Received: 24 May 2021 | Revised: 24 May 2021 | Accepted: 21 June 2021

DOI: 10.1002/psp4.12695

ARTICLE

CKD subpopulations defined by risk-factors: A longitudinal analysis of electronic health records

Rajagopalan Ramaswamy[1] | **Soon Nan Wee**[1] | **Kavya George**[1] | **Abhijit Ghosh**[1] | **Joydeep Sarkar**[1] | **Rolf Burghaus**[2] | **Jörg Lippert**[2]

[1]Advanced Analytics, Holmusk, Singapore

[2]Pharmacometrics, Bayer AG – Pharmaceuticals, Wuppertal, Germany

Correspondence
Jörg Lippert, Bayer AG - Pharmaceuticals, Pharmacometrics, 42096 Wuppertal, Germany.
Email: joerg.lippert@bayer.com

Funding information
This work was supported by Bayer AG Pharmaceuticals, Wuppertal, Germany.

Abstract

Chronic kidney disease (CKD) is a progressive disease that evades early detection and is associated with various comorbidities. Although clinical comprehension and control of these comorbidities is crucial for CKD management, complex pathophysiological interactions and feedback loops make this a formidable task. We have developed a hybrid semimechanistic modeling methodology to investigate CKD progression. The model is represented as a system of ordinary differential equations with embedded neural networks and takes into account complex disease progression pathways, feedback loops, and effects of 53 medications to generate time trajectories of eight clinical biomarkers that capture CKD progression due to various risk factors. The model was applied to real world data of US patients with CKD to map the available longitudinal information onto a set of time-invariant patient-specific parameters with a clear biological interpretation. These parameters describing individual patients were used to segment the cohort using a clustering approach. Model-based simulations were conducted to investigate cluster-specific treatment strategies. The model was able to reliably reproduce the variability in biomarkers across the cohort. The clustering procedure segmented the cohort into five subpopulations – four with enhanced sensitivity to a specific risk factor (hypertension, hyperlipidemia, hyperglycemia, or impaired kidney) and one that is largely insensitive to any of the risk factors. Simulation studies were used to identify patient-specific strategies to restrain or prevent CKD progression through management of specific risk factors. The semimechanistic model enables identification of disease progression phenotypes using longitudinal data that aid in prioritizing treatment strategies at individual patient level.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Chronic kidney disease (CKD) progression involves complex biological pathways resulting in a large variability in progression characteristics. Although deep learning models can capture complex associations, they lack clear interpretability, and

do not account for causal biological relationships, thus offering insights that are not clinically actionable.

WHAT QUESTION DID THIS STUDY ADDRESS?

This study presents a semimechanistic modeling methodology and applies it to multidimensional longitudinal real-world electronic health record (EHR) data to segment patients into subpopulations based on dominant drivers of CKD progression.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

The model mapped the longitudinal EHR data onto time invariant patient-specific parameters and clustering on these parameters yield five CKD subpopulations: four with enhanced sensitivity to a specific risk factor and one that is largely insensitive to any risk factor.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT AND/OR THERAPEUTICS?

Assignment of patients with CKD to clusters based on driving risk factor will aid in designing patient-specific treatment strategies.

INTRODUCTION

Chronic kidney disease (CKD) is a growing cause of global concern affecting ~ 753 million patients[1] globally in 2016 and accounting for ~ 1.2 million deaths[2] in 2017 worldwide. CKD seldomly occurs in isolation and is almost always associated with various cardiovascular and metabolic comorbidities, like hypertension,[3] hyperlipidaemia,[4] and hyperglycemia.[5,6] Clinical management is challenging as CKD gradually progresses almost imperceptibly over years and is usually diagnosed in advanced stages when kidney function is already irreversibly impaired.[7] Although the management of progression involves treating the associated comorbidities, CKD in turn increases the risks of exacerbation of the comorbidities through feedback loops that pervade biological and physiological systems.[8–11] This makes the analysis of longitudinal data on CKD progression difficult using regular statistical methods; any analytical model for CKD progression should incorporate effects of these physiological feedback loops.

In most cases, the individual risk factors for CKD work through multiple mechanisms.[6,12,13] The risk factors themselves evolve with time and so does their effect on the kidneys.[14] The management of risk factors varies significantly across patients due to various factors, including differences in care delivered. Consequently, across a large cohort of patients with CKD, a wide variability in progression rates is observed.[15,16] Due to large variations in the risk factors themselves, it is difficult to gauge whether the variability in CKD progression rates arises due to variability in risk factors or are there any patient-specific aspects, which further affect the CKD progression rates. Planning interventions and treatment strategies for CKD thus requires a better understanding of disease progression and contribution of different comorbidities; hence, a methodology that takes into account the time varying effects of risk factors along with their inter-relationships and feedbacks simultaneously is desired for generating insights for better patient management.

In literature, prior studies have developed risk prediction models[17–20] (primarily based on Cox regression) for onset and progression for CKD. There are also animal models[21] to understand CKD pathophysiology, which come with their own class of problems[22] in translation to humans. Alternatively, there are pure mechanistic models[23–25] based on quantitative systems pharmacology (QSP) that incorporate physiological inter-relationships and feedbacks and have been primarily applied to drug development. In this paper, we present a novel hybrid semimechanistic modeling approach that combines differential equations-based mechanistic QSP modeling with modern machine learning techniques to investigate CKD subpopulations in real-world data. Unlike pure deep learning-based black-box models, a large part of our model is built upon mechanistic QSP modeling to explicitly represent established causal inter-dependencies and feedback loops for the simulation of time trajectories of clinical biomarkers and risk factors associated with CKD. Machine learning components, such as neural networks, were used to complement the mechanistic equations in parts where the exact functional form for the biological interactions are intractable. Using this semimechanistic modeling methodology, we have mapped longitudinal patient data from electronic health records (EHRs) onto a set of time-invariant patient-specific parameters that have

a clear biological interpretation (i.e., sensitivity of renal damage to different risk factors). These mechanistic parameters describing individual patients were subsequently utilized to understand the dominant drivers of CKD progression in different patient subpopulations.

METHODS

Analysis cohort definition

For this study, longitudinal EHR data of patients diagnosed with CKD within the period of data availability (January 2014–December 2017: 4 years) was utilized. The data were collected from 468,998 different hospitals in the United States with ~ 1.6 million unique patients with CKD. As illustrated in Figure 1, the analysis cohort was defined by a stepwise application of the following exclusion criteria: (1) patients with no laboratory/vitals measurements, (2) patients who suffer from acute kidney injuries (AKIs) or have undergone surgical interventions since the study's focus is gradual long-term progression of CKD and not acute renal events that can cause abrupt changes in kidney functioning; also in addition, the mechanism of CKD progression would be different between patients with and without AKI, hence modeling AKI-CKD interaction would involve developing a largely different model structure given the complex etiology of AKI and may also potential require additional biomarkers[26,27] not frequently captured in our EHR data, (3) patients who did not have sufficient amount of longitudinal data for biomarkers captured by the model (see Supplementary Material, Section S1). These criteria also ensure that patients in the analysis dataset satisfy a minimal set of data quality and quantity requirements. We checked for potential biases introduced by our selection process by a comparison of descriptive statistics between the complete and the analysis datasets (see Results).

Model structure

Figure 2a shows a schematic of our semimechanistic model with its different components and their interrelationships. A large part of the model is developed mechanistically considering the underlying biological causality driven by physiological interactions and feedback loops. Black-box machine learning components complement the above approach in sections where the exact functional form representing the interaction of biological mechanisms is intractable, for example, blood

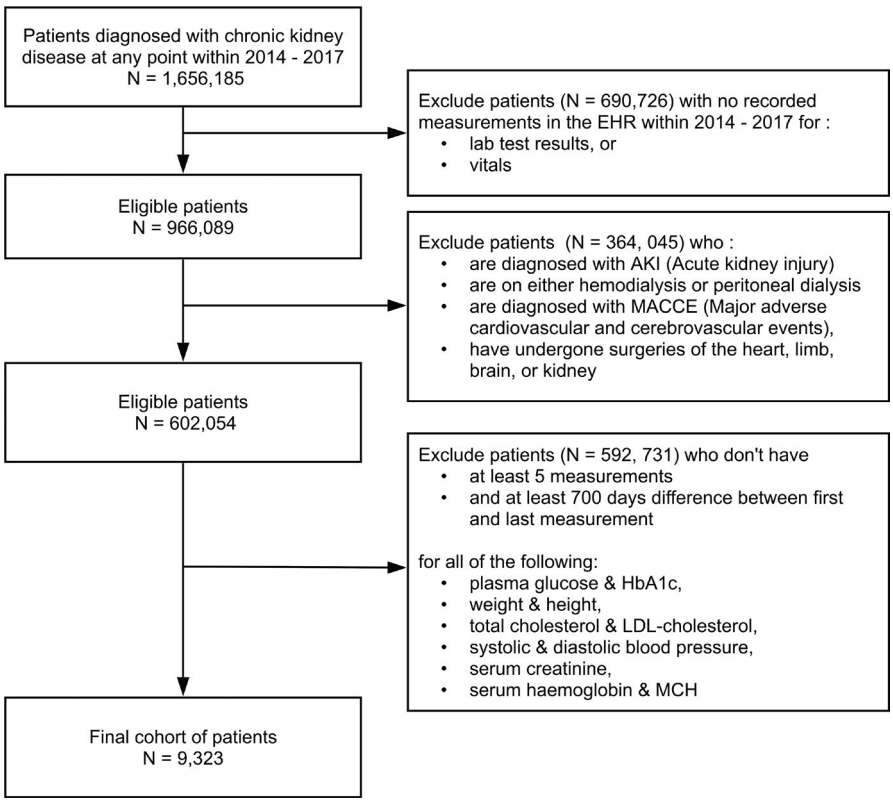

FIGURE 1 Cohort selection process. Flowchart of cohort selection process showing exclusion criteria applied on the population of patients diagnosed with chronic kidney disease from 2014 to 2017. EHR, electronic health record

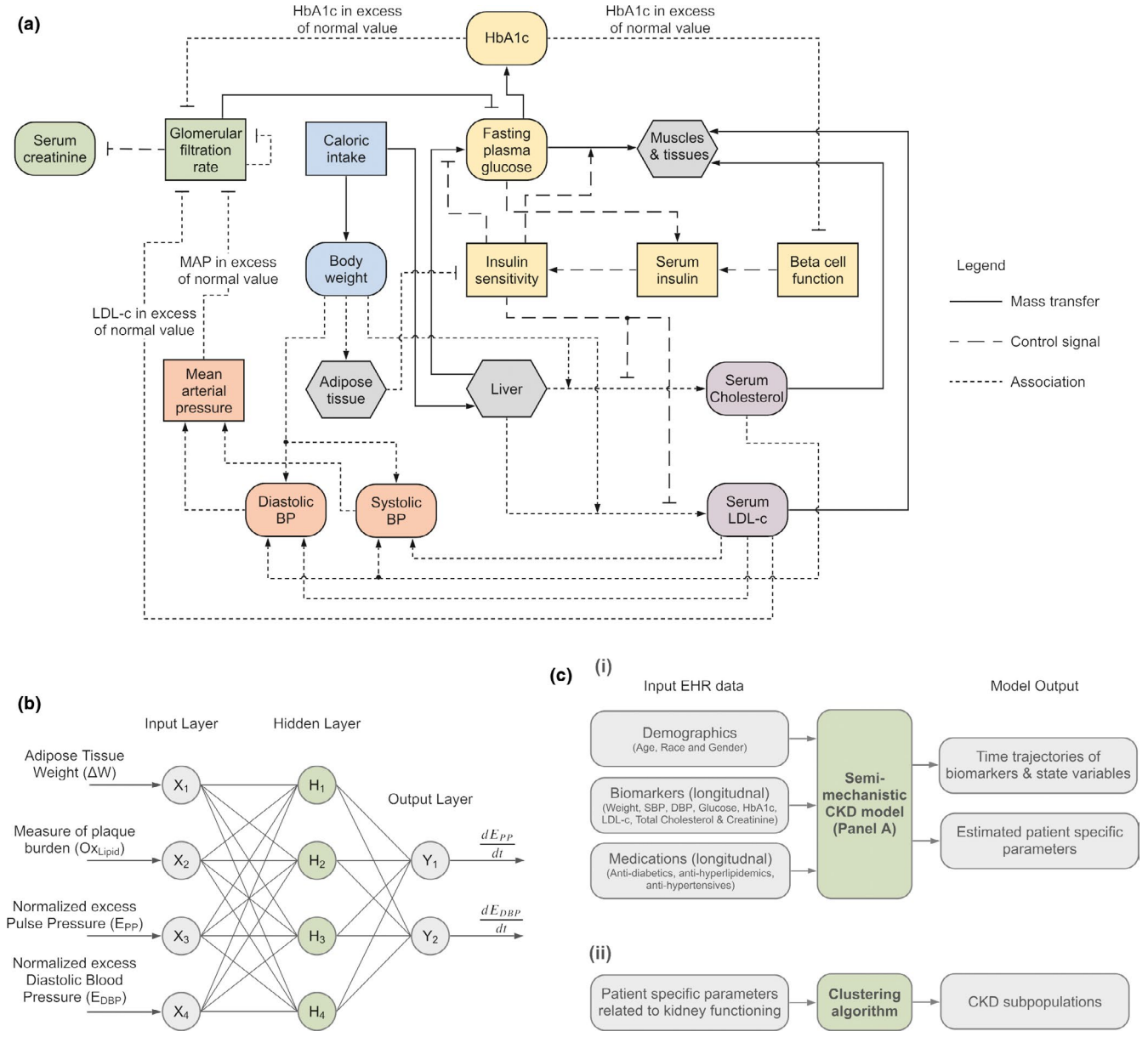

FIGURE 2 Illustration of model. (a) Schematic flow-diagram of semimechanistic model showing the inter-relationship between the different components. The rounded rectangles represent the biomarkers that are captured in the electronic health record (her) data; ordinary rectangles represent the internal state variables and components with physiological meaning but are not captured in the EHR data; the hexagons in dark grey background represent the abstract systems, like the liver, etc. The model components are grouped into different subsystems, based on the underlying physiology, which are visually distinguished through different colors – body weight (blue), glucose-insulin-HbA1c (yellow), lipids (purple), blood pressure (orange), and renal functioning (green). (b) Architecture of neural network representing the time evolution of blood pressure. The neural network used a single hidden layer with four neurons with sigmoid activation and an output layer with two neurons with ReLU activation. (c) (i) Schematic illustrating the inputs and outputs of the semimechanistic model. The patient-specific parameters of the semimechanistic model are estimated using the inputs - time-invariant patient data (demographics) as well as longitudinal patient data (Biomarkers and Medications) from the EHR. (c) (ii) Schematic illustrating the clustering procedure. The patient-specific parameters representing sensitivities of chronic kidney disease (CKD) to different risk factors were analyzed using clustering algorithms to divide the patient cohort into CKD subpopulations

pressure trajectories. The model components can be grouped into subsystems (Figure 2a) representing body weight, glucose metabolism, lipid metabolism, blood pressure, and kidney function itself. Table 1 lists down each subsystem and their main state variables (formal descriptions in Supplementary Material, Section S3). Within each subsystem, the effects of individual medications were incorporated as Hill functions at appropriate physiological pathways based on mechanism of action (Supplementary Material S3). The constants within the

TABLE 1 Subsystems of the semi-mechanistic model and their main state variables

Subsystem	State variable	Meaning	Physiological processes
Body weight	Lean body weight	Weight of all organs except adipose tissue	Constant over time
	Adipose tissue	Weight of adipose tissue	Increases with excess caloric input and decreases when caloric input is lower than maintenance demand
Glucose-insulin-HbA1c system	Glucose	Serum glucose concentration	Influx components: carb. intake from food, glucose released by liver, re-absorption from the kidneys Outflux components: absorption by tissues mediated by insulin and insulin sensitivity and removal through the kidneys
	Insulin	Serum insulin level	Increases with elevated serum glucose level and higher pancreatic beta cell functioning
	Beta cell function	Levels of beta cell functioning determining the rate of insulin production	Increases with elevated glucose levels due to beta cell replication; prolonged exposure to hyperglycemia causes reduction due to beta cell death
	Insulin sensitivity	Sensitivity of cells to reduce blood glucose levels in response to insulin	Decreases with excess adipose tissue mass; inversely related with insulin resistance
	HbA1c	Serum HbA1c concentration	Increases with increased serum glucose levels and is removed with RBC turnover
Lipids	Serum cholesterol	Total serum cholesterol level	Increases with excess adipose tissue mass and increased insulin resistance
	Serum LDL-C	Serum LDL cholesterol level	Increases with excess adipose tissue mass and increased insulin resistance
BP	Systolic BP Diastolic BP		Increases with age due to arterial stiffness, which is further increased by plaque burden
	Mean arterial pressure	Time-weighted average arterial pressure during single cardiac cycle	Increases with increase in either systolic or diastolic BP
Kidney functioning	GFR	Represents flow rate of filtered fluids by kidney; indicator of kidney functioning	Decreases with age, increase in levels of risk factors (MAP, LDL-C, and glucose)
	Creatinine	Serum creatinine levels	Increases with decreased GFR

Abbreviations: BP, blood pressure; carb., carbohydrate; GFR, glomerular filtration rate; RBC, red blood cell.

Hill function (maximal response, half maximal response dosage, and Hill exponent) for each medication is estimated from dose-response curves from literature (details and references in Supplementary Section S6).

A brief description of each subsystem is furnished as follows.

Body weight

The time evolution of body weight (equations in Supplementary Material, Section S3B) is modeled as a function of net daily caloric intake[28] and was decomposed into two parts:(a) lean body weight (LBW) and (b) adipose tissue weight (ATW). The model captures the rate of change of ATW while assuming that LBW remains constant over time.[29] When excess net caloric intake (caloric intake in excess of the person's total daily energy expenditure) increases, the ATW increases; whereas in the absence of additional net caloric intake, the body weight converges to LBW. Increase in ATW triggers the development of insulin resistance as described below.

Glucose-insulin-HbA1c

The evolution of serum glucose levels (equations in Supplementary Material, Section S3C–E) takes into

account the glucose-insulin homeostatic feedback[30] and depends on glucose released from the liver,[31] serum insulin levels, and insulin sensitivity. Serum insulin levels are modeled to rise when there is excess glucose in blood as observed in an oral glucose tolerance test (OGTT).[32] However, prolonged exposure to hyperglycemia results in damage to and ultimately death of pancreatic beta cells[33] and associated reduction in fasting insulin levels and insulin response to glucose.[34] In addition, insulin sensitivity is modulated by excess adipose tissue,[35] which in turn depends on net caloric intake. The effect of diabetes medications (listed in Supplementary Material, Section S6) including metformin, DPP-4I, SGLT-2I, sulfonylurea, and GLP1-A are included to modulate the appropriate parts of model dictated by mechanism of action. Referring to Figure 2a, it is noted that we have associated HbA1c as a chronic measure of glucotoxicity because HbA1c levels are reflective of the exposure to blood glucose levels over a few months.

Lipids

The primary biomarkers captured in the lipids system are low-density lipoprotein cholesterol (LDL-C) and total cholesterol. The model includes the effects of excess weight and insulin sensitivity on lipid synthesis (equations in Supplementary Material, Section S3F) and the effect of statins (listed in Supplementary Material, Section S6) in suppressing lipid synthesis.[36] Although high density lipoprotein levels varied between patients, they did not vary significantly over time, as evidenced in the data; and hence was not included.

Blood pressure

The time evolution of blood pressure depends on many factors, which includes arterial stenosis, obesity, stress, dysregulation of Renin-Angiotensin-Aldosterone system, and possibly hitherto unidentified drivers.[37,38] As illustrated in Figure 2b, the time evolution of blood pressure was modeled with a neural network model with the following inputs: excess body weight,[39] an estimate of plaque burden using cumulative exposure[40] to oxidized lipid levels, and a feedback of prevailing normalized blood pressure levels (equations in Supplementary Material, Section S3G). The neural network contained a single hidden layer with four neurons with sigmoid activation and a output layer with two neurons with ReLU activation. The effect of antihypertensive medications (listed in Supplementary Material, Section S6) is also considered. Although both systolic and diastolic blood pressure can show significant variation over extremely short time scales (approximately a few hours) depending on various factors, including stress, this short time variation of blood pressure is not incorporated.

Glomerular filtration rate and serum creatinine

In this study, we represent kidney function by the glomerular filtration rate (GFR). The time evolution of GFR (equations in Supplementary Material, Section S3H) incorporates the role of different risk factors, including hyperglycemia, hyperlipidemia, hypertension, and autocrine inflammation, as well as accounts for the natural decline associated with aging. The evolution of serum creatinine levels is expressed as a function of GFR assuming a quasi-steady approximation, as the dynamics of serum creatinine levels are much faster than the gradual decline rate of GFR.[37] We directly modeled serum creatinine as it is in a continuous range as compared to CKD stage information that is available in EHR data in the form of International Classification of Diseases (ICD) codes.

Simulations of differential equations

The model simulation of different patient biomarkers is posed as an initial value problem, which was numerically solved using the odeint module from SciPy[41] using the "LSODA" method.[42] The initial condition corresponds to that of a healthy lean individual at the age of 20 years, which is the simulation start timepoint. The majority of initial conditions were known from the healthy values of different biomarkers available from literature; whereas the remaining were calculated using the structure of the equations assuming steady-state at initial condition. A list of state variables and their initial conditions are listed in Supplementary Material, Section S4.

Modeling summary

Figure 2c summarizes how the semimechanistic model was applied to EHR data to identify CKD subpopulations. Specifically, the semimechanistic model uses the EHR data (demographics and medications) to simulate the time-trajectories of different state variables and biomarkers. The simulated time-trajectories were fit to the longitudinal biomarker data from EHR to estimate the model parameters (described in Section "Model parameters and estimation procedure"). The patient-specific parameters

related to renal functioning were then applied to the clustering algorithm (described in Section "Clustering") to segment the patient population into different CKD subgroups.

Model parameters and estimation procedure

The model parameter values, which are related to physiological pathways, vary from patient to patient and has substantial variability in the EHR data. As such, a single set of values for the model parameters is insufficient to simulate the biomarker trajectories for all the patients in the EHR data. Consequently, we have two sets of parameters:

- Population-level parameters, which are assumed as constants across the entire patient cohort. The values for these constants are directly available in literature (e.g., red blood cell turnover rate and insulin removal rate) or were calibrated based on data from literature (e.g., constants in serum glucose clearance equations were calibrated using OGTT data). These population-level parameters are listed for each physiological system as constants in Supplementary Material, Section S3.
- Patient-specific parameters, which exhibit significant variation over the population and thus vary on a per-patient basis (e.g., rate constant of damage to pancreatic β-cells). The value of these parameters is estimated through optimization process. Patient-specific parameters for each physiological system are listed in Table S2 (also indicated as patient-specific parameters in Supplementary Material, Section S3).

To estimate the patient-specific parameters, we used the Differential Evolution algorithm.[43] As the total number of parameters to be estimated is high, the optimization process was carried out for each patient in a sequential fashion for each component as listed below.

- Body weight
- Glucose-HbA1c-insulin
- Lipids
- Systolic and diastolic blood pressure
- estimated GFR (eGFR) and creatinine.

For each optimization step, the best parameters obtained are assumed to be constant and used in the next step.

The loss function for optimization at each step is defined as the sum of scaled mean square error (MSE) of the biomarkers in that step:

$$\text{Loss function} = \sum_{j=1}^{M} \frac{1}{\left[\max\left(y_{1j}, y_{2j}, \cdots, y_{N_j j}\right)\right]^2} \cdot \frac{1}{N_j} \sum_{i=1}^{N_j} \left(y_{ij} - \hat{y}_{ij}\right)^2$$

where y_{ij} and $\hat{y}_{ij}$ are the values of j^{th} biomarker at i^{th} timepoint from EHR data and model's simulated trajectory respectively, N_j = number of available timepoints for j^{th} biomarker, and M = number of biomarkers considered in that optimization step. In essence, given a patient, the MSE for each biomarker is scaled by the inverse of squared maximum value of that biomarker in order to account for the differences in scales of different biomarkers. After parameter estimation, robustness of parameter estimates and goodness-of-fit for each biomarker were evaluated (See Supplementary Material, Sections S7, S8).

Clustering

The patient-specific parameters have a biological interpretation and were subsequently analyzed using clustering to understand different patient phenotypes. Clustering was performed on the following estimated parameters used in differential equations for eGFR progression (Supplementary Material Section S3H, Eq. 29), which correspond to the sensitivity of GFR decline to the following specific risk factors:

$k_{E_{egfr},LDL}$: Sensitivity to excess LDL (LDL in excess of 30 mg/dL)

$k_{E_{egfr},HbA1c}$: Sensitivity to excess HbA1c (HbA1c in excess of 4.5%)

$k_{E_{egfr},MAP}$: Sensitivity to excess MAP (MAP in excess of 83.33 mmHg)

$k_{E_{egfr},E_{egfr}}$: Sensitivity to existing renal impairment (E_{egfr} =1 for a healthy kidney).

Clustering on these patient parameters allows for identification of groups of patients who have similar response to risk factors that drive renal impairment. Prior to clustering, the parameters were standardized to have zero mean and unit variance. Spectral clustering with second degree polynomial kernel was used. To determine the optimum number of clusters, Silhouette score[44] and Calinski-Harabasz score[45] were used. Higher values of these scores indicate better clustering quality and the optimum number of clusters was determined when these scores were maximized. In addition, the Jaccard score[46] index was used to establish the robustness of clustering method. Refer to Supplementary Material, Section S9 for more details on clustering and evaluation metrics.

Simulation studies

We performed simulation studies to explore changes in eGFR decline across different clusters in response to reduction in specific risk factors as described below:

- GFR decline is quantified as ΔeGFR = (eGFR at the end of 4 years − eGFR at baseline). Simulation end timepoint was chosen as 4 years as the EHR data approximately contained 4 years' worth of data.
- The semimechanistic model is used to simulate biomarker trajectories for each patient with a particular risk factor (MAP, LDL-C, or HbA1c) reduced at one time.
- For each simulation scenario, ΔeGFR is estimated and is used to study the differences in GFR decline among the different clusters.

RESULTS

Analysis of bias in cohort selection

To understand if there is any bias introduced due to cohort selection, we compared the distributions of demographics, biomarkers, and vitals for the population cohort (N = 966,089) and the analysis cohort (N = 9323), as shown in Figure 3. Because the number of patients in these cohorts is large, the statistical difference in the distributions of these cohorts have been

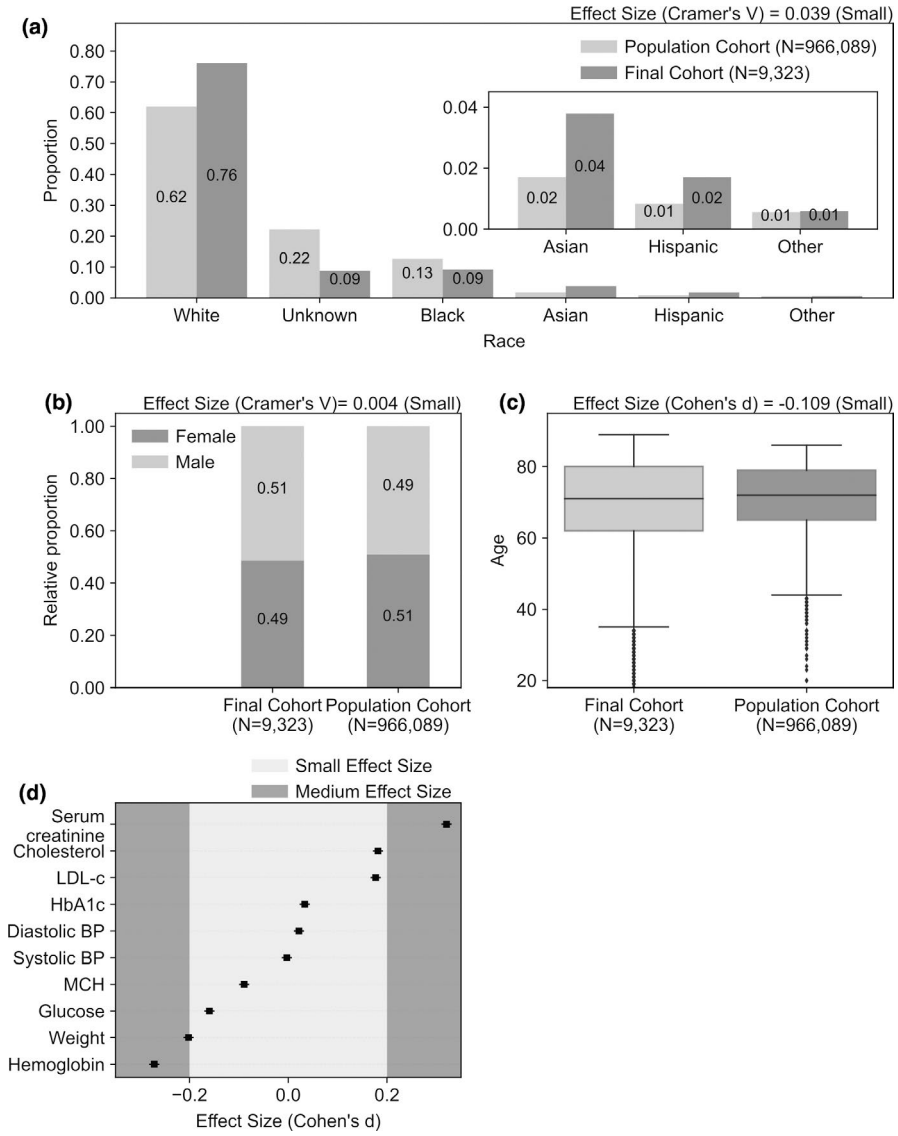

FIGURE 3 Analysis of bias in cohort selection. (a) Proportion of individual races in the two cohorts (inset shows magnified graph for the Asian, Hispanic, and other races), (b) Relative proportion of gender in the two cohorts. (c) Box and whiskers plot showing the age distribution for the two cohorts, (d) Forest plot of the effect size (Cohen's d) of biomarkers and vitals for the two cohorts

quantified in terms in effect size (Cramer's V[47] for categorical variables and Cohen's d[48] for continuous variables) rather than looking at statistical tests of difference between their means. Figure 3a–c show the distributions of demographics in the two cohorts and Figure 3d shows a forest plot of effect size of the differences in the biomarkers and vitals used in the cohort selection process between the two cohorts. Except for serum creatinine and hemoglobin, the effect sizes of the differences between the two cohorts are small. The medium effect size (>0.2) of serum creatinine and hemoglobin indicate that our analysis cohort is slightly more diseased than complete cohort. For demographics, although there is a slight increase in Asian, Hispanic, and White patients and a slight decrease in Black patients and patients with unknown race in the analysis cohort, the effect sizes of the differences were small.

Simulated trajectories from the semimechanistic model

Figure 4 illustrates the simulated biomarker trajectories from the semimechanistic model along with the raw EHR data for an example patient after parameter estimation. We see that the simulated trajectories follow the observed biomarker data quite closely. The model also captures the effect of changes in medications on target biomarkers. For example, the patient in Figure 4 starts a combinatory antihypertensive therapy at around $t = 18$ months (indicated by drug

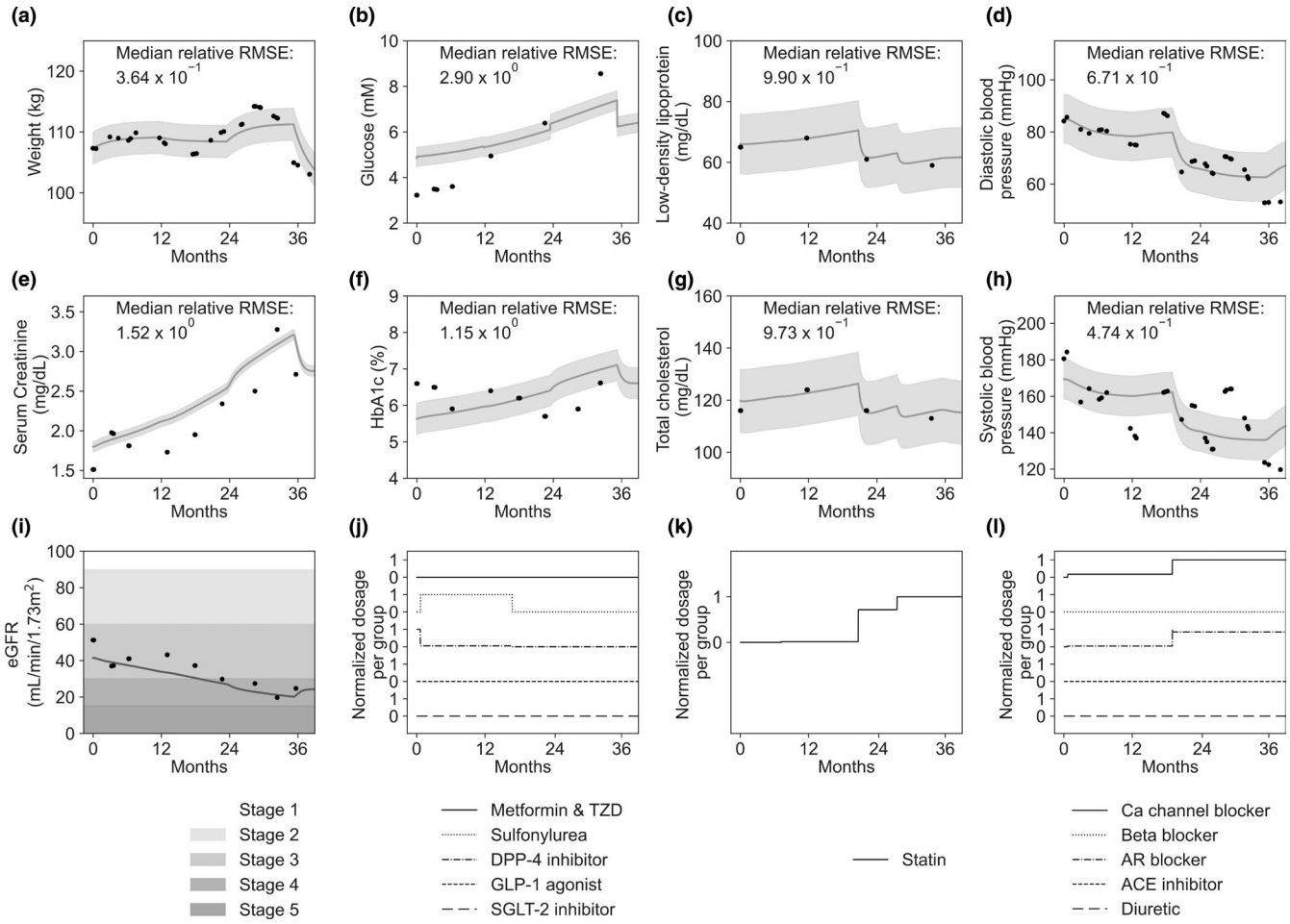

FIGURE 4 Illustration of biomarker trajectories from the model simulations along with the electronic health record (EHR) data and medication data for an example patient. The different panels correspond to (a) weight (kg), (b) glucose (mM), (c) LDL-C (mg/dl), (d) diastolic blood pressure (mmHg), (e) serum creatinine (mg/dl), (f) HbA1c (%), (g) total cholesterol (mg/dl), (h) systolic blood pressure (mmHg), and (i) estimated glomerular filtration rate (eGFR; ml/min/1.73 m^2). Patient medication data obtained from EHR (dosage normalized to each subgroup) are also plotted for (j) antidiabetics, (k) statins, and (l) antihypertensives. In panels (a–h), black circles represent the biomarker data from EHR data and grey lines represent the simulated biomarker trajectory from the semimechanistic model. The shaded bands around the simulated trajectories indicate the literature reported variability in clinical measurements for each biomarker. Conversion factors for units: serum creatinine in mg/dl to μmol/L, × 88.4; LDL-C in mg/dl to μmol/L, × 0.02586; total cholesterol in mg/dl to μmol/L, ×0.02586. RMSE, root mean squared error; TZD, thiazolidinedione

dosage increase in Figure 4l). This results in a corresponding decrease in systolic (Figure 4d) and diastolic (Figure 4h) blood pressure in EHR data beyond $t = 18$ months, which is also reproduced in the model's simulated trajectory. Similarly, introduction of statins at $t = 18$ months for this patient (Figure 4k) corresponds to a decrease in lipid levels (Figure 4c,g).

Some patients were dropped at each optimization step due to non-robust parameter estimates (Supplementary Material, Section S7) or high fitting errors. After parameter estimation, the cohort consisted of 7792 patients for whom the model can satisfactorily fit. The median coefficient of variation in the estimates of patient-specific parameters related to CKD progression was less than 10% indicating they are fairly robust. The goodness of fit varies with each biomarker and Figure 4 indicates the median relative root mean square error (RMSE; refer to Figure S3 and Supplementary Material, Section S8) for each biomarker. The median relative RMSE is ~ 1 for most biomarkers, indicating that the model is able to track the EHR data well.

Clustering results

The clustering methodology applied to patient-specific parameters related to the CKD progression yielded five clusters. Figure 5 shows the clustering results through

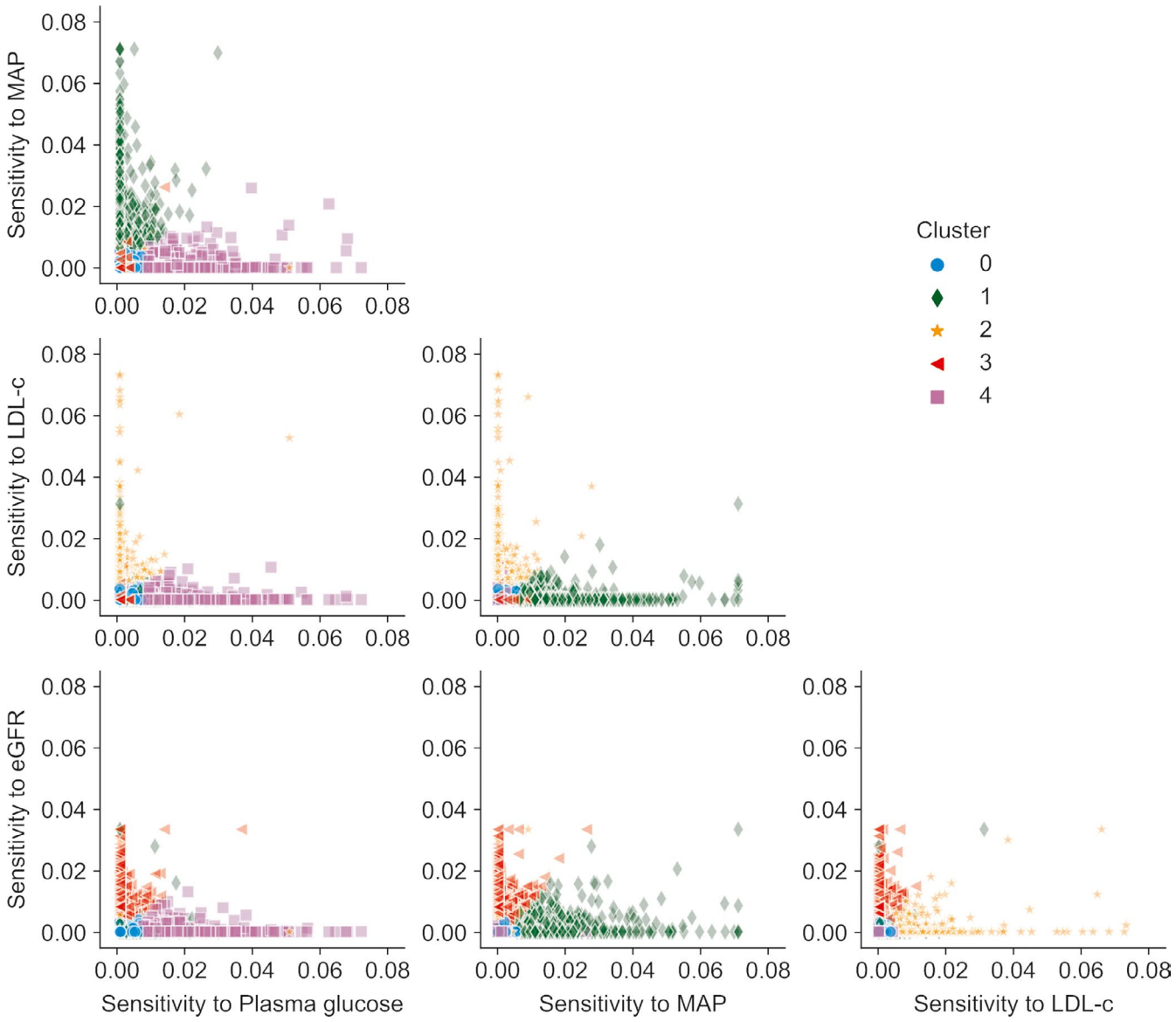

FIGURE 5 Visualization of clusters. The 2D scatterplots for each combination of sensitivity parameters used in the clustering procedure. The different clusters (0, 1, 2, 3, and 4) are illustrated using different colors and symbols. Cluster 0 consists of patients who are insensitive to the modeled risk factors, whereas clusters 1, 2, 3, and 4 consist of patients whose chronic kidney disease progression is driven by hypertension, hyperlipidemia, impaired estimated glomerular filtration rate and hyperglycemia, respectively

2D scatter plots for different combinations of scaled parameters used in clustering (see Supplementary Material, Section S9). The different clusters are visually distinguished using different symbols and colors.

From Figure 5, we see that majority of patients in clusters 1, 2, 3, and 4 are concentrated along a particular axis (i.e., they have high values for one of the 4 parameters used in clustering) indicating their CKD progression is predominantly sensitive to a specific risk factor (hypertension, hyperglycemia, hyperlipidemia, or existing renal impairment). There is also a fifth cluster (cluster-0) that is not particularly sensitive to any of these risk factors. These results are also illustrated via. The 1D marginal distribution plots for each scaled parameter in Figure S5. Based on these observations, the characteristics of each cluster (driving risk factor for CKD progression) are summarized below:

- Cluster-0: Not particularly sensitive to any risk factor ($N = 2780$)
- Cluster-1: More sensitive to hypertension ($N = 848$)
- Cluster-2: More sensitive to hyperlipidemia ($N = 622$)
- Cluster-3: More sensitive to existing renal impairment ($N = 2827$)
- Cluster-4: More sensitive to hyperglycemia ($N = 715$).

Simulation studies

Figure 6 shows the simulation results studying GFR response (quantified by ΔeGFR defined in Methods) due to reduction in specific risk factors across different clusters. When there is no risk factor reduction (x-axis value = 0 in Figure 6), mean ΔeGFR is negative for all clusters (i.e., on an average), eGFR declines over 4 years for all clusters. This is expected because the cohort is composed of patients with CKD. When a specific risk factor is reduced, ΔeGFR becomes less negative; whereas when a specific risk factor is increased, ΔeGFR becomes more negative. Specifically, when a particular risk factor (MAP, LDL-C, and HbA1c) is changed, the cluster identified to be sensitive to the corresponding risk factor (cluster-1, cluster-2, and cluster-4, respectively) shows the largest change in ΔeGFR. Changes in HbA1c show the biggest effect in ΔeGFR, because HbA1c also affects LDL-C and MAP indirectly. Alternatively, simulation studies allow for quantifying how much a specific risk factor needs to be managed for a specific cluster to prevent eGFR decline. For example, when MAP is reduced by 5 mmHg (Figure 6b), the mean ΔeGFR for cluster-1 becomes ~ 0 whereas the mean ΔeGFR for other clusters does not change much.

DISCUSSIONS

Insights

A major part of the semimechanistic model is defined in a biological causal fashion using the QSP approach and feedforward neural networks supplement the parts with not well-established biology (e.g., variations in blood pressure data). Alternatively, the GFR decline rate could be

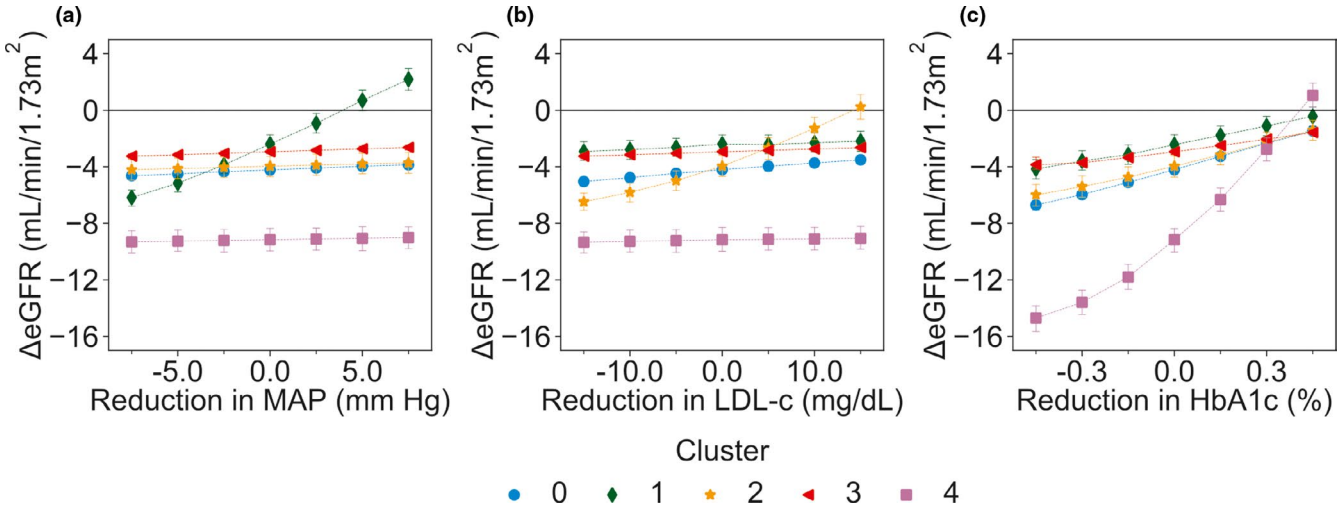

FIGURE 6 Results of simulation studies. Mean ΔeGFR (quantifying GFR decline over 4 years) for each cluster for different simulation scenarios: (a) Reduction in MAP, (b) reduction in LDL-C, and (c) reduction in HbA1c. The different clusters, which are sensitive to specific risk factors, are indicated in different colors and symbols. Cluster 0 consists of patients who are insensitive to the modelled risk factors, whereas clusters 1, 2, 3, and 4 consist of patients whose CKD progression is driven by hypertension, hyperlipidemia, impaired eGFR, and hyperglycemia, respectively. The error bars indicate the 95% confidence interval estimate for mean ΔeGFR of each cluster. Conversion factors for units: LDL-C in mg/dL to μmol/L, × 0.02586

satisfactorily explained with a weighted additive contribution of its risk factors, which suggests that a simpler linear model is appropriate to explain serum creatinine/eGFR dynamics. Moreover, the model's linear nature offered clear physiological interpretability of the parameters.

Our methodology incorporates the effects of different medication classes on appropriate physiological pathways taking into account appropriate mechanism of action and other drug-specific information (dose-response curves from literature). By looking retrospectively at the estimated medication adherence, this feature allows in assessing patients' response to a drug with realistic adherence.

The model's mechanistic parameters have a clear physiological meaning and clustering on patient-specific parameters corresponding to CKD progression segmented the patient cohort into groups with similar progression characteristics. Treatment strategies will vary for the patients belonging to the different clusters; the different clusters characterized as sensitive to specific risk factors will benefit through control of the corresponding risk factors, as shown through simulations.

The analysis also revealed a cluster in which CKD progression is driven by feedback of GFR to itself indicating a disease progression pathway related to autocrine effect, wherein control of comorbidities would not help the patients. Further investigation is required to understand this pathway. The analysis also revealed a cluster insensitive to any of the risk factors suggesting that there are patients whose GFR decline is very gradual even though their risk factor values may be high. Such patients may have other unmeasured mechanisms slowing down their progression and require further investigation. We also note that the identified clusters are characterized by disease progression driven largely by a dominant risk factor. This does not imply that adverse changes in other risk factors will not have any effect, rather the maximal effect will be due to the dominant risk factor.

Limitations

First, the model does not attempt to incorporate exhaustively all the risk factors that contribute to CKD progression but the most important ones regularly captured in EHR. Any attempts to make the model more extensive by adding more mechanisms will increase the number of parameters to be estimated, which in turn would require a more comprehensive data coverage. For example, physical activities apart from affecting body weight can affect insulin sensitivity, glucose, blood pressure, and lipids; however, the data associated with physical activities and exercise were not available in our EHR data and are rarely captured in a traditional healthcare setting. Likewise, our model is able to discriminate among the different patients with distinct disease progression characteristics; however, the same risk factor (e.g., hypertension) can cause CKD progression through different pathways. The development of a model that can discern between different pathways requires additional data typically not captured in the standard healthcare setting. A "big data" approach incorporating more biomarkers, dense measurements, and metabolomics can allow such analysis.

The effects of medications are currently considered to be independent of each other. Combination therapy is considered as a simple simultaneous effect of the individual drugs, which is not universally true. Incorporating exact responses of combination therapy is harder due to a large number of combinations and sparsity in available data for many combinations. The effects of drugs outside the direction mechanism or biomarker are also not incorporated. For example, the reported increase in fasting glucose due to high-dose statins is not incorporated.[49] There is plenty of evidence for a patient being on multiple drugs with different mechanisms of action for the same disease, like diabetes; interactions between drug classes are not included.

Currently, the model requires sufficient longitudinal data density for cluster assignment and it would be beneficial to perform this with minimal follow-up data to expedite treatments based on model's insights. Ideally, it is desirable to assign clusters using baseline data only; however, the model currently needs at least one follow-up measurement to estimate patient-specific parameters as they depend on the derivative of GFR.

The large changes in risk factors used in our simulation studies are often unrealistic due to other side effects or even adverse effects. For example, a large reduction in HbA1c in patients with diabetes increases risk of hypoglycemic shock. The model does not consider these other effects that a clinician needs to consider when choosing to manage a risk factor.[50]

ACKNOWLEDGEMENTS

The electronic health record data for this study was obtained from Decision Resources Group (https://decisionresourcesgroup.com). The authors thank Anisha Balani for the extraction of drug-efficacy relationships from published literature.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

R.R., S.N.W., K.G., J.L., R.B., A.G., and J.S. wrote the manuscript. J.L., R.B., J.S., A.G., and R.R. designed the

research. S.N.W., K.G., and R.R. performed the research. S.N.W. and K.G. analyzed the data.

REFERENCES

1. Bikbov B, Perico N, Remuzzi G. Disparities in chronic kidney disease prevalence among males and females in 195 countries: analysis of the global burden of disease 2016 study. *Nephron*. 2018;139:313-318.
2. Bikbov B, Purcell CA, Levey AS, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2020;395:709-733.
3. Klag MJ, Whelton PK, Randall BL, et al. Blood pressure and end-stage renal disease in men. *N Engl J Med*. 1996;334:13-18.
4. Muntner P, Coresh J, Smith JC, Eckfeldt J, Klag MJ. Plasma lipids and risk of developing renal dysfunction: the atherosclerosis risk in communities study. *Kidney Int*. 2000;58:293-301.
5. de Boer IH. Temporal trends in the prevalence of diabetic kidney disease in the United States. *JAMA*. 2011;305:2532-2539.
6. Anders H-J, Huber TB, Isermann B, Schiffer M. CKD in diabetes: diabetic kidney disease versus nondiabetic kidney disease. *Nat Rev Nephrol*. 2018;14:361-377.
7. Remuzzi G, Benigni A, Remuzzi A. Mechanisms of progression and regression of renal lesions of chronic nephropathies and diabetes. *J Clin Invest*. 2006;116:288-296.
8. Ku E, Lee BJ, Wei J, Weir MR. Hypertension in CKD: core curriculum 2019. *Am J Kidney Dis*. 2019;74:120-131.
9. Afkarian M, Sachs MC, Kestenbaum B, et al. Kidney disease and increased mortality risk in type 2 diabetes. *J Am Soc Nephrol*. 2013;24:302–308.
10. Chan CM. Hyperlipidaemia in chronic kidney disease. *Ann Acad Med Singapore*. 2005;34:31-35.
11. Schnaper HW. Remnant nephron physiology and the progression of chronic kidney disease. *Pediatr Nephrol*. 2014; 29:193-202.
12. Mennuni S, Rubattu S, Pierelli G, Tocci G, Fofi C, Volpe M. Hypertension and kidneys: unraveling complex molecular mechanisms underlying hypertensive renal damage. *J Hum Hypertens*. 2014;28:74-79.
13. Udani S, Lazich I, Bakris GL. Epidemiology of hypertensive kidney disease. *Nat Rev Nephrol*. 2011;7:11-21.
14. Li L, Astor BC, Lewis J, et al. Longitudinal progression trajectory of GFR among patients with CKD. *Am J Kidney Dis*. 2012;59:504-512.
15. Tsai C-W, Ting IW, Yeh H-C, Kuo C-C. Longitudinal change in estimated GFR among CKD patients: a 10-year follow-up study of an integrated kidney disease care program in Taiwan. *PLoS One*. 2017;12:e0173843.
16. Go AS, Yang J, Tan TC, et al. Contemporary rates and predictors of fast progression of chronic kidney disease in adults with and without diabetes mellitus. *BMC Nephrol*. 2018;19:146.
17. Keane WF, Zhang Z, Lyle PA, et al. Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. *Clin J Am Soc Nephrol*. 2006;1:761-767.
18. Tangri N. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA*. 2011;305:1553-1559.
19. Tangri N, Inker LA, Hiebert B, et al. A dynamic predictive model for progression of CKD. *Am J Kidney Dis*. 2017;69:514-520.
20. Ramspek CL, de Jong Y, Dekker FW, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrol Dial Transplant*. 2019;35:1527-1538.
21. Bao Y-W, Yuan Y, Chen J-H, Lin W-Q. Kidney disease models: tools to identify mechanisms and potential therapeutic targets. *Zool Res*. 2018;39:72.
22. Becker GJ, Hewitson TD. Animal models of chronic kidney disease: useful but not perfect. *Nephrol Dial Transplant*. 2013;28:2432-2438.
23. Hallow K, Gebremichael Y. A quantitative systems physiology model of renal function and blood pressure regulation: model description. *CPT Pharmacometrics Syst Pharmacol*. 2017;6:383-392.
24. Hallow KM, Greasley PJ, Helmlinger G, Chu L, Heerspink HJ, Boulton DW. Evaluation of renal and cardiovascular protection mechanisms of SGLT2 inhibitors: model-based analysis of clinical data. *Am J Physiol Renal Physiol*. 2018;315:F1295-F1306.
25. Musante C, Ramanujan S, Schmidt BJ, Ghobrial OG, Lu J, Heatherington AC. Quantitative systems pharmacology: a case for disease models. *Clin Pharmacol Ther*. 2017;101:24-27.
26. Hsu C-Y, Chinchilli VM, Coca S, et al. Post-acute kidney injury proteinuria and subsequent kidney disease progression: the assessment, serial evaluation, and subsequent sequelae in acute kidney injury (ASSESS-AKI) study. *JAMA Intern Med*. 2020;180:402-410.
27. Vaidya VS, Ferguson MA, Bonventre JV. Biomarkers of acute kidney injury. *Annu Rev Pharmacol Toxicol*. 2008;48:463-493.
28. Chow CC, Hall KD. The dynamics of human body weight change. *PLoS Comput Biol*. 2008;4:e1000045.
29. Hall KD. Computational model of in vivo human energy metabolism during semistarvation and refeeding. *Am J Physiol Endocrinol Metab*. 2006;291:E23-E37.
30. Winter WD, DeJongh J, Post T, et al. A mechanism-based disease progression model for comparison of long-term effects of pioglitazone, metformin and gliclazide on disease processes underlying type 2 diabetes mellitus. *J Pharmacokinet Pharmacodyn*. 2006;33:313-343.
31. König M, Bulik S, Holzhütter H-G. Quantifying the contribution of the liver to glucose homeostasis: a detailed kinetic model of human hepatic glucose metabolism. *PLoS Comput Biol*. 2012;8:e1002577.
32. Tura A, Ludvik B, Nolan JJ, Pacini G, Thomaseth K. Insulin and C-peptide secretion and kinetics in humans: direct and model-based measurements during OGTT. *Am J Physiol Endocrinol Metab*. 2001;281:E966-E974.
33. Gilbert ER, Liu D. Epigenetics: the missing link to understanding β-cell dysfunction in the pathogenesis of type 2 diabetes. *Epigenetics*. 2012;7:841-852.
34. Gaetano AD, Hardy T, Beck B, et al. Mathematical models of diabetes progression. *Am J Physiol Endocrinol Metab*. 2008;295:E1 462-E1479.
35. Sears B, Perry M. The role of fatty acids in insulin resistance. *Lipids Health Dis*. 2015;14:121.
36. Davidson MH, Robinson JG. Lipid-lowering effects of statins: a comparative review. *Expert Opin Pharmacother*. 2006;7:1701-1714.
37. Hall JE, Guyton AC. *Guyton and Hall Textbook of Medical Physiology*, 12th edn. Saunders/Elsevier; 2011.
38. Sun Z. Aging, arterial stiffness, and hypertension. *Hypertension*. 2015;65:252-256.

39. Marion RW, John EH. Pathophysiology and treatment of obesity hypertension. *Curr Pharm Des*. 2004;10:3621-3637.
40. Lao D, Parasher PS, Cho KC, Yeghiazarians Y. Atherosclerotic renal artery stenosis - diagnosis and treatment. *Mayo Clin Proc*. 2011;86:649-657.
41. Virtanen P, Gommers R, Oliphant TE. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
42. Hindmarsh, AC ODEPACK, a systematized collection of ODE solvers. In: Stepleman, RS, et al., *Scientific Computing*. North-Holland, Amsterdam, 1983: 55-64.
43. Storn R, Price K. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J Global Optim*. 1997;11:341-359.
44. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
45. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun. Stat. Theory Methods*. 1974;3:1-27.
46. Jaccard P. The distribution of the flora of the alpine zone. *New Phytol*. 1912;11:37-50.
47. Cramér H. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press; 1999.
48. Cohen J. The t test for means. In: Cohen, J, eds. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press; 1977: 19-74.
49. Kim J, Lee HS, Lee K-Y. Effect of statins on fasting glucose in non-diabetic individuals: nationwide population-based health examination in Korea. *Cardiovasc Diabetol*. 2018;17:155.
50. Sauer AJ, Cole R, Jensen BC, et al. Practical guidance on the use of sacubitril/valsartan for heart failure. *Heart Fail Rev*. 2019;24:167-176.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ramaswamy R, Wee SN, George K, et al. CKD subpopulations defined by risk-factors: A longitudinal analysis of electronic health records. *CPT Pharmacometrics Syst Pharmacol*. 2021;10:1343–1356. https://doi.org/10.1002/psp4.12695