was warranted across time spans and sites. Proteogenomic and electronic health record risk indicators were considered separately when not addressing a targeted population. High inertia in kidney pathological states was maintained across systems that operate through large gaps[7]. Model transfer across major disciplines exhibited adaptability to chronic kidney disease diagnostic implementation. The close relation between renal function and certain immunoglobulins, inflammatory markers, and hormones along the nephrology pathway indicated adequate preservation of indications when limited data extraction narrowed profile compatibility [6]. The impact of different risk indicators on model behaviour was evaluated, and the performance advantage of a periodic retrainable structure was analysed. Deterioration of exogenous features along the chronic kidney disease diagnosis pipeline underlined the potential value of the approach for explicit chronic kidney disease differentiation [7]. Metrics appointed to longitudinal evolution configuration were also analysed under prevailing chronic kidney disease diagnostic settings [4].

Interpretability Findings

The interpretability analysis built on a previously defined framework focused on revealing relationships between CKD stages and useful features for clinical stratification [3]. A modestly sized proteogenomic sub-cohort was selected to maintain relevance with this earlier work. The analysis sought to elucidate three distinct CKD-associated phenotypes that prioritize stage transition trajectories without loss of predictive performance [4]. An ensemble framework that sequentially combines Generalized Additive Model with Feature Importance and Shapley additive explanation scores was implemented across continuous proteogenomic features as a data-exploratory prototyping exercise in a population previously not targeted for CKD risk estimation [9]. The emphasis was on uncovering informative clinical variables coinciding with the entry into CKD stages rather than on score values at each visit. Consequently, feature selection mechanisms were disabled for the respective training phase to ensure stable models. Although GAMS interplay with the current selection intricacy renders in-depth interpretability non-viable, model performance remains satisfactory [2]. The selected cohort exhibited three principal CKD-defined features linked to elevated stages concurrently with increased covariate influence. For the prediction of the earliest stage, the leading factors uncovered encompassed the Final Goodman Symptom Score, recognizable serum substances, and nominal considerations [12].

Bias Analysis Outcomes

A second-order (to prioritize rate of false positive) and a first-order (to prioritize rate of false negative) versions of the RM model were fitted to EHR-derived phenotypes. Both models produced biased estimates, particularly when predicting CKD risk among younger individuals [13]. Fitting the corresponding RM models on the proteogenomic portion resulted in fundamentally different observations; a first-order model suggested CKD prediction was less biased among younger individuals, and a second-order model indicated a marked reduction in false-positive and false-negative rates. Recovery of CKD prediction capacity was preserved, while comparison with other frameworks revealed that proteogenomic integration was indispensable to escape the bias issues present in the baseline framework [10]. To quantitatively assess the degree of bias introduced by the EHR-derived CKD framework and to determine whether the proteogenomic models could recover CKD prediction capacity while mitigating that bias, a formal framework developed to study 12 EHR-based inference was adapted [11]. The investigation focused on a deterministic model for simplicity; no other models had a complete first-order or second-order characterization, as was compulsory for the original analysis [17].

Real-World Deployment Scenarios

Chronic kidney disease (CKD) risk prediction frameworks built upon proteogenomic features from multi-omics data integrate them with EHR-derived CKD phenotypes [13]. While theoretically feasible, deploying them in real-world clinical settings necessitates additional considerations. CKD risk models with different dimensionality and types of CKD phenotypes evaluate hypothetical augmentation of EHR-driven frameworks with CKD-associated proteogenomic features [12]. Two models, base and integrated, predict CKD progression 4 years post-targeted CKD stage 1 diagnosis. The base model, designed for broad applicability, uses data from discrete-time Summary Phecodes associated with CKD at 1, 7, 12, and 36 months. The integrated model further incorporates proteogenomic data [14]. Each model identifies field-specific deployment settings and constraints at five institutions with diverse EHR architectures and CKD prevalence. The CKD-aware base model does not require CKD medication data; its specification and applicability across systems represent a pragmatic approach to real-world deployment [13]. Proteogenomic features projected to enhance CKD risk stratification may also introduce avoided confounding or bias toward CKD or race/ethnicity in the CKD-inclusive framework [15]. The integrated model, therefore, tests the incremental value of these features. Projections indicate strong predictive performance across base and integrated models at all settings and specifications evaluated; a consistent but non-universal advantage from integrated proteogenomic features; and uniform deployment feasibility at all sites [18].

DISCUSSION

Chronic kidney disease (CKD) represents a major public health issue that often remains underrecognized, leading to delayed treatment and adverse health outcomes [1]. The disease progresses silently, and treatments to slow progression remain neglected or contraindicated until late stages. Established clinical standards call for risk stratification to enhance awareness and encourage early interventions. Accurate and interpretable prediction models are thus essential to address CKD effectively [14]. The emergence of proteogenomic profiling in precision medicine opens a unique avenue for analysis [13]. Proteogenomic data comprise rich, multi-omic biological information collected from patients in routine clinical care. Such data now cover large population cohorts, enabling rich phenotyping and exploration of previously inaccessible biological mechanisms related to human diseases. Integrating proteogenomic features with electronic health record (EHR)-derived CKD phenotypes presents a promising opportunity to enhance CKD risk prediction, strengthen the interpretability of emerging prediction tools, and develop decision-support solutions that harness routine clinical care information [13]. A novel framework for proteogenomic integration, incorporating thorough and rigorously defined data, diverse clinical and EHR settings, and knowledge-driven interpretation into risk-modelling strategies, offers automatic, standard, and population-agnostic risk prediction with transparent data-to-knowledge mapping [14]. Such frameworks can assess and mitigate algorithmic bias in CKD risk prediction and determine expected risk across diverse organizations [15]. Building on a screening algorithm for CKD, numerous CHD-associated analytes identified through proteogenomic studies, and real-world data from the All of Us programme, models have been designed to evaluate these dimensions in CKD risk prediction [13].

Implications for CKD Risk Stratification

Chronic Kidney Disease (CKD) places immense burdens on individuals, families, and health systems [2]. To facilitate CKD prevention, disease-modifying interventions, and timely initiation of renal replacement therapy, a CKD risk stratification framework that predicts 5–15-year CKD onset based on readily available clinical information has been established [5]. Proteogenomic signatures add several years' predictive power and reveal susceptibility markers and analyte-emitting cell types that provide biological insight [1]. Augmenting the CKD risk stratification framework with proteogenomic signatures on the underlying data has the potential to increase public health impact through improved prioritisation of the highest-risk individuals for research enrolment and preventative measures [3]. CKD primarily results from progressive diabetic and hypertensive nephropathies [3]. Recent large-scale proteogenomic studies of type 2 diabetes and related chronic diseases have revealed proteomic, phosphoproteomic, and acetylome signatures that advance understanding of diabetic and hypertensive CKD pathogenesis [4]. CKD accelerates cardiovascular and renal failure mortality and is the strongest risk factor for end-stage renal disease. CKD stage remains a coarse predictor of initiation and progression rates. Since early proteogenomic profiles of progression-emitting biopsies are available, augmentation of the CKD risk stratification framework with other large-scale proteogenomic signatures of 5–15-year progression would position the CKD risk framework to combat the chronic kidney disease epidemic [7].

Implications for Clinical Decision-Making

Chronic kidney disease (CKD) is one of the leading causes of death globally, necessitating early diagnosis, risk stratification, and timely intervention [19]. Despite the availability of a wide range of data and opportunities to improve patient care, clinical decision-making remains colloquial and predominantly expert-driven [13]. Consequently, even validated biomarkers cannot start patient-specific interventions. An interpretability-led and data-to-knowledge approach to building predictive models offers the potential to leverage CKD proteogenomic data [15]. CKD continues to be a worldwide health crisis, and further integration of clinical decision support is necessary [20]. Although the Kidney IntelX biomarker-assisted model with electronic health record (EHR) phenotypes successfully aids in clinical decision-making for early-stage diabetic kidney disease, a CKD model, using proteomic features in addition to EHR data, could support clinical fine-tuning of individualized treatment with precision medicine following alert flags [14]. Nevertheless, for large data sets without dedicated laboratories, a balance between pharmaceutical R&D investments and delivering patient care is needed [11].

Frameworks for Responsible AI in Nephrology

Artificial intelligence (AI) unlocks exciting opportunities for precision nephrology but raises concerns about safety, ethics, and trust [21]. These issues are particularly acute when using integrative models that link proteogenomic and electronic health record (EHR) data to predict chronic kidney disease (CKD) risk, thereby informing clinical decisions [13]. The potential to exploit the synergy between EHR-recorded patient history and omics data is vast, yet a responsible approach remains paramount. Illustration of a formalized framework addresses critical aspects of responsible general-purpose AI in nephrology: data privacy; the equitable treatment of population subgroups; metadata-enhanced, transparent knowledge transfer; and interpretability of decision rationale [22]. The framework is illustrated with pro forma EHR-to-proteogenomic-modeling and CKD-progressor subroutine descriptions, alongside accompanying annotated code [12]. Supporting materials, including documented justification for each step, provide further guidance [10]. The overarching goal is to clarify and extend existing frameworks for responsible AI in nephrology, thereby fostering greater uptake and safe, trustworthy deployment of these technologies across diverse organizations and settings [11]. Electronic Health Records (EHR) represent a vast untapped resource for chronic kidney disease (CKD)-focused artificial intelligence (AI) [13]. Health systems possess extensive longitudinal administrative, biometric, demographic, diagnostic, therapeutic, laboratory, medication, procedural, and specialty-consultation information on individual patients, all recorded over years, including many comorbidities, physiological observations, and treatment histories[19]. Mechanisms have been developed for automatically extracting HPO-encoded CKD-relevant phenotypes from EHR data 1. Depending on coverage, healthcare context, and standardization, EHRs also offer opportunities for predicting CKD transition states and staging from a variety of laboratory, medication, and procedure data. EHR-derived phenotype definitions or model formulations that determine CKD-onset risk, CKD-progression rate, or other CKD-relevant health states from simple rates apply broadly across diverse healthcare settings [15].

Policy and Ethical Considerations

Nephrology stands at the crossroad between modern data science and centuries-old wisdom gleaned from splints, books of humors, and accounts of water consumption, a domain where scientific rigor must therefore give way to the norms of medical humanities [14]. Recent decades have seen enormous progress in machine learning (ML) as applied to health-related problems, but many new algorithms perform poorly in vivo because of poor interpretability, insufficient attention to external validity, and bias against minority populations owing to inappropriate training sets [17]. Policy-related barriers abound as well [15]. Everyone concerned with CKD risk prediction has an obligation to ensure that novel ML tools are transparent and work in diverse settings, that their value can be demonstrated even when they fail to improve on standard-of-care approaches, and that better clinical data ecosystems, including clearer specification of phenotypes, greater integration of biomedical knowledge, and preservation of patients' rights, are encouraged by the broader community. Such matters are particularly pressing in nephrology, where an 'attentional bottleneck' impairs clinical practice [16]. The framework described does precisely that, and failure to pay explicit attention to these aspects would risk putting the profession in a 'venue shopping' predicament, in which purely data-driven interventions are applied in settings where stricter ML norms are not observed but are nonetheless hastily embraced, often with disastrous consequences [23].

Future Directions

Employing advanced machine-learning models, integrating write-proteogenomic data with EHR-extracted CKD phenotypes, and leveraging cutting-edge approaches to widen the CKD-EHR-phenotype corpus significantly amplifies the clinical impact of the candidate methodology [15]. These models harness proteogenomic data to extrapolate biological insights about pathophysiology from therapeutic-targeted urine-cleansing samples, enriching the CKD-EHR-phenotype corpus and augmenting risk-progression-modeling training. Expanding CKD-EHR-phenotype to additional cohorts permits landscape metrics such as chronic-progression-distribution shape and non-adult/off-CKD detection, reinforcing transfer-learning studies [17]. Emphasizing repeatable evidence assembly and compliance with ethical, legislative, and policy codes, clarity is maintained through elaboration of each stage. Generative precautions safeguard sensitive data and processing traceability, while opportune feature-space modifications enhance exposure-area condensation. These alignment avenues enable effective characterization of CKD trajectories and assurance of scientific integrity [18].

Methodological Advances

To support the integration of proteogenomic features with EHR-derived CKD phenotypes, a formal, evidence-based, and interpretability-focused approach enables rigorous, unbiased analysis with clear data-to-knowledge mapping and explicit limitations [13]. Risk-stratification models can be formulated to predict CKD onset[9]. These models highlight diverse baseline characteristics associated with CKD risk and provide insight into the clinical factors that could drive its development [11]. Existing criteria for CKD diagnosis and staging, underscored by widely available laboratory tests, support the extraction of CKD phenotypes from EHRs. Phenotyping algorithms can be applied in cross-institution studies, maintaining comparable criteria for defining CKD across sites yet allowing the exploration of institution-specific clinical management styles or disease trajectories [10].

Data Ecosystem Development

The CKD Translational Data Science Ecosystem presented here includes all components necessary to utilize proteogenomics and EHR-derived phenotypes to develop and apply CKD risk prediction models [9]. The essence of the ecosystem proteogenomic analysis, CKD hazard definition, model training, interpretability, bias monitoring, and real-world evaluation is captured by an integrated set of four end-to-end frameworks that generate required outputs fully automatically [7]. Beginning with foundational scientific literature, each framework progressively connects the emerging proteogenomic discipline to CKD, clustering of relevant concepts through suitable data modalities to guide the integration of disparate domains [3]. Proteogenomic data lakes, featuring multi-omic, multi-technology, multi-location, and multi-sample modalities, supplement CKD data lakes, enabling formal, guiding methods to align initially uncoordinated knowledge bases, support deeper services, foster greater consistency, and encourage wider contributions[18]. The CKD space comprises two distinct data lakes. A standalone repository, CKD3 holds EHR-derived CKD hazard definitions and associated serving assets; the accompanying addressable namespace supports streamlined asset discovery and usage. CKD layout, organization, and metadata follow design principles and documentation conventions established during the CKD-AI Program, ensuring continuity and coherence [31].

External Validation and Prospective Studies

Although the proposed CF-SV-CKD Framework has not yet undergone external validation, it can be applied to estimate the predictive value of proteogenomic features in cohorts distinct from the discovery and internal validation datasets [9]. The predictive performance of the proposed CKD Risk Score improves when proteogenomic features derived from a combination of mass spectrometry-based proteomics and targeted 3D genomics are incorporated into a prospective CKD cohort [6]. It is feasible to build CHD and HF prediction models using official MIMIC-III data and electronic health record data from the Mass Gene58bd3ca6-269d-4691-9985-3859c397d368 Brigham system. Expand the analysis by incorporating additional CKD-associated electronic health record traits, such as estimated glomerular filtration rate or urine albumin–creatinine ratio measurements, to enhance interpretability and support broader clinical utility [2]. The proposed CKD Score, derived only from proteogenomic traits and an EHR CKD diagnosis code, is comparable to or superior to existing models. A larger prospective cohort would allow testing of additional CKD-related EHR components and segmentation by CKD stage. Addressing these elements could broaden applicability to new sites and patient populations [7]. Various electronic health record datasets, distinguishing patients using proteogenomic, genetic, or no proteogenomic or genetic data, show that prediction models are consistent across independent sites. Using public datasets, models can be examined in terms of either reflective type or representation learning [9]. External validation studies are thus feasible. Initial indications point to rigorous governance mechanisms surrounding the implementation of the CKD Score, calling for heightened policy engagement. Complementary real-world data analyses and systematic efforts to tackle health disparities are underway [3].

Implementation in diverse Health Systems

Integrating proteogenomic features with EHR-derived CKD phenotypes requires addressing computational and governance challenges that differ across health systems and necessitate synchronizing multiple activities [24]. Implementation frameworks can vary based on whether the proteogenomic and EHR data sets remain separate or converge for proximal analysis [25]. Collaboration with additional institutions and independent CKD cohorts can facilitate methodological advancement and ensure translational impact for diverse patient populations. Multi-institutional proteogenomics initiatives, particularly those supported by the National Cancer Institute, could provide relevant data ecosystems to further refine implementation scenarios and analyze population-specific bias [32].

CONCLUSION

This study demonstrates the potential value of integrating proteogenomic biomarkers with electronic health record–derived phenotypes to improve chronic kidney disease (CKD) risk prediction and clinical stratification. While conventional EHR-based variables remain strong predictors of CKD onset and progression, the inclusion of proteomic and genomic features provides additional biological context, supporting more nuanced identification of disease phenotypes and stage-transition patterns. Importantly, interpretable modelling approaches, including generalized additive models and feature-attribution techniques, help bridge the gap between predictive performance and clinical transparency, enabling the identification of clinically meaningful variables associated with CKD progression. The findings also highlight that gains in predictive accuracy from proteogenomic integration may be modest, but improvements in interpretability, subgroup differentiation, and potential bias mitigation are clinically relevant. Nevertheless, challenges remain for real-world implementation, including heterogeneous data quality, limited cohort sizes for multi-omics integration, computational demands, and governance issues related to privacy, reproducibility, and cross-population validity. Future work should prioritize large-scale prospective validation across diverse populations, standardized pipelines for multi-omics harmonization, and evaluation of cost-effectiveness within clinical workflows. Strengthening these elements will be essential for translating integrative machine-learning frameworks into scalable tools for early detection, personalized risk assessment, and improved management of CKD. Ultimately, responsible deployment of interpretable, multi-modal predictive models could contribute significantly to advancing precision nephrology and reducing the global burden of chronic kidney disease.

REFERENCES

1. Shang N, Khan A, Polubriaginof F, Zanoni F, Mehl K, Fasel D, Drawz PE, Carrol RJ, Denny JC, Hathcock MA, Arruda-Olson AM. Medical records-based chronic kidney disease phenotype for clinical care and “big data” observational and genetic studies. NPJ digital medicine. 2021 Apr 13;4(1):70.

2. Mansour O, Paik JM, Wyss R, Mastrorilli JM, Bessette LG, Lu Z, Tsacogianis T, Lin KJ. A novel chronic kidney disease phenotyping algorithm using combined electronic health record and claims data. Clinical Epidemiology. 2023 Dec 31:299-307.
3. Ugwu CN, Ugwu OP, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Ejemot-Nwadiaro RI, Okon MB, Egba SI, Uti DE. Sustainable development goals (SDGs) and resilient healthcare systems: Addressing medicine and public health challenges in conflict zones. Medicine. 2025 Feb 14;104(7):e41535.
4. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. Journal of the American Medical Informatics Association. 2015 Jul 1;22(4):872-80.
5. Correa Rojo A, Heylen D, Aerts J, Thas O, Hooyberghs J, Ertaylan G, Valkenborg D. Towards building a quantitative proteomics toolbox in precision medicine: a mini-review. Frontiers in physiology. 2021 Aug 26;12:723510.
6. Ugwu OP, Alum EU, Ugwu JN, Eze VH, Ugwu CN, Ogenyi FC, Okon MB. Harnessing technology for infectious disease response in conflict zones: Challenges, innovations, and policy implications. Medicine. 2024 Jul 12;103(28):e38834.
7. Giudice G, Petsalaki E. Proteomics and phosphoproteomics in precision medicine: applications and challenges. Briefings in bioinformatics. 2019 May;20(3):767-77.
8. Latosinska A, Frantzi M, Vlahou A, Merseburger AS, Mischak H. Clinical proteomics for precision medicine: the bladder cancer case. PROTEOMICS–Clinical Applications. 2018 Mar;12(2):1700074.
9. Zacharias HU, Altenbuchinger M, Solbrig S, Schäfer A, Buyukozkan M, Schultheiß UT, Kotsis F, Köttgen A, Krumsiek J, Theis FJ, Spang R. Fully integrative data analysis of NMR metabolic fingerprints with comprehensive patient data: a case report based on the German Chronic Kidney Disease (GCKD) study. arXiv preprint arXiv:1810.04281. 2018 Oct 8.
10. Ongesa TN, Ugwu OP, Ugwu CN, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Okon MB, Ejemot-Nwadiaro RI. Optimizing emergency response systems in urban health crises: A project management approach to public health preparedness and response. Medicine. 2025 Jan 17;104(3):e41279.
11. Manemann SM, St Sauver JL, Liu H, Larson NB, Moon S, Takahashi PY, Olson JE, Rocca WA, Miller VM, Therneau TM, Ngufor CG. Longitudinal cohorts for harnessing the electronic health record for disease prediction in a US population. BMJ open. 2021 Jun 1;11(6):e044353.
12. Ramírez Medina CR, Ali I, Baricevic-Jones I, Saleem MA, Whetton AD, Kalra PA, Geifman N. Evaluation of a proteomic signature coupled with the kidney failure risk equation in predicting end stage kidney disease in a chronic kidney disease cohort. Clinical Proteomics. 2024 Dec;21(1):34.
13. Ugwu CN, Ugwu OP, Alum EU, Eze VH, Basajja M, Ugwu JN, Ogenyi FC, Ejemot-Nwadiaro RI, Okon MB, Egba SI, Uti DE. Medical preparedness for bioterrorism and chemical warfare: A public health integration review. Medicine. 2025 May 2;104(18):e42289.
14. Song W, Huang H, Zhang CZ, Bates DW, Wright A. Using whole genome scores to compare three clinical phenotyping methods in complex diseases. Scientific reports. 2018 Jul 27;8(1):11360.
15. Altenbuchinger M, Zacharias HU, Solbrig S, Schäfer A, Büyüközkan M, Schultheiß UT, Kotsis F, Köttgen A, Spang R, Oefner PJ, Krumsiek J. A multi-source data integration approach reveals novel associations between metabolites and renal outcomes in the German Chronic Kidney Disease study. Scientific reports. 2019 Sep 27;9(1):13954.
16. Beesley LJ, Mukherjee B. Bias reduction and inference for electronic health record data under selection and phenotype misclassification: three case studies. medRxiv. 2020 Dec 23.
17. Paul-Chima UO, Ugwu CN, Alum EU. Integrated approaches in nutraceutical delivery systems: optimizing ADME dynamics for enhanced therapeutic potency and clinical impact. RPS Pharmacy and Pharmacology Reports. 2024 Oct;3(4):rqae024.
18. Zawistowski M, Sussman JB, Hofer TP, Bentley D, Hayward RA, Wiitala WL. Corrected ROC analysis for misclassified binary outcomes. Statistics in Medicine. 2017 Jun 15;36(13):2148-60.
19. Bellocchio F, Lonati C, Ion Titapiccolo J, Nadal J, Meiselbach H, Schmid M, Baerthlein B, Tschulena U, Schneider M, Schultheiss UT, Barbieri C. Validation of a novel predictive algorithm for kidney failure in patients suffering from chronic kidney disease: The Prognostic Reasoning System for Chronic Kidney Disease (PROGRES-CKD). International Journal of Environmental Research and Public Health. 2021 Nov 30;18(23):12649.
20. Wang L, Olson JE, Bielinski SJ, St. Sauver JL, Fu S, He H, Cicek MS, Hathcock MA, Cerhan JR, Liu H. Impact of diverse data sources on computational phenotyping. Frontiers in genetics. 2020 Jun 3;11:556.
21. Brauneck A, Schmalhorst L, Weiss S, Baumbach L, Völker U, Ellinghaus D, Baumbach J, Buchholtz G. Legal aspects of privacy-enhancing technologies in genome-wide association studies and their impact on performance and feasibility. Genome Biology. 2024 Jun 13;25(1):154.

22. Ugwu OP, Ogenyi FC, Ugwu CN, Basajja M, Okon MB. Mitochondrial stress bridge: Could muscle-derived extracellular vesicles be the missing link between sarcopenia, insulin resistance, and chemotherapy-induced cardiotoxicity?. Biomedicine & Pharmacotherapy. 2025 Dec 1;193:118814.
23. La Cava W, Bauer C, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. InAMIA annual symposium proceedings 2020 Mar 4 (Vol. 2019, p. 572).
24. Agius R, Riis-Jensen AC, Wimmer B, da Cunha-Bang C, Murray DD, Poulsen CB, Bertelsen MB, Schwartz B, Lundgren JD, Langberg H, Niemann CU. Deployment and validation of the CLL treatment infection model adjoined to an EHR system. NPJ digital medicine. 2024 Jun 5;7(1):147.
25. Ortiz A. Proteomics for clinical assessment of kidney disease. PROTEOMICS–Clinical Applications. 2019 Mar;13(2):1900004.
26. Paul-Chima UO, Basajja M, Fabian CO, Chinyere NU, Ben OM, Mustafa MM. Neuro-entero-cardiac bridge: could gut-derived catecholamine-loaded extracellular vesicles synchronize the pathogenesis of Parkinson's disease, irritable bowel syndrome, and stress-triggered arrhythmias?. Medical Hypotheses. 2026 Feb 7:111896.
27. Tokita J, Vega A, Sinfield C, Naik N, Rathi S, Martin S, Wang S, Amoruso L, Zabetian A, Coca SG, Nadkarni GN. Real world evidence and clinical utility of KidneyIntelX on patients with early-stage diabetic kidney disease: interim results on decision impact and outcomes. Journal of Primary Care & Community Health. 2022 Nov;13:21501319221138196.
28. Khalid F, Alsadoun L, Khilji F, Mushtaq M, Eze-Odurukwe A, Mushtaq MM, Ali H, Farman RO, Ali SM, Fatima R, Bokhari SF. Predicting the progression of chronic kidney disease: a systematic review of artificial intelligence and machine learning approaches. Cureus. 2024 May 12;16(5).
29. Song X, Yu AS, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, Hu Y, Liu M. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nature communications. 2020 Nov 9;11(1):5668.
30. Mann SP, Treit PV, Geyer PE, Omenn GS, Mann M. Ethical principles, constraints, and opportunities in clinical proteomics. Molecular & Cellular Proteomics. 2021 Jan 1;20:100046.
31. Linder JE, Bastarache L, Hughey JJ, Peterson JF. The role of electronic health records in advancing genomic medicine. Annual review of genomics and human genetics. 2021 Aug 31;22(1):219-38.
32. Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, Liu S, Huang M, Wang C, Jiang G. Leveraging genetic reports and electronic health records for the prediction of primary cancers: algorithm development and validation study. JMIR Medical Informatics. 2021 May 25;9(5):e23586.

CITE AS: Atukunda Derrick (2026). Integrating Proteogenomics with Electronic Health Record Phenotypes for Chronic Kidney Disease Risk Prediction: Interpretability, Bias, and Real-World Performance. RESEARCH INVENTION JOURNAL OF SCIENTIFIC AND EXPERIMENTAL SCIENCES 6(1):55-66. https://doi.org/10.59298/RIJSES/2026/615566