# Identifying subtypes of chronic kidney disease with machine learning: development, internal validation and prognostic validation using linked electronic health records in 350,067 individuals

Ashkan Dashtban,[a] Mehrdad A. Mizani,[a,b] Laura Pasea,[a] Spiros Denaxas,[a] Richard Corbett,[c] Jil B. Mamza,[d] He Gao,[d] Tamsin Morris,[d] Harry Hemingway,[a,e] and Amitava Banerjee[a,f,g,*]

[a]Institute of Health Informatics, University College London, London, UK
[b]British Heart Foundation Data Science Centre, Health Data Research UK, London, UK
[c]Imperial College Healthcare NHS Trust, London, UK
[d]Medical and Scientific Affairs, BioPharmaceuticals Medical, AstraZeneca, London, UK
[e]Health Data Research UK, University College London, London, UK
[f]Barts Health NHS Trust, London, UK
[g]University College London Hospitals NHS Trust, London, UK

eBioMedicine
2023;89: 104489
Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104489

## Summary

**Background** Although chronic kidney disease (CKD) is associated with high multimorbidity, polypharmacy, morbidity and mortality, existing classification systems (mild to severe, usually based on estimated glomerular filtration rate, proteinuria or urine albumin-creatinine ratio) and risk prediction models largely ignore the complexity of CKD, its risk factors and its outcomes. Improved subtype definition could improve prediction of outcomes and inform effective interventions.

**Methods** We analysed individuals ≥18 years with incident and prevalent CKD (n = 350,067 and 195,422 respectively) from a population-based electronic health record resource (2006–2020; Clinical Practice Research Datalink, CPRD). We included factors (n = 264 with 2670 derived variables), e.g. demography, history, examination, blood laboratory values and medications. Using a published framework, we identified subtypes through seven unsupervised machine learning (ML) methods (K-means, Diana, HC, Fanny, PAM, Clara, Model-based) with 66 (of 2670) variables in each dataset. We evaluated subtypes for: (i) *internal validity* (within dataset, across methods); (ii) *prognostic validity* (predictive accuracy for 5-year all-cause mortality and admissions); and (iii) *medications* (new and existing by British National Formulary chapter).

**Findings** After identifying five clusters across seven approaches, we labelled CKD subtypes: 1. Early-onset, 2. Late-onset, 3. Cancer, 4. Metabolic, and 5. Cardiometabolic. *Internal validity*: We trained a high performing model (using XGBoost) that could predict disease subtypes with 95% accuracy for incident and prevalent CKD (Sensitivity: 0.81–0.98, F1 score:0.84–0.97). *Prognostic validity:* 5-year all-cause mortality, hospital admissions, and incidence of new chronic diseases differed across CKD subtypes. The 5-year risk of mortality and admissions in the overall incident CKD population were highest in cardiometabolic subtype: 43.3% (42.3–42.8%) and 29.5% (29.1–30.0%), respectively, and lowest in the early-onset subtype: 5.7% (5.5–5.9%) and 18.7% (18.4–19.1%). *Medications:* Across CKD subtypes, the distribution of prescription medication classes at baseline varied, with highest medication burden in cardiometabolic and metabolic subtypes, and higher burden in prevalent than incident CKD.

**Interpretation** In the largest CKD study using ML, to-date, we identified five distinct subtypes in individuals with incident and prevalent CKD. These subtypes have relevance to study of aetiology, therapeutics and risk prediction.

**Funding** AstraZeneca UK Ltd, Health Data Research UK.

*Corresponding author. Institute of Health Informatics, University College London, 222 Euston Road, London NW1 2DA, UK.
*E-mail address:* ami.banerjee@ucl.ac.uk (A. Banerjee).

**Keywords:** CKD subtype; Cluster analysis; Machine learning; Unsupervised clustering; Survival analysis

## Research in context

### Evidence before this study

We searched Pubmed, Web of Science, medRxiv, bioRxiv, and arXiv on 1 July 2022 for clustering or subtyping of chronic kidney disease using the terms "CKD", "kidney", "renal", "dialysis", "transplant", "ESRD" in combination with "clustering", "subtyping", "machine learning", "deep learning", "artificial intelligence", with no language restrictions. we used the bibliographies of the retrieved articles for literature review. We identified only one relevant study of machine learning in CKD (n = 2696). There were no studies in large-scale, representative, population-level data, across incident and prevalent CKD, all stages of CKD, and using multiple ML methods.

### Added value of this study

In a large, nationally representative, population-based dataset, we used seven machine learning methods for subtyping and risk prediction with 66 of a possible 2670 variables in cohorts of individuals with incident and prevalent CKD, including all stages of CKD (mild: stage 3a, moderate: stage 3b, and severe: stages 4 and 5). We identified and validated (internally and prognostically) five subtypes in incident and prevalent CKD: (1) early-onset; (2) late-onset; (3) metabolic; (4) cancer; and (5) cardiometabolic. These subtypes differentially predicted all-cause hospitalisation and mortality, which were highest in cardiometabolic subtype: 43.3% (42.3–42.8%) and 29.5% (29.1–30.0%), and lowest in the early-onset subtype: 5.7% (5.5–5.9%) and 18.7% (18.4–19.1%), respectively. We also demonstrated clinically relevant differences across subtypes by new and existing medications.

### Implications of all the available evidence

Identifying clinically relevant subtypes in complex conditions such as CKD could inform targeted management, healthcare resource utilization and future clinical trials, and our methods using machine learning in electronic health records are transferable to other disease areas.

## Introduction

Chronic kidney disease (CKD) has high prevalence (9.1% globally; affecting 700 million individuals), multimorbidity (as both risk factor and outcome), and burden of disease (1.2 million annual deaths worldwide).[1–4] A 2017 roadmap highlighted major global gaps in care, research and policy for CKD, recommending a ten-point action plan, mostly related to better data and personalised management, such as "improve understanding of the natural course of CKD".[5] Despite such calls for "precision nephrology", existing classification systems (mild to severe; usually based on laboratory measures, such as estimated glomerular filtration rate [eGFR], proteinuria and urine albumin-creatinine ratio)[6] and risk prediction models[7] largely ignore the complexity of CKD and its risk factors.[8] Improved subtyping of CKD offers potential for improvement in risk prediction, planning for prevention and policy. Moreover, better subtyping is necessary for new aetiologic insights and potential therapies for CKD.

Widespread implementation of electronic health records (EHR) and machine learning (ML) methods provide opportunities for better subtype definition and risk prediction across diseases.[9] However, observational studies in CKD have not made full use of available longitudinal EHR data, focusing on either incident or prevalent CKD, rather than both,[10–13] leading to concerns about the applicability and quality of existing risk prediction models. Across diseases, ML has been used for subtype definition and risk prediction,[14,15] which could ultimately facilitate early intervention and better targeted management. Like heart failure (HF) and many long-term conditions, CKD is predominantly a broad diagnosis focused on staging which often ignores aetiology, which is usually based on clinical information (e.g. diabetic or hypertensive nephropathy) rather than considering all available EHR data for a given individual with CKD. Only one study has used ML to identify subtypes in CKD, but not in a large-scale, population-based population using routine EHR data. This study used 72 baseline characteristics in 2696 individuals to define three distinct subgroups, strongly associated with future risks of CKD, and cardiovascular events, independent of established CKD risk factors.[16] Such ML methods need to be used in larger routine datasets (e.g., EHR) which are representative at national level, to derive generalisable and scalable subtypes of CKD.

Given high rates of multimorbidity and polypharmacy in individuals with CKD, a better understanding of all medications being taken at baseline and over time in CKD may inform research, whether pathophysiology or targeted intervention trials.[8] Moreover, nephrotoxicity is a common side effect of medications and in the context of CKD, medications may need to be adjusted or stopped. For example, the reported worldwide incidence of nephrotoxicity with non-steroidal anti-inflammatory drugs (NSAIDs) is 1–5%.[17] Although some EHR studies to-date have considered all prescription medication classes, they have neither considered specifically CKD, nor included all such variables in ML analyses.[18]

Using a published framework for ML studies of subtype definition and risk prediction,[9] we used seven

ML methods, in a UK population-based cohort of 350,067 individuals with incident and prevalent CKD across 2670 variables, to:

(i) Generate clinically relevant subtypes throughout the course of CKD and low risk of bias for patient selection and algorithms.
(ii) Demonstrate validity: internal (across methods) and prognostic (all-cause hospital admissions, mortality, incident diseases over five years).
(iii) Investigate distribution of prescription medication classes at baseline and over time.

## Methods

We used our published framework for ML implementation to inform our methods[9] (Fig. S1).

### To generate subtypes (development)

*Clinical relevance*

By aiming to improve diagnostic and prognostic prediction of CKD, our research concerned "patient benefit". We used population-based primary care EHR with validity for CKD research ("target condition applicability": whether the disease defined in data matches research questions). Clinical Practice Research Datalink (CPRD-GOLD), was linked by unique national healthcare identifiers, with hospital admissions (Hospital Episodes Statistics, HES), and death registry (Office for National Statistics, ONS). CPRD is representative of the UK population, with prospective recording and follow-up ("data suitability").[19]

*Patients*

We ensured "patient applicability" (to study aims), minimising patient selection bias, including individuals ≥18 years with incident (n = 350,067) and prevalent CKD (n = 195,422) and ≥1 year of follow-up in CPRD, registered with a practice from January 2006 to April 2020 in the UK. We defined "prevalent CKD" as having at least 6 months history of CKD from April 2014, and "incident CKD" as new onset from January 2006 to April 2020. We used the KDIGO (Kidney Disease Improving Global Outcomes) definition of CKD (either eGFR <60 mL/min/1.73 $m^2$ with ≥2 screening measures in 6 months, or diagnosis codes recorded in general practice), the MDRD-4 algorithm for eGFR, and code list as previously reported.[20] Phenotypes and laboratory measures were extracted using reproducible, validated algorithms (HDR UK CALIBER Phenotype Library).[21]

*Algorithm*

Applying rule-based phenotyping algorithms (n = 75), and medication chapters based on the British National Formulary (BNF), we generated 2670 variables, reflecting information before and after index date for 264 distinct factors, including: (I)socio-demographic (e.g. age; n = 7); (II)aetiology (e.g. Type 2 diabetes, T2D[3]; n = 48); (III)examination (e.g. blood pressure; n = 27), and derived events (e.g. High/Low glucose level; n = 31); and (IV)medication use and persistence (by 90-day prescription gap over 1 year[22]) (n = 151). Factors were excluded if records were incomplete or data were missing, redundant (high correlation) or highly sparse (<4% prevalence) and checked with clinical experts (non-nephrology: AB and nephrology: RC). Events created based on laboratory measures were treated as 0 if no test result was available or the available measure was not within clinically normal range; and for conditions without recorded diagnosis codes. We chose K-means clustering to obtain clusters, refining by excluding low-associated factors (to the disease subtypes) using multiple GLM (Generalised Linear Model; Poisson distribution) models. Clustering was performed using standard score and relative prevalence for baseline factors across all subtypes.

### To demonstrate validity (validation)

*Internal (within dataset and across methods)*

Seven clustering algorithms [(K-means, Diana, Hierarchical clustering (HC), fuzzy clustering (Fanny), partition around medoids (PAM), Clara, Model-based] were used in 30 subsamples of data to investigate best algorithms via analysing stability [average proportion of non-overlap (APN), average distance between means (ADM), figure of merit (FOM)], compactness (connectivity, Dunn, Silhouette), and computational complexity. We obtained the optimal number of clusters where the minimum overlap between all pairs of clusters was maximised jointly for both incident and prevalent CKD. The external validity (Stability) of subtypes was investigated (a) within clinical stages of CKD by independently choosing each CKD stage, running the clustering algorithm on, and comparing the resulting clusters with our identified subtypes using Jaccard similarity, Purity index, and survival trajectory, and (b) on whole data comparing with Clara algorithm. Four distinguished machine learning classifiers (Naïve Bays, KNN, Decision Tree, XGBoost) were trained and cross validated (5-fold; Sensitivity, Balanced accuracy, F1-score, No information rate, Kappa) to predict the identified subtypes in prevalent and incident CKD in a supervised manner.

*Prognostic (predictive accuracy for admissions and all-cause mortality)*

We analysed prevalence of risk factors and diseases in each cluster at baseline in incident and prevalent CKD, comparing Kaplan–Meier 5-year hospital admissions and all-cause mortality (log-rank for differences; p < 0.01). We also considered new-onset chronic diseases, namely, cardiovascular disease (CVD), cancer, T2D, dementia, anaemia, asthma, chronic obstructive pulmonary disease (COPD), gout, lipid disorders and anaemia. As previously published, we defined CVD as a composite of heart failure, arrhythmias, acute myocardial infarction, cardiomyopathy, atrial fibrillation, deep vein thrombosis, isolated calf vein thrombosis, pulmonary embolism, and stroke (ischaemic, transient ischaemic attack, haemorrhagic, subarachnoid haemorrhagic, and non-specified). Obesity was defined as body mass index (BMI) > 40kg/m$^{2}$.[14,20,21] Details of outcome definition are in **Panel S1**.

### Investigate distribution of prescription medication classes at baseline and over time

We examined baseline classes of medications as per BNF chapters (using methods in a previous study[23]), including the absolute and relative prevalence of the most commonly prescribed medications vs medications selected based on the standard score. The rate of new prescription was investigated for medication classes.

### Ethical approval

Study approvals were by: (i) MHRA Independent Scientific Advisory Committee [18_217R]: Section 251 (NHS Social Care Act 2006), (ii) Scientific Review Committee [17THIN038-A1] and (iii) UKB 15422: Patient informed consent was not required or provided.

### Data availability

All data produced in the present work are contained in the manuscript.

### Role of funders

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

## Results

### Development

*Clinical relevance and patients*

The study (Fig. S1) included 350,067 individuals (Age: 74.2 ± 11.9, 60.1% female, CKD stages 3a: n = 315,792, 3b: n = 10,289, 4: n = 17,373, 5: n = 6613) with incident CKD and 195,422 individuals (72.2 ± 11.8, 59.8%, 3a: n = 152,158, 3b: n = 21,244, 4: n = 15,726, 5: n = 6294) with prevalent CKD.

*Algorithm*

We selected 66 out of 2670 available variables after dimensionality reduction. The selected factors had significant associations with at least one of the disease subtypes (Table S1). Characterising clusters (Fig. S2) based on baseline aetiologic factors (Table 1) demonstrated distinct comorbidity profiles across the subtypes (Fig. 1).

### Validation

*Internal*

The optimal number of clusters with minimum overlap was 5 (Table S2, Figs. S3 and S4). In the stability analysis, identified clusters were (a) representative within clinical stages of CKD in terms of similarity (Jaccard: Incident: 83.9–99.2%, Prevalent: 82.6–95.1%, Table S3) and prognosis (Fig. S5), and (b) reproducible using at least one other clustering algorithm (Jaccard: Incident 84.6%, Prevalent 70%, Table S4). In terms of generalisability, we identified a high performing model using XGBoost that could predict disease subtypes with 95% accuracy for incident and prevalent CKD (Sensitivity: 0.81–0.98, F1:0.84–0.97, Table 2).

Five clusters were identified based on demography, risk factor burden, CKD severity, medications and laboratory factors. We labelled clusters as subtypes, after studying each cluster's characteristics: (1) early-onset; (2) late-onset; (3) metabolic; (4) cancer and (5) cardiometabolic. Among individuals with incident and prevalent CKD respectively, the proportions of subtypes were 16.9% and 14.1% for early-onset, 43.1% and 37.8% for late-onset, 12.1% and 16.6% for metabolic, 15.0% and 16.4% for cancer, and 12.9% and 15.1% for cardiometabolic subtypes (Table 1).

Age and sex varied across subtypes (oldest: Late-onset; youngest: early-onset; most females: late-onset; and least females: cardiometabolic). Severe CKD was more common in the early-onset subtype (9.2%) and least common in the late-onset subtype (5.2%). Prevalence of CVD was highest in the cardiometabolic subtype, e.g. in incident CKD, CVD: 99.9%, HF: 35.8%, and AF: 49.3%. Prevalence of hypertension (79.4%), obesity (6.8%) and T2D (99.2%) were highest in the metabolic subtype (Table 1, Fig. 1). Age, laboratory measures, BMI and blood pressure did not discriminate well between subtypes (Table 1, Table S5).

*Prognostic*

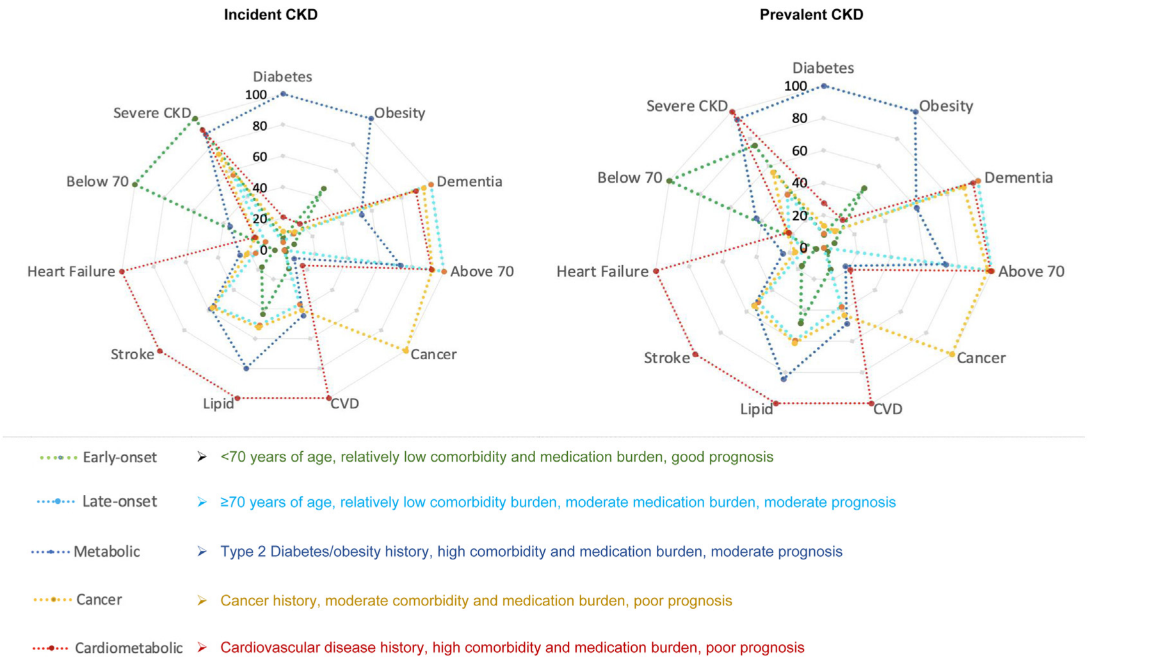
In incident CKD, 10-year mortality risks for early-onset, late-onset, metabolic, cancer and cardiometabolic subtypes were 14.1% (95% CI 13.7–14.5%), 57.4% (57.1–57.8%), 56.4% (55.7–57.0%), 66.1% (65.5–66.7%) and 71.1% (70.5–71.6%), respectively (Fig. S5). The 5-year risk of mortality and admissions in the overall incident CKD population were highest in cardiometabolic subtype: 43.3% (42.3–42.8%) and 29.5% (29.1–30.0%) (Fig. 1) respectively, and lowest in the early-onset subtype: 5.7% (5.5–5.9%) and 18.7% (18.4–19.1%). By CKD stage, 5-year risk of mortality and admissions was 62.9% (61.2–64.6%) and 34.5% (32.7–36.3%) in severe; 46.3% (43.4–49.1%) and 27.2% (25.0–29.7%) in moderate; 40.7% (40.2–41.2%) and 29.1% (28.6–29.7%) in mild CKD for the cardiometabolic subtype; and 11.0% (10.0–11.9%) and 27.3% (26.0–28.5%) in severe; 9.0% (6.2–11.6%) and 17.9% (14.6–21.2%) in moderate; 5.1% (4.9–5.3%) and 17.9% (17.5–18.2%) in mild CKD, for the early onset subtype. Risk of mortality and admission was higher in incident than prevalent CKD, across subtypes, diverging over time (Fig. 2, Figs. S5 and S6). Compared

| Risk Factor % | CKD Cohort | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Incident | | | | | | Prevalent | | | | | |
| | Early-onset | Late-onset | Metabolic | Cancer | Cardiometabolic | Overall | Early-onset | Late-onset | Metabolic | Cancer | Cardiometabolic | Overall |
| N (%) | 59,134 (16.9) | 151,053 (43.1) | 42,208 (12.1) | 52,546 (15.0) | 45,126 (12.9) | **350,067** | 27,521 (14.1) | 73,885 (37.8) | 32,357 (16.6) | 32,085 (16.4) | 29,574 (15.1) | **195,422** |
| Age (mean ± SD) | 56.0 ±9.5 | 79.2 ±7.3 | 73.0 ±9.0 | 78.5 ± 9.1 | 77.7 ±8.6 | 74.2 ±11.9 | 53.0 ±10 | 76.6 ±7.5 | 70.8 ±9.8 | 76.2 ±9.0 | 76.1 ±8.6 | 72.2 ±11.8 |
| >70 years (%) | 0.01 | 88.3 | 64.5 | 82.0 | 81.3 | 68.7 | 0.01 | 77.6 | 56.3 | 75.7 | 77.2 | 62.8 |
| Female | 57.5 | 69.1 | 55.0 | 58.5 | 40.3 | 60.1 | 57.5 | 70.5 | 53.8 | 59.1 | 42.3 | 59.8 |
| **CKD stage, eGFR < 60 mL/min/1.73 m$^2$** | | | | | | | | | | | | |
| Stage 3a, eGFR 45–59) | 89.7 | 92.6 | 84.5 | 90.5 | 88.1 | 90.2 | 84.8 | 84.2 | 64.2 | 80.2 | 67.9 | 77.9 |
| Stage 3b, eGFR 30–44) | 1.1 | 2.3 | 7.5 | 2.8 | 3.5 | 2.9 | 2.2 | 9.0 | 19.5 | 10.2 | 4.8 | 10.9 |
| Stage 4 (eGFR 15–29) | 3.5 | 4.4 | 6.5 | 5.3 | 7.1 | 5.0 | 3.9 | 5.6 | 12.3 | 7.4 | 13.8 | 8.05 |
| Stage 5 (eGFR <15) | 5.8 | 0.79 | 1.60 | 1.37 | 1.36 | 1.90 | 9.0 | 1.1 | 3.9 | 2.1 | 3.4 | 3.2 |
| Severe (4,5, eGFR<30) | 9.2 | 5.2 | 8.1 | 6.7 | 8.4 | 6.9 | 13.0 | 6.8 | 16.3 | 9.6 | 17.3 | 11.3 |
| **Smoking** | | | | | | | | | | | | |
| Smoker | 21.8 | 14.2 | 20.5 | 14.0 | 21.0 | 17.1 | 18.5 | 11.0 | 16.3 | 11.0 | 15.3 | 13.6 |
| Ever smoker | 41.9 | 50.7 | 58.1 | 53.5 | 58.1 | 51.5 | 47.3 | 57.6 | 65.5 | 60.5 | 67.3 | 59.4 |
| Non-smoker/NA | 36.3 | 35.1 | 21.4 | 32.5 | 20.9 | 31.4 | 34.2 | 31.4 | 18.2 | 28.5 | 17.4 | 27.0 |
| **Ethnicity (unknown: 45.6% for incident CKD, 50.4% for prevalent CKD)** | | | | | | | | | | | | |
| White | 51.3 | 51.2 | 53.9 | 54.2 | 53.4 | 52.3 | 45.6 | 46.8 | 44.8 | 50.0 | 50.6 | 47.4 |
| Black | 2.6 | 0.5 | 1.7 | 0.4 | 0.3 | 1.0 | 2.7 | 0.6 | 1.6 | 0.4 | 0.4 | 1.0 |
| Asian | 2.0 | 0.7 | 2.6 | 0.3 | 1.1 | 1.1 | 2.0 | 0.7 | 2.4 | 0.4 | 1.1 | 1.2 |
| **Index of multiple deprivation (unknown: 68.8% for incident CKD, 73.2% for prevalent CKD)** | | | | | | | | | | | | |
| 1 (most deprived) | 21.2 | 20.4 | 17.2 | 22.9 | 18.6 | 20.3 | 21.2 | 20.4 | 16.4 | 22.4 | 18.8 | 20.1 |
| 2 | 19.8 | 21.5 | 19.8 | 22.1 | 20.6 | 21.0 | 19.1 | 20.9 | 18.0 | 21.8 | 19.8 | 20.1 |
| 3 | 22.0 | 24.1 | 22.3 | 24.7 | 23.5 | 21.0 | 21.7 | 24.3 | 22.5 | 24.6 | 23.7 | 20.1 |
| 4 | 20.0 | 19.6 | 21.7 | 17.7 | 20.0 | 21.0 | 20.2 | 20.2 | 23.2 | 18.4 | 21.2 | 20.1 |
| 5 (least deprived) | 16.9 | 14.4 | 19.1 | 12.5 | 17.3 | 21.0 | 17.4 | 14.2 | 19.9 | 12.5 | 16.6 | 20.1 |
| **Circulatory disease** | | | | | | | | | | | | |
| Overall CVD | 13.0 | 37.0 | 44.5 | 41.0 | 99.9 | 42.6 | 13.8 | 37.6 | 38.6 | 43.3 | 99.7 | 46.4 |
| Heart failure (HF) | 1.8 | 6.1 | 9.5 | 8.0 | 35.8 | 9.9 | 1.7 | 4.6 | 9.4 | 6.6 | 38.8 | 10.5 |
| Atrial fibrillation (AF) | 3.1 | 13.3 | 15.0 | 17.0 | 49.3 | 17.0 | 3.2 | 13.0 | 16.9 | 17.3 | 59.3 | 20.0 |
| Any stroke | 3.8 | 12.3 | 12.8 | 12.6 | 21.9 | 12.2 | 4.3 | 13.0 | 13.7 | 13.8 | 25.4 | 13.9 |
| Coronary Heart Disease (CHD) | 3.4 | 10.5 | 18.0 | 12.5 | 78.7 | 19.3 | 3.8 | 9.8 | 20.1 | 13.3 | 79.8 | 21.8 |
| Myocardial infarction (MI) | 0.8 | 0.2 | 2.1 | 1.9 | 72.3 | 10.1 | 1.1 | 0.9 | 5.6 | 2.6 | 58.6 | 10.8 |
| Venous thromboembolism (VTE) | 3.0 | 5.0 | 5.3 | 7.2 | 7.3 | 5.3 | 3.6 | 5.7 | 6.4 | 8.2 | 8.9 | 6.4 |
| Cardiac Valve disorder | 1.7 | 4.8 | 4.4 | 5.9 | 13.7 | 5.5 | 2.1 | 5.1 | 5.3 | 6.4 | 9.2 | 7.1 |
| Unstable angina (UA) | 0.7 | 1.2 | 2.8 | 1.9 | 16.5 | 3.4 | 0.7 | 1.2 | 3.2 | 1.9 | 17.2 | 4.0 |
| Peripheral Vascular Disease and Abdominal aortic aneurysm (PAD) | 1.5 | 5.0 | 10.0 | 5.8 | 16.3 | 6.6 | 1.5 | 4.9 | 10.7 | 6.1 | 18.0 | 7.6 |
| Stable angina (SA) | 5.4 | 13.4 | 22.3 | 15.5 | 73.8 | 21.3 | 5.9 | 13.3 | 25.3 | 15.9 | 74.0 | 23.8 |
| Hypertension | 50.3 | 68.9 | 79.4 | 66.4 | 68.8 | 66.7 | 57.2 | 75.2 | 83.4 | 74.7 | 77.6 | 74.3 |
| Bradycardia or Tachycardia | 2.3 | 4.4 | 17.1 | 6.1 | 14.3 | 7.1 | 2.8 | 5.4 | 20.4 | 7.5 | 19.0 | 9.9 |
| **Cancer** | | | | | | | | | | | | |
| Any Cancer | 2.4 | 0.0 | 9.2 | 100 | 15.8 | 18.6 | 3.2 | 0.0 | 17.1 | 100 | 20.9 | 22.9 |
| Cancer – Charlson indexed | 1.2 | 0.0 | 5.3 | 73.0 | 8.8 | 12.9 | 1.5 | 0.0 | 11.3 | 72.0 | 11.2 | 15.6 |
| Cancer – non-Charlson | 1.2 | 0.0 | 4.3 | 36.1 | 8.2 | 7.2 | 1.8 | 0.0 | 7.2 | 39.0 | 11.7 | 9.6 |
| Skin Biopsy | 21.9 | 23.1 | 24.3 | 30.1 | 31.1 | 25.5 | 22.5 | 21.4 | 22.6 | 27.0 | 29.5 | 23.9 |

(Table 1 continues on next page)

| Risk Factor % | CKD Cohort | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Incident | | | | | | Prevalent | | | | | |
| | Early-onset | Late-onset | Metabolic | Cancer | Cardiometabolic | Overall | Early-onset | Late-onset | Metabolic | Cancer | Cardiometabolic | Overall |
| (Continued from previous page) | | | | | | | | | | | | |
| **Respiratory disease** | | | | | | | | | | | | |
| Asthma | 15.0 | 12.7 | 15.5 | 13.4 | 16.2 | 14.0 | 17.2 | 14.2 | 17.7 | 15.0 | 18.0 | 15.9 |
| COPD | 3.4 | 8.2 | 7.7 | 8.8 | 14.3 | 8.2 | 4.2 | 9.5 | 10.1 | 10.3 | 17.0 | 10.1 |
| **Endocrine, nutritional, metabolic** | | | | | | | | | | | | |
| Type 2 Diabetes | 6.9 | 4.5 | 99.2 | 1.5 | 20.8 | 19.5 | 8.2 | 8.7 | 99.2 | 13.5 | 27.6 | 27.2 |
| Lipid disorders | 15.0 | 17.5 | 27.4 | 18.0 | 34.3 | 20.0 | 18.0 | 23.2 | 32.9 | 23.9 | 39.0 | 26.7 |
| Chronic Anaemia | 5.4 | 8.2 | 10.4 | 10.1 | 11.2 | 8.7 | 7.7 | 9.9 | 17.4 | 12.2 | 16.6 | 12.2 |
| Thyroid disorders | 11.7 | 15.5 | 14.7 | 14.1 | 14.5 | 14.4 | 13.8 | 18.2 | 17.3 | 16.5 | 18.2 | 17.2 |
| Obesity: BMI $\geq$ 40 | 3.1 | 0.8 | 6.8 | 0.9 | 1.3 | 2.0 | 3.2 | 0.9 | 7.3 | 0.9 | 1.5 | 2.4 |
| Underweight: BMI$\leq$18.5 | 0.5 | 1.1 | 0.5 | 1.0 | 1.0 | 0.9 | 0.5 | 1.2 | 0.4 | 1.2 | 1.1 | 1.0 |
| **Skin and subcutaneous tissue diseases** | | | | | | | | | | | | |
| Lupus | 4.4 | 4.0 | 4.8 | 4.5 | 5.0 | 4.4 | 5.3 | 4.7 | 5.9 | 5.2 | 6.0 | 5.3 |
| Systemic lupus erythematosus (SLE) | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.3 | 0.8 | 0.8 | 0.2 | 0.3 | 0.3 | 0.3 |
| **Musculoskeletal diseases** | | | | | | | | | | | | |
| Rheumatoid arthritis | 5.9 | 4.3 | 5.8 | 5.5 | 5.8 | 5.4 | 9.4 | 0.8.0 | 8.5 | 8.1 | 9.7 | 8.6 |
| Osteoarthritis | 19.2 | 40.3 | 37.1 | 41.3 | 41.8 | 36.7 | 22.4 | 45.5 | 42.6 | 46.6 | 49.1 | 42.5 |
| Gout | 6.8 | 7.4 | 11.0 | 9.0 | 15.1 | 9.0 | 10.6 | 10.3 | 15.3 | 13.0 | 24.4 | 13.8 |
| **Other risk factors** | | | | | | | | | | | | |
| History of Influenza | 8.8 | 15.8 | 19.9 | 17.7 | 23.5 | 16.4 | 16.3 | 23.1 | 30.7 | 25.7 | 33.6 | 25.4 |
| Dementia | 0.6 | 7.9 | 4.2 | 7.5 | 7.1 | 6.1 | 0.8 | 11.5 | 6.9 | 10.4 | 11.1 | 9.0 |
| **Albumin to creatinine ratio: 89.9% for incident CKD, 56.6% for prevalent CKD)** | | | | | | | | | | | | |
| uACR >30 mg/g | 0.67 | 0.23 | 2.8 | 0.38 | 0.70 | 0.69 | 2.0 | 1.22 | 7.6 | 2.0 | 3.60 | 2.90 |

*Table 1:* **Baseline characteristics by subtype in incident (n = 350,067) and prevalent (n = 195,422) chronic kidney disease.**



*Axes show relative prevalence of baseline characteristics, scaled to 100 for the subtype with highest prevalence of that factor. For example, having cancer is most frequent in the Cancer subtype. CKD: chronic kidney disease.*

***Fig. 1:*** Baseline characteristics by disease subtype for incident and prevalent chronic kidney disease.

with the early-onset subtype, the relative risk for 5-year mortality was highest for the cardiometabolic subtype (Relative risk [RR] 8.38, 8.14–8.63, and 11.80, 11.15–12.51 for incident and prevalent CKD respectively) and lowest for the late-onset subtype (5.56, 5.41–5.72 and 6.68, 6.31–7.07 for incident and prevalent CKD respectively) (Table S6).

Generally, risk of developing new chronic disease was greater in prevalent CKD than incident CKD, except for heart failure, atrial fibrillation, stroke and lipid disorders. In incident CKD, 5-year risk of developing any cancer (7.2%, 14.5%, 13.1%, 0.0% and 15.9%), CVD (10.0%, 24.4%, 25.4%, 26.0% and 0.0%), heart failure (1.9%, 7.6%, 9.4%, 9.0% and 17.0%), atrial fibrillation (3.4%, 11.8%, 11.2%, 12.9% and 16.5%), coronary heart disease (4.1%, 6.5%, 9.0%, 7.7% and 27.4%), stroke (2.8%, 7.9%, 8.1%, 8.6% and 11.4%) and anaemia (3.9%, 6.9%, 11.3%, 8.2% and 9.9%) varied across early-onset, late-onset, metabolic, cancer and cardiometabolic subtypes, respectively (Fig. S7).

### *Medications*

Across CKD subtypes, the distribution of prescription medication classes at baseline varied, with highest medication burden in cardiometabolic and metabolic subtypes, and higher burden in prevalent than incident CKD. For example, for hypertension and heart failure medications, across early-onset, late-onset, metabolic, cancer and cardiometabolic subtypes, the prescription rate was 24.3% and 39.8%, 36.3% and 49.5%, 69.1% and 78.2%, 36.3% and 48.4%, and 64.3% and 71.4%, respectively in incident (overall: 41.0%) and prevalent (overall: 56.0%) CKD (Table S7).

The most commonly prescribed medication class differed by CKD subtype: early-onset: psychoses and hormones; late-onset: musculoskeletal and joint diseases; metabolic: hypertension and heart failure, and endocrine, genito-urinary disorders; cancer: immunosuppression and local anaesthesia; and cardiometabolic: hypertension and heart failure, antiarrhythmics and anticoagulants (Fig. 3, Fig. S8). Overall, the most commonly prescribed medication classes were hypertension and heart failure, lipid-regulating drugs, analgesics, antibacterial drugs, antisecretory drugs, nitrates and calcium channel blockers, antiplatelet drugs, beta-adrenoceptor blockers, diuretics, rheumatic diseases and gout, and antidepressant drugs, and the highest use of these medications was generally in cardiometabolic and metabolic subtypes (Fig. 3 and Fig. S9). In general, the baseline medication distribution was similar across late-onset and cancer subtypes, and across metabolic and cardiometabolic subtypes.

In terms of new prescription medication classes over 5 years, the rates for antibacterial drugs (69.1% and 56.5%); analgesics (57.3% and 43.3%); anti-secretory drugs and mucosal protectants (43.0% and 30.5%); hypertension and heart failure (35.9% and 18.7%); diuretics (30.9% and 22.2%); lipid-regulating drugs

| Cohort | Classifier | Metric | Subtypes | | | | | All-data | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Early-onset | Late-onset | Metabolic | Cancer | Cardiometabolic | Overall accuracy | Kappa |
| Incident[b] NIR: 0.4315 $P<\varepsilon$ | Naïve Bays | Sensitivity | 0.844 | 0.616 | 0.307 | 0.152 | 0.221 | 0.624 (0.57–0.65) | 0.307 (0.26–0.32) |
| | | F1-score | 0.826 | 0.687 | 0.616 | 0.120 | 0.443 | | |
| | | B-Accuracy | 0.945 | 0.785 | 0.720 | 0.554 | 0.586 | | |
| | KNN | Sensitivity | 0.967 | 0.967 | 0.444 | 0.484 | 0.546 | 0.777 (0.75–0.79) | 0.676 (0.66–0.69) |
| | | F1-score | 0.861 | 0.833 | 0.592 | 0.628 | 0.682 | | |
| | | B-accuracy | 0.955 | 0.849 | 0.718 | 0.737 | 0.769 | | |
| | Decision Tree | Sensitivity | 0.915 | 0.957 | 0.883 | 0.972 | 0.772 | 0.920 (0.90–0.93) | 0.889 (0.88–0.90) |
| | | F1-score | 0.915 | 0.943 | 0.883 | 0.956 | 0.831 | | |
| | | B-accuracy | 0.949 | 0.951 | 0.934 | 0.981 | 0.880 | | |
| | XGBoost | Sensitivity | 0.959 | 0.965 | 0.921 | 0.980 | 0.842 | 0.945 (0.94–0.95) | 0.924 (0.92–0.93) |
| | | F1-score | 0.951 | 0.957 | 0.924 | 0.974 | 0.934 | | |
| | | B-Accuracy | 0.973 | 0.963 | 0.955 | 0.987 | 0.915 | | |
| Prevalant NIR: 0.3781 $P<\varepsilon$ | Naïve Bays | Sensitivity | 0.810 | 0.530 | 0.256 | 0.233 | 0.270 | 0.574 (0.55–0.59) | 0.212 (0.20–0.22) |
| | | F1-score | 0.847 | 0.654 | 0.616 | 0.09 | 0.547 | | |
| | | B-Accuracy | 0.888 | 0.764 | 0.756 | 0.521 | 0.575 | | |
| | KNN | Sensitivity | 0.498 | 0.500 | 0.975 | 0.959 | 0.563 | 0.751 (0.73–0.77) | 0.657 (0.64–0.67) |
| | | F1-score | 0.629 | 0.637 | 0.847 | 0.794 | 0.702 | | |
| | | B-accuracy | 0.740 | 0.744 | 0.960 | 0.841 | 0.778 | | |
| | Decision Tree | Sensitivity | 0.900 | 0.801 | 0.889 | 0.935 | 0.882 | 0.897 (0.87–0.92) | 0.864 (0.84–0.88) |
| | | F1-score | 0.927 | 0.787 | 0.900 | 0.926 | 0.900 | | |
| | | B-accuracy | 0.945 | 0.881 | 0.941 | 0.942 | 0.936 | | |
| | XGBoost | Sensitivity | 0.959 | 0.814 | 0.954 | 0.965 | 0.926 | 0.933 (0.93–0.94) | 0.916 (0.91–0.92) |
| | | F1-score | 0.960 | 0.844 | 0.945 | 0.950 | 0.936 | | |
| | | B-Accuracy | 0.976 | 0.897 | 0.972 | 0.962 | 0.958 | | |

[a]Five-fold cross validation. [b]NIR denotes "No Information Rate" for the data set. B-accuracy denotes balanced accuracy i.e., the average of sensitivity plus specificity per class. KNN is the K-nearest neighbour algorithm. $P<\varepsilon$ denotes p-value <0.0001 using 5-fold cross validation.

*Table 2:* **Performance of four supervised machine learning models for predicting disease subtypes in incident (n = 350,067) and prevalent (n = 195,422) chronic kidney disease[a].**

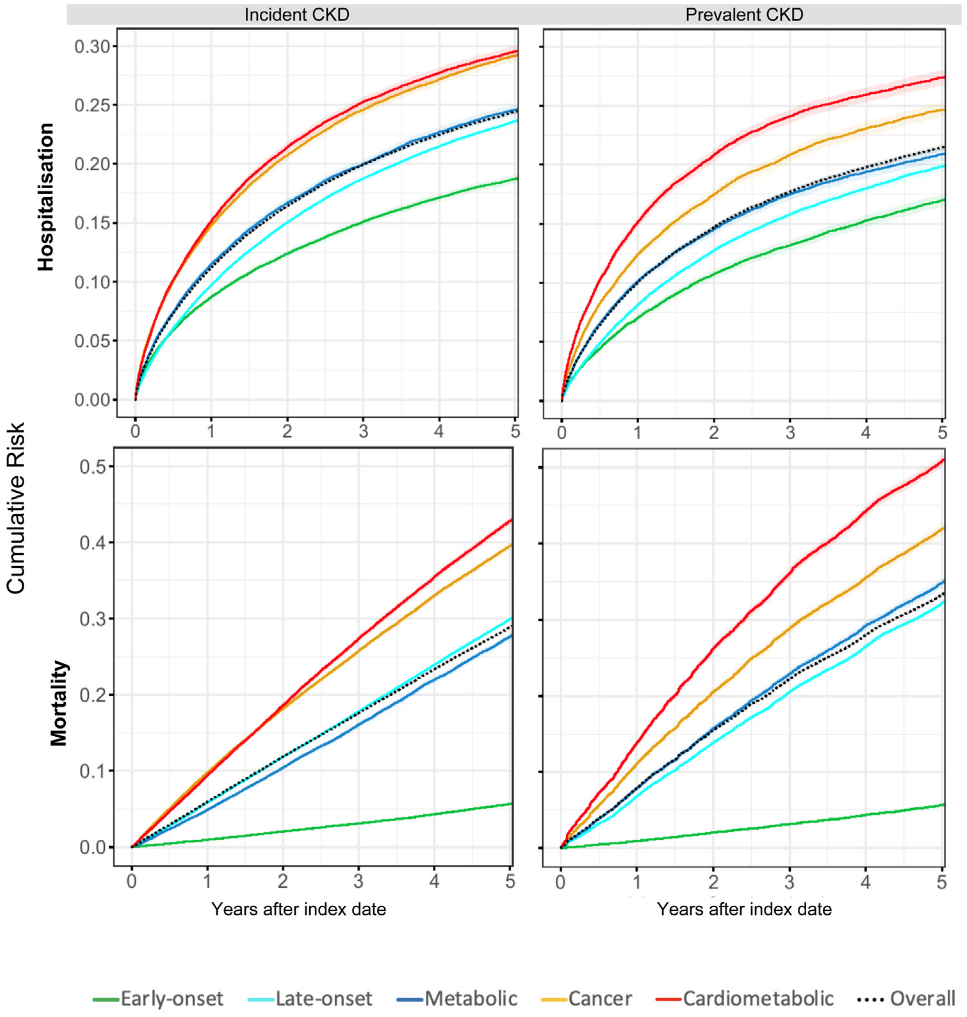
(30.8% and 18.5%); nitrates, calcium-channel blockers (29.4% and 19.2%); and medications used in rheumatic diseases and gout (27.4% and 19.4%) were high in incident and prevalent CKD respectively. Other than medications used in rheumatic diseases and gout (highest in the late-onset subtype), and hypertension and heart failure (highest in the metabolic subtype), rates of new prescription were highest across medication classes in the cardiometabolic subtype and lowest in the early onset subtype, for incident and prevalent CKD (Fig. 4).

## Discussion

In the largest and most representative study to-date of machine learning-informed subtype definition and risk prediction in CKD, we have three major findings. First, we identified five distinct subtypes across incident and prevalent CKD with clinically important differences across baseline characteristics: early-onset, late-onset, metabolic, cancer and cardiometabolic subtypes, with rigorous internal validation. Second, we highlighted high 5-year rates of all-cause hospital admissions, mortality and incident chronic diseases in individuals with both incident and prevalent CKD and demonstrated important differences across subtypes. Third, we comprehensively showed high medication burden, with differences across CKD subtypes and across incident and prevalent CKD.

Classification systems for severity of CKD are well-established in guidelines and in clinical practice,[6,7] like subtypes in other disease areas, such as left ventricular ejection fraction-based classification in heart failure.[14] However, there is substantial scope for improvement for clinical, public health and research applications in primary and secondary prevention, including precision medicine.[13] Based on easily available clinical characteristics from routine EHR, including socio-demographic, aetiologic and laboratory investigation variables, we developed 5 subtypes with validity which we showed in incident and prevalent CKD, the overall population and subgroups (mild vs severe CKD). Whether epidemiology during the pandemic[22] or novel therapeutic approaches for chronic diseases (e.g. SGLT2-inhibitors[24]), links across traditional, organ- and disease-specific silos are becoming more apparent. Our ML-informed subtypes signal the focus for integrated CKD primary prevention: T2D, hypertension, CVD,[1,5] cancer[2] and age.

The internal validation conducted in our study is robust and reproducible; using multiple ML methods (n = 7), variables (n = 66, selected from n = 2670

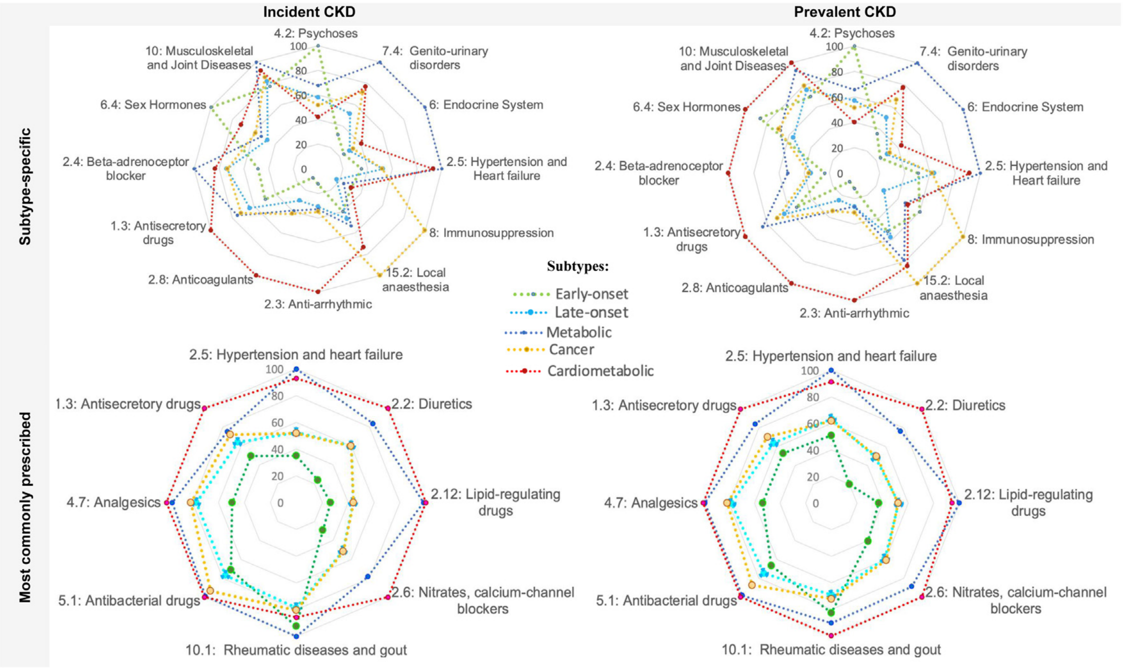


*CKD: chronic kidney disease, P-value < 0.01 in all plots.*

***Fig. 2:*** Five-year risk of all-cause hospitalisation and mortality by disease subtypes in prevalent and incident chronic kidney disease.

variables derived from 264 factors), classifiers (n = 4) and metrics (n = 3), following our framework for use of ML in subtyping and risk prediction research.[9] It is now important to conduct external validation of these subtypes in an independent dataset. Moreover, as we have suggested for subtypes in HF,[14] genetic validation of the CKD subtypes may help to understand the importance and role of particular aetiologic factors. For example, apolipoprotein L1 (APOL1) is a relevant genotype[25] associated with albuminuria, subclinical atherosclerosis, incident myocardial infarction, and mortality[26]; with two kidney risk variants (APOL1 G1 and G2) associated with nondiabetic CKDs specially in people with African origin.[27] The cardiometabolic subtype with highest incidence of CVD, MI and mortality may have an association with these genetic loci. To show clinical utility of this subtype classification, we will need to assess the effectiveness and cost-effectiveness. The methodology and framework we used is applicable, generalisable and scalable to other diseases where EHR data are available, particularly where there are already existing subtype classifications with scope for improvement.

The cardiometabolic subtype has the worst prognosis: associated with relatively higher incidence of cancer, CVD, and highest mortality (5- and 10-year mortality rates of 43.3% and 71.1%), in line with existing data supporting mutual associations between severity of CKD and CVD.[2–4] Relatively lower rate of comorbidities, admissions and mortality in the late-onset subtype, which accounts for two-fifths of incident and prevalent CKD, suggests that age alone is an important predictor of onset and progression of CKD. Current criteria for CKD using the same eGFR



*Axes show relative prevalence of BNF sections of medications, scaled to 100 for the subtype with highest prevalence of that factor. For example, Antisecretory drugs were most often prescribed for the Cardiometabolic subtype. CKD: chronic kidney disease.*

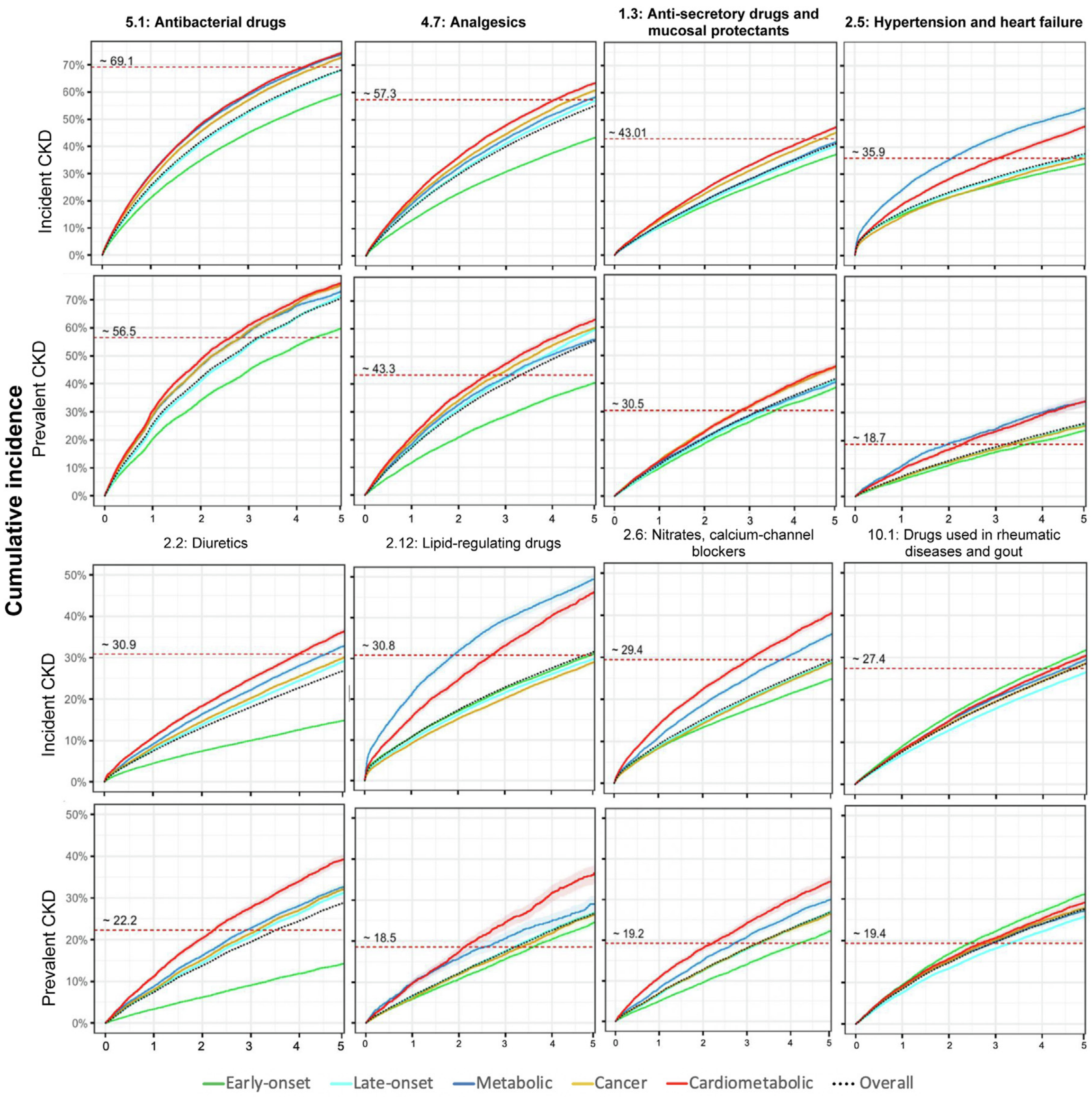
***Fig. 3:*** Relative prevalence of British National Formulary sections of prescribed medication by disease subtype for incident and prevalent chronic kidney disease.

threshold for all ages "may result in overestimation of the CKD burden in an aging population, overdiagnosis, and unnecessary interventions in many elderly people who have age-related loss of eGFR", based on analyses showing similar rates of progression to end-stage renal disease as the non-CKD population is individuals >65 years.[28] Our finding of relatively lower all-cause mortality in the late-onset subtype (especially in prevalent CKD) supports these analyses. Further research should prospectively test and validate predictive accuracy of subtypes and outcomes, and the preventable burden of disease following CKD diagnosis, perhaps informing clinical guidelines and resource allocation. Like our prior study,[20] we showed higher risk of mortality and admission in incident than prevalent CKD. Potential explanations are greater severity and progression at CKD incidence, lower rates of treatment in incident disease and better control of comorbidities in prevalent CKD (e.g. hypertension, CVD), which future studies must address. Guidelines already emphasise early diagnosis and targeted management of hypertension, T2D and CVD,[5] but not necessarily in cancer or other chronic diseases. As in primary prevention, our analyses support personalised, multidisciplinary approaches to secondary prevention in those with CKD at individual and population levels.

We show the scale of prescribed medication burden at baseline and over 5 years after diagnosis in people with CKD, which highlights the multimorbidity and complexity of management of CKD and associated comorbidities and underlines the need for more evidence to inform management guidelines.[29] Moreover, in combination with data about new prescribing and incident chronic disease, we believe there are four applications for these data. First, such data could help clinicians and policymakers in planning care and resource utilisation for people with CKD. Second, the large-scale data regarding new and existing medications, and new and existing diseases by CKD stage and CKD subtype could have direct clinical applications and inform guidelines, where inappropriate prescribing before and after diagnosis are common.[30,31] Third, epidemiology and data science approaches in such large-scale longitudinal data could facilitate knowledge and action to prevent medication-induced kidney disease, e.g. the "6R framework" (risk, recognition, response, renal support, rehabilitation and research[30]), and could help in understanding trajectory and pathophysiology of CKD. Fourth, the fact that new and existing medications seem to vary by our identified CKD subtypes is of interest, showing cross-medication analysis as a potential form of validation of ML-informed subtypes.

## Strengths and limitations

This is the largest study to-date to develop and validate ML-informed subtypes in CKD. Moreover, we used



** This analysis only included individuals who did not receive any prescription for the specified medication in the two years prior to index date. The dashed horizontal line shows the prescription rate in the overall population (without considering censoring time). CKD: chronic kidney disease.*

*Fig. 4:* New prescribed medication rate by disease subtype for most commonly prescribed medication classes in individuals with chronic kidney disease over 5 years.

rigorous ML methods, a published framework for implementation of ML in subtype definition and risk prediction, a large number of variables, and a large, nationally representative study population. However, there were several limitations. First, we did not externally validate our subtypes or undertake genetic analyses as in our previous work in HF subtypes.[14] Second, we investigated neither the acceptability and clinical utility of the CKD subtypes nor their effectiveness and cost-effectiveness, which future research should consider. Third, we considered all-cause mortality and hospital admissions and therefore cannot comment on specific causes. Fourth, we analysed by medication chapters in the BNF and did not evaluate individual medications. Fifth, we emphasised mortality, hospitalisation and incidence of other chronic diseases, but did not consider renal replacement therapy, which is important clinically,[32] and should be considered as an outcome in future validation analyses. Sixth, although we have labelled subtypes as "early onset" and "late onset", it is important to note that early or late presentation may, at least partially, be due to differences in health-seeking behaviour, social deprivation or access to primary care as much as physiological differences, but were not investigated in this study. Seventh, our data may not represent all populations, e.g. ethnic minorities were under-represented. Finally, misclassification of diseases due to incorrect coding is possible as in any electronic health record study, although the risk and impact is likely to be minimal in such a large-scale dataset.

## Conclusion

Current classifications of CKD do not capture the complexity of comorbidities, medications, disease trajectory and outcomes for research or clinical practice. In the largest study of machine learning in CKD to-date, we confirmed high burden of comorbidities and medications, poor prognosis by mortality, admissions and new chronic disease, and defined and validated five subtypes, which may have both academic and clinical applications.

### Contributors

AB conceived the research question. AB, JBM, and TM obtained funding. MAM, LP, AB, and AD wrote the study protocol. AD designed and led analyses. AB and MAM verified the data. AB and AD drafted the initial and final versions of manuscript. RC provided nephrology input to analyses and their interpretation. SD and HH contributed expertise in EHR research, particularly phenotyping. HG provided additional epidemiology input. All authors critically reviewed early and final versions of the manuscript.

### Data sharing statement

We used Clinical Practice Research Datalink, CPRD-GOLD[19] that is not publicly available. However, all analyses and resulting data are included in this manuscript and supplementary materials.

### Declaration of interests

JBM, HG and TM are employed by AstraZeneca UK Ltd, a biopharmaceutical company and declare stock and stock options. Other authors report no conflicts of interest.

### Acknowledgments

AB is supported by research funding from the National Institute for Health Research (NIHR), British Medical Association, AstraZeneca, and UK Research and Innovation, and Trustee of the South Asian Health Foundation. HH is a National Institute for Health Research (NIHR) Senior Investigator. HH is funded by the National Institute for Health Research University College London Hospitals Biomedical Research Centre, Health Data Research UK (grant No. LOND1, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care [England], Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division [Welsh Government], Public Health Agency [Northern Ireland], British Heart Foundation and Wellcome Trust). AB, SD and HH are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC.

### Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104489.

### References

1 GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2020;395(10225):709–733.

2 Stengel B. Chronic kidney disease and cancer: a troubling connection. *J Nephrol*. 2010;23(3):253.

3 Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, et al. Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. *Lancet*. 2013;382:339–352.

4 van der Velde M, Matsushita K, Coresh J, et al. Lower estimated glomerular filtration rate and higher albuminuria are associated with all-cause and cardiovascular mortality: a collaborative meta-analysis of high-risk population cohorts. *Kidney Int*. 2011;79: 1341–1352.

5 Levin A, Tonelli M, Bonventre J, et al. Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet*. 2017 Oct 21;390(10105):1888–1917.

6 Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Inter*. 2013;3: 1–150.

7 Grams ME, Sang Y, Ballew SH, et al. Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. *Kidney Int*. 2018;93(6): 1442–1451.

8 Provenzano M, De Nicola L, Pena MJ, et al. Precision nephrology is a non-negligible state of mind in clinical research: remember the past to face the future. *Nephron*. 2020;144(10):463–478.

9 Banerjee A, Chen S, Fatemifar G, et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Med*. 2021;19(1):85.

10 Perotte A, Ranganath R, Hirsch JS, et al. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *J Am Med Inform Assoc*. 2015;22(4):872–880.

11 Fraccaro P, van der Veer S, Brown B, et al. An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK. *BMC Med*. 2016 Jul 12;14:104.

12 Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017 Jan;24(1):198–208.

13 Collins GS, Omar O, Shanyinde M, et al. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol*. 2013;66(3):268–277.

14 Banerjee A, Chen S, Dashtban A, et al. Identifying subtypes of heart failure with machine learning: external, prognostic and genetic validation in three electronic health record sources with 320,863 individuals. *medRxiv*; 2022 [Preprint and under peer review)] https://www.medrxiv.org/content/10.1101/2022.06.27.22276961v1.full.

15 Pikoula M, Quint JK, Nissen F, et al. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. *BMC Med Inform Decis Mak*. 2019;19(1):1–14.

16 Zheng Z, Waikar SS, Schmidt IM, et al. Subtyping CKD patients by consensus clustering: the chronic renal insufficiency cohort (CRIC) study. *J Am Soc Nephrol*. 2021 Mar;32(3):639–653.

17 Pazhayattil GS, Shirali AC. Drug-induced impairment of renal function. *Int J Nephrol Renovasc Dis*. 2014;7:457–468.

18 Stafford G, Villén N, Roso-Llorach A, et al. Combined multimorbidity and polypharmacy patterns in the elderly: a cross-sectional study in primary health care. *Int J Environ Res Public Health*. 2021;18(17):9216.

19 Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Informatics Assoc*. 2019;26(12): 1545–1559.

20 Dashtban A, Mizani MA, Denaxas S, et al. A retrospective cohort study measured predicting and validating the impact of the COVID-19 pandemic in individuals with chronic kidney disease. *Kidney Int*. 2022;102(3):652–660.

21 Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million

individuals in the English National Health Service. *Lancet Digit Heal*. 2019;1(2):e63–e77.

22 Norris T, Razieh C, Zaccardi F, et al. Impact of cardiometabolic multimorbidity and ethnicity on cardiovascular/renal complications in patients with COVID-19. *Heart*. 2022;108(15):1200–1208.

23 Torabi F, Akbari A, Bedston S, et al. Impact of COVID-19 pandemic on community medication dispensing: a national cohort analysis in Wales, UK. *Int J Popul Data Sci*. 2022;5(4):1715.

24 Heerspink HJL, Stefánsson BV, Correa-Rotter R, et al. Dapagliflozin in patients with chronic kidney disease. *N Engl J Med*. 2020;383(15):1436–1446.

25 Limou S, Nelson GW, Kopp JB, Winkler CA. APOL1 kidney risk alleles: population genetics and disease associations. *Adv Chron Kidney Dis*. 2014;21(5):426–433.

26 Mukamal KJ, Tremaglio J, Friedman DJ, et al. APOL1 genotype, kidney and cardiovascular disease, and death in older adults. *Arterioscler Thromb Vasc Biol*. 2016;36(2):398–403.

27 Hung AM, Shah SC, Bick AG, et al. APOL1 Risk variants, acute kidney injury, and death in participants with African ancestry hospitalized with COVID-19 from the million veteran program. *JAMA Intern Med*. 2022;182(4):386–395.

28 Liu P, Quinn RR, Lam NN, et al. Accounting for age in the definition of chronic kidney disease. *JAMA Intern Med*. 2021 Oct 1;181(10):1359–1366.

29 MacRae C, Mercer SW, Guthrie B, et al. Comorbidity in chronic kidney disease: a large cross-sectional study of prevalence in Scottish primary care. *Br J Gen Pract*. 2021 Feb 25;71(704):e243–e249.

30 Awdishu L, Mehta RL. The 6R's of drug induced nephrotoxicity. *BMC Nephrol*. 2017 Apr 3;18(1):124.

31 Sommer J, Seeling A, Rupprecht H. Adverse drug events in patients with chronic kidney disease associated with multiple drug interactions and polypharmacy. *Drugs Aging*. 2020 May;37(5):359–372.

32 Go AS, Chertow GM, Fan D, McCulloch CE, Hsu CY. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *N Engl J Med*. 2004 Sep 23;351(13):1296–1305.