Projecting the Epidemiological and Economic Impact of Chronic Kidney Disease Using Patient-Level Microsimulation Modelling: Rationale and Methods of *Inside CKD*

Navdeep Tangri · Steven Chadban · Claudia Cabrera · Lise Retat · Juan José García Sánchez

N. Tangri
University of Manitoba, Winnipeg, MB R3T 2N2, Canada

S. Chadban
Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia

C. Cabrera
Real World Science and Analytics, BioPharmaceuticals Medical, AstraZeneca, Gothenburg, SE-431 83, Sweden

L. Retat
HealthLumen Limited, London, EC3N 2PJ, UK

J. J. G. Sánchez
Global Market Access and Pricing, BioPharmaceuticals, AstraZeneca, Cambridge, CB2 0AA, UK.

Corresponding Author:
J. J. G. Sánchez
Email: juanjose.garciasanchez@astrazeneca.com

SUPPLEMENTARY MATERIALS

Appendix S1 Extended methodology for determining regression analysis

A linear mixed-effects regression analysis was fitted to all estimated glomerular filtration rate (eGFR) measurements within the study period, irrespective of the number of measurements per patient and the timings of these measurements. The linear mixed-effects model was fitted to eGFR measurements in terms of time in years and baseline covariates. Each patient was given their own distinct eGFR trajectory over time, by including a random patient-level intercept and a random patient-level effect of time. The effect of missing data was coded as 1 (missing) versus 0 (not missing) as an indicator (missing data indicator) and accounted for in the modelling. Within the microsimulation, the decline rate for each individual was based on the generated β coefficients.

Patients were grouped into slow versus rapid progressors based on their unadjusted eGFR trajectories. The cut-offs were based on published literature (e.g., rapid progressor with eGFR > 4 mL/minute/1.73 m^2). eGFR measures of less than 5 mL/minute/1.73 m^2 were excluded (eGFR only, not the patient). To produce the coefficients presented in Table 5, an unadjusted linear mixed-effects regression model was run overall, for patients grouped into normal[a] eGFR decline, for patients grouped into rapid eGFR decline, and for patients with each of the following:

- overall[b]
- type 2 diabetes (T2D) at baseline versus no T2D at baseline[b]
- heart failure (HF) at baseline versus no HF at baseline[b]
- T2D + HF at baseline versus no T2D + HF at baseline[b]
- hypertension (HTN) at baseline versus no HTN at baseline[b]
    - overall[c]
    - controlled (defined as blood pressure (BP) < 140/90 with/without therapy [renin–angiotensin–aldosterone system inhibitors, calcium channel blockers, β blockers, α blockers, diuretics, and mineralocorticoid receptor antagonists])[c]
    - uncontrolled (defined as BP > 140/90 with/without therapy [renin–angiotensin–aldosterone system inhibitors, calcium channel blockers, β blockers, α blockers, diuretics, and mineralocorticoid receptor antagonists])[e]
- patients without T2D, HF, and HTN at baseline[b].

[a]Based on the overall slope analysis, patients were allocated to normal and rapid progressor groups and mean slopes within groups were calculated based on the overall slope analysis.

[b]Patients were allocated to these specific groups based on the presence or absence of comorbidities at baseline and mean slopes were calculated based on the overall slope analysis.

[c]Defined as BP and/or HTN code and/or therapy. Those with missing BP values could not be included in the controlled/uncontrolled groups.

Appendix S2 Data sources for input parameters

Parameter	Example sources	Year
Population (total population, age and sex break down, population growth rate)	United Nations population statistics, National Population Database	Year 2020 and projections
CKD prevalence by stage (eGFR/albuminuria)	National dataset or published sources e.g., NHANES, HSE, CPRD-HES linked	Most recent 5+ years of data to plot trends
Prevalence and relative/absolute risk for impact of CKD and other comorbidities on study outcomes	Literature review	Most recent/robust
Prevalence and relative/absolute risk of complications (stroke, MI, heart failure) on study outcomes	Literature review	Most recent/robust
Renal replacement therapy data	National Renal Registry data	Most recent/robust
Direct and indirect health costs (including care burden, cost of screening)	Literature review	Most recent/robust
Health state utility weights	Literature review	Most recent/robust
Screening intervention data	Literature review	Most recent/robust

CKD chronic kidney disease, *CPRD* Clinical Practice Research Datalink, *eGFR* estimated glomerular filtration rate, *HSE* Health Survey England, *MI* myocardial infarction, *NHANES* National Health and Nutrition Examination Survey

Appendix S3 Microsimulation model framework

The HealthLumen software consists of two models. The first model is a sophisticated regression model that calculates the predictions of risk factor trends over time based on data from rolling cross-sectional studies. The second model performs the microsimulation of a virtual population, generated with demographic characteristics matching those of the observed data. The health trajectory of each individual from the population is simulated over time allowing them to contract, survive, or die from a set of diseases or injuries related to the analysed risk factors. A detailed description of the two models is presented below.

Model One: Predictions of Risk Factors Over Time

A logistic multinomial regression was applied to eGFR and albumin distributions to obtain regression coefficients using the following assumptions.

For the risk factor, let N be the number of categories for a given risk factor, e.g., $N = 3$ for albuminuria and $N = 6$ for eGFR. Let $k = 1, 2, \ldots, N$ number these categories and $p_k(t)$ denote the prevalence of individuals with risk factor values that correspond to the category k at time t. We estimate $p_k(t)$ using a multinomial logistic regression model with prevalence of risk factor category k as the outcome and time t as a single explanatory variable. For $k < N$, we have:

$$\ln\left(\frac{p_k(t)}{p_1(t)}\right) = \beta_0^k + \beta_1^k t \qquad (1.1)$$

The prevalence of the first category is obtained by using the normalization constraint $\sum_{k=1}^{N} p_k(t) = 1$. Solving equation (1.1) for $p_k(t)$, we obtain:

$$p_k(t) = \frac{\exp\left(\beta_0^k + \beta_1^k t\right)}{1 + \sum_{k'=1}^{N} \exp\left(\beta_0^{k'} + \beta_1^{k'} t\right)}, \qquad (1.2)$$

which respects all constraints on the prevalence values, i.e., normalization and [0, 1] bounds.

Measured data consist of sets of probabilities, with their variances, at specific time values (typically the year of the survey). For any particular time, the sum of these probabilities is unity. Each data point is treated as a normally distributed random variable; together they are a set of N groups (number of years) of K probabilities $[63K - 1]] \mid i \in 0, N - 1]\}$.

For each year the set of K probabilities form a distribution and their sum is equal to unity.
The regression consists of fitting a set of logistic functions $[28K - 1]]$ to these data – one function for each k-value. At each time value, the sum of these functions is unity. Thus, for example, when measuring albuminuria in the three states already mentioned, the $k = 0$ regression function represents the probability of no albuminuria (A1) over time, $k = 1$ the probability of microalbuminuria (A2), and $k = 2$ the probability of macroalbuminuria (A3).

The regression equations are most easily derived from a familiar least-squares minimization. In the following equation set, the weighted difference between the measured and predicted probabilities is written as S; the logistic regression functions $p_k(\mathbf{a},\mathbf{b};t)$ are chosen to be ratios of sums of exponentials (this is equivalent to modelling the log probability ratios, p_k/p_0, as linear functions of time).

$$S(\mathbf{a},\mathbf{b}) = \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \qquad (1.3)$$

$$p_k(\mathbf{a},\mathbf{b},t) \equiv \frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}$$

$$\mathbf{a} \equiv (a_0,a_1,..,a_{K-1}), \quad \mathbf{b} \equiv (b_0,b_1,..,b_{K\text{-}1})$$

$$A_0 \equiv 0, \quad A_k \equiv a_k + b_k t \qquad (1.4)$$

The parameters A_0, a_0, and b_0 are all zero and are used merely to preserve the symmetry of the expressions and their manipulation. For a K-dimensional set of probabilities there will be $2(K-1)$ regression parameters to be determined.

For a given dimension K, there are $K-1$ independent functions p_k – the remaining function being determined from the requirement that a complete set of K forms a distribution and sum to unity. Note that the parameterization ensures the necessary requirement that each p_k be interpretable as a probability – a real number lying between 0 and 1.

The minimum of the function S is determined from the equations:

$$\frac{\partial S}{\partial a_j} = \frac{\partial S}{\partial b_j} = 0 \qquad \text{for j=1,2,....,k-1} \qquad (1.5)$$

noting the relations:

$$\frac{\partial p_k}{\partial A_j} = \frac{\partial}{\partial A_j}\left(\frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}\right) = p_k\delta_{kj} - p_k p_j$$

$$\frac{\partial}{\partial a_j} = \frac{\partial}{\partial A_j}$$

$$\frac{\partial}{\partial b_j} = t\frac{\partial}{\partial A_j} \qquad (1.6)$$

The values of the vectors **a**, **b** that satisfy these equations are denoted $\hat{\mathbf{a}},\hat{\mathbf{b}}$. They provide the trend lines $p_k(\hat{\mathbf{a}},\hat{\mathbf{b}};t)$, for the separate probabilities.

The regression equations are most easily derived from a familiar least-squares minimization. In the following equation set the weighted difference between the measured and predicted probabilities is written as S; the logistic regression functions $p_k(\mathbf{a},\mathbf{b};t)$ are chosen to be ratios of sums of exponentials (this is equivalent to modelling the log probability ratios, p_k/p_0, as linear functions of time).

$$S(\mathbf{a},\mathbf{b}) = \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \qquad (1.7)$$

$$p_k(\mathbf{a},\mathbf{b},t) \equiv \frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}$$

$$\mathbf{a} \equiv (a_0,a_1,..,a_{K-1}), \quad \mathbf{b} \equiv (b_0,b_1,..,b_{K-1}) \tag{1.8}$$

$$A_0 \equiv 0, \quad A_k \equiv a_k + b_k t$$

The parameters A_0, a_0, and b_0 are all zero and are used merely to preserve the symmetry of the expressions and their manipulation. For a K-dimensional set of probabilities there will be $2(K-1)$ regression parameters to be determined.

For a given dimension K, there are $K-1$ independent functions p_k – the remaining function being determined from the requirement that a complete set of K forms a distribution and sum to unity. Note that the parameterization ensures the necessary requirement that each p_k be interpretable as a probability – a real number lying between 0 and 1.

The minimum of the function S is determined from the equations:

$$\frac{\partial S}{\partial a_j} = \frac{\partial S}{\partial b_j} = 0 \qquad \text{for j=1,2,....,k-1} \tag{1.9}$$

noting the relations.

$$\frac{\partial p_k}{\partial A_j} = \frac{\partial}{\partial A_j}\left(\frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}\right) = p_k\delta_{kj} - p_k p_j$$

$$\frac{\partial}{\partial a_j} = \frac{\partial}{\partial A_j} \tag{1.10}$$

$$\frac{\partial}{\partial b_j} = t\frac{\partial}{\partial A_j}$$

The values of the vectors **a**, **b** that satisfy these equations are denoted $\hat{\mathbf{a}},\hat{\mathbf{b}}$. They provide the trend lines, $p_k(\hat{\mathbf{a}},\hat{\mathbf{b}};t)$, for the separate probabilities.

Bayesian Interpretation

The $2K-2$ regression parameters are regarded as random variables of which posterior distribution is proportional to the function $\exp(-S(\mathbf{a},\mathbf{b}))$. The maximum likelihood estimate of this probability distribution function, the minimum of the function S, is obtained at the values $\hat{\mathbf{a}},\hat{\mathbf{b}}$. Other properties of the $(2K-2)$-dimensional probability distribution function are obtained by first approximating it as a $(2K-2)$-dimensional normal distribution, the mean of which is the maximum likelihood estimate. This amounts to expanding the function $S(\mathbf{a},\mathbf{b})$ in a Taylor series as far as terms quadratic in the differences $(\mathbf{a}-\hat{\mathbf{a}}),(\mathbf{b}-\hat{\mathbf{b}})$ about the maximum likelihood estimate $\hat{\mathbf{S}} \equiv S(\hat{\mathbf{a}},\hat{\mathbf{b}})$. Hence:

$$
\begin{aligned}
S(\mathbf{a},\mathbf{b}) &= \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \\
&\equiv S\left(\hat{a},\hat{b}\right)+\tfrac{1}{2}\left(a-\hat{a},b-\hat{b}\right)P^{-1}\left(a-\hat{a},b-\hat{b}\right)+\ldots \\
&\approx S\left(\hat{a},\hat{b}\right)+\tfrac{1}{2}\sum_{i,j}\left(a_i-\hat{a}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{a}_j}\left(a_j-\hat{a}_j\right)+\tfrac{1}{2}\sum_{i,j}\left(a_i-\hat{a}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{b}_j}\left(b_j-\hat{b}_j\right)+ \\
&+\tfrac{1}{2}\sum_{i,j}\left(b_i-\hat{b}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{a}_j}\left(a_j-\hat{a}_j\right)+\tfrac{1}{2}\sum_{i,j}\left(b_i-\hat{b}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{b}_j}\left(b_j-\hat{b}_j\right)
\end{aligned}
\tag{1.11}
$$

The $(2K - 2)$-dimensional covariance matrix P is the inverse of the appropriate expansion coefficients. This matrix is central to the construction of the confidence limits for the trend lines.

Model Two: Microsimulation Model

Simulated people are generated with the correct demographic statistics in the simulation's start year. In this year, females are stochastically allocated a number of children and the birth dates of their children – these are generated from known fertility and mothers' age at birth statistics (valid in the start year). If a female has children, those children are generated as members of the simulation in the appropriate birth year. The microsimulation is provided with a list of relevant diseases. These diseases are allocated using the best available data regarding incidence, mortality, survival, relative risk, and prevalence statistics (by age and gender). The virtual population is initialized with diseases by simulating each individual from birth until the start year of the model simulation. It assumes that a person can die before the model start year. It is assumed that at initialization the diseases are independent random variables. In the course of their lives, simulated people can die from one of the diseases caused by a particular risk factor that they might have acquired or from some other cause. The probability that a person of a given age and gender dies from a cause other than the disease is calculated in terms of known death and disease statistics valid in the start year. This rate is constant over the course of the simulation. The survival rates from the risk factor-related diseases will change as a consequence of the changing distribution of the risk factor in the population.

The microsimulation incorporates a sophisticated economic module. The module employs Markov-type simulation of long-term health benefits, healthcare costs, and non-healthcare-related costs of specified interventions. It synthesizes and estimates evidence for the cost-utility analysis. The model is used to project the differences in quality-adjusted life-years (QALYs), lifetime healthcare costs, premature mortality costs, and indirect costs as a consequence of interventions over a specified timescale. Outputs can be discounted for any specific discount rate.

The confidence limits that accompany the sets of output data represent the accuracy of the microsimulation (stochastic or aleatoric uncertainty) as opposed to the confidence of the input data itself (parameter uncertainty). Errors around the input data were not available.

Population Module

The population module contains several data sets that can be edited by the end user through a user interface. The population is created in the start year and propagated forward in time by allowing females to give birth, and population projections can be incorporated (i.e., migration through

minimum arrivals and departures). People within the model can die from specific diseases or from other causes. The <deaths by year by sex by age> file is a necessary input to the model when valid population projections are being used in the start year and is usually referred to as the 'deaths from all causes' file. This module is flexible and allows the user to run open and closed populations with no births.

Distributions

Distribution name	**Symbol**	**Note**
MalesByAgeByYear	$p_m(a)$	Input in year$_0$ – probability of a male having age *a*
FemalesByAgeByYear	$p_f(a)$	Input in year$_0$ – probability of a female having age *a*
BirthsByAgeofMother	$p_b(a)$	Input in year$_0$ – conditional probability of a birth at age *a*/the mother gives birth
NumberOfBirths	$p_l(n)$	λ≡TFR, Poisson distribution, probability of giving birth to *n* children
MaleDeathByAge	$p_{Wm}(a)$	Input in year$_0$, probability of a male dying at age *a*
FemaleDeathByAge	$p_{Wf}(a)$	Input in year$_0$, probability of a female dying at age *a*

TFR total fertility rate

Birth Model

Any female of childbearing age is deemed capable of giving birth. The number of children, *n*, that she has in her life is dictated by the Poisson distribution $p_l(n)$ where the mean of the Poisson distribution is the total fertility rate (TFR) parameter.[d]

[d]This could be made to be time dependent; in the baseline model it is constant.

The probability that a mother (who does give birth) gives birth to a child at age *a* is determined from the BirthsByAgeOfMother distribution as $p_b(a)$. For any particular mother, the births of multiple children are treated as independent events, so that the probability that a mother who produces *N* children produces *n* of them at age *a* is given as the binomially distributed variable,

$$p_b\left(n\,at\,a \mid N\right) = \frac{N!}{n!\left(N-n\right)!}\left(p_b\left(a\right)\right)^n\left(1-p_{bm}\left(a\right)\right)^{N-n} \tag{1.15}$$

The probability that the mother gives birth to *n* children at age *a* is

$$p_b\left(n\,at\,a\right) = e^{-\lambda}\sum_{N=n}^{\infty}\frac{\lambda^N}{N!}p_b\left(n\,at\,a \mid N\right) = e^{-\lambda}\sum_{N=n}^{\infty}\frac{\lambda^N}{n!\left(N-n\right)!}\left(p_b\left(a\right)\right)^n\left(1-p_b\left(a\right)\right)^{N-n} \tag{1.16}$$

Performing the summation in this equation gives the simplifying result that the probability $p_b(n \text{ at } a)$ is itself Poisson distributed with mean parameter $\lambda p_b(a)$,

$$p_b\left(n\,at\,a\right) = e^{-\lambda p_b(a)}\frac{\left(\lambda p_b\left(a\right)\right)^n}{n!} = p_{\lambda p_b(a)}\left(n\right) \tag{1.17}$$

Thus, on average, a mother at age a will produce $\lambda p_b(a)$ children in that year. The gender of the children is determined by the probability $p_{male}=1-p_{female}$. In the baseline model, this is taken to be the probability $N_m/(N_m+N_f)$. The probability of child gender can be made time dependent.

The 'Population Editor' menu item Population Editor\Tools\Births\Show Random Birth list creates an instance of the *T*Population class and uses it to generate and list a (selectable) sample of mothers and the years in which they give birth.

Time-Dependent Birth Rates

The TFR parameter for future years can be input from the input file if known – or otherwise modelled. In general, the TFR parameter is kept constant over time. In each year of their simulated life (*y* at age *a*), mothers of childbearing age can use the appropriate Poisson parameter $\lambda(a)p_b(a)$ to generate the number of children in that year. Each child is recorded in the mother's 'Life Event' list and processed as part of the current family at the end of the mother's life.

Population Dynamics

In one year, *Y*, the population will consist of N_m males and N_f females with their respective age distributions. In the next year, *Y*', the numbers will have been depleted by deaths and augmented by the $N_{newborn}$ births. The new, primed population is determined from the old population by the following equation set:

$$N_{newborn} = \lambda N_f \sum_{a=AgeAtChild.lo}^{a=AgeAtChild.hi} p_f\left(a\right)\left(1-p_f\left(a\right)\right)p_b\left(a\right) \tag{1.18}$$

$$N'_m = N_m \sum_{a=1}^{a=Age.hi} p_m\left(a\right)\left(1-p_m\left(a\right)\right) + p_{male}N_{newborn} \tag{1.19}$$

$$N'_f = N_f \sum_{a=1}^{a=Age.hi} p_f(a)\left(1 - p_f(a)\right) + p_{female} N_{newborn} \tag{1.20}$$

$$p'_m(a+1) = \frac{N_m}{N'_m} p_m(a)\left(1 - p_m(a)\right) \tag{1.21}$$

$$p'_m(a+1) = \frac{N_m}{N'_m} p_m(a)\left(1 - p_{\Omega m}(a)\right) \tag{1.22}$$

$$p'_f(a+1) = \frac{N_f}{N'_f} p_f(a)\left(1 - p_{\Omega f}(a)\right) \tag{1.23}$$

$$p'_m(0) = \frac{1}{N'_m} p_{male} N_{newborn} \tag{1.24}$$

$$p'_f(0) = \frac{1}{N'_f} p_{female} N_{newborn} \tag{1.25}$$

The 'Population Editor' menu item Population Editor\View\Population Dynamics\Male implements these equations and draws projected populations year by year.

Deaths From Modelled Diseases

The simulation models any number of specified diseases, some of which may be fatal. In the start year the simulation's death model uses the mortality statistics of diseases to adjust the probabilities of death by age and gender. In the start year the net effect is to maintain the same probability of death by age and gender as before; in subsequent years, however, the rates at which people die from modelled diseases will change as modelled risk factors change. The population dynamics outlined above will be only an approximation of the simulated population's dynamics. The latter will be known only on completion of the simulation.

Multiple Population Processing

Multiple populations can be used in a simulation provided they are non-overlapping (people cannot belong to both).

In a simulation, Monte Carlo trials are allocated between different current populations in proportion to their total person count (malesCount + femalesCount). The idea being to provide a representative sample of the combined population. In a simulation, a population (pop) is current if the simulated year Y satisfies

$$pop \rightarrow startYear \leq Y \leq pop \rightarrow stopYear \tag{1.26}$$

Open Populations

This model is an open population model that allows people to enter and to depart from the population (departure probability $p_\delta(t)$).

Open Population, Births, and Deaths

In the year y, the number of males and females in the population are denoted as $\{N_m(a,y), N_f(a,y)\}$, and we suppose that they have departure probabilities $\{p_{m\delta}(a,y), p_{f\delta}(a,y)\}$.The number of new arrivals into each age in the year Y are denoted $\{N_{mArr}(a,y), N_{fArr}(a,y)\}$. The following analysis applies equally to males and females, and we drop the gender suffix. The male and female populations grow according to the recursion relations:

$$N(a+1,y+1) = N(a,y)\left(1-p_{\Omega}(a)\right)\left(1-p_{\delta}(a,y)\right) + N_{Arr}(a,y) \quad (a>1) \tag{1.27}$$

$$N(1,y+1) = N_{Newborn}(y)\left(1-p_{\Omega}(0)\right)\left(1-p_{\delta}(0,y)\right) + N_{Arr}(0,y) \quad (a=0) \tag{1.28}$$

The Longitudinal Modelling of Populations Having Known Cross-Sectional Data

Given a set of X-sectional population projections $\{K_m(a,y), K_f(a,y) | 0 <= a <= 100;\ Y_0 <= y <= Y_1\}$ (the K-population) the question arises of how to model the lives of individuals within the population (the N-population). In the absence of precise arrival (immigration) and departure (emigration) statistics, many solutions exist. The population is constructed iteratively: given the population in year Y the population of the next year is calculated from the known birth and death rates. Thus, the departure probabilities and arrival numbers are found by matching with the projected K-population.

Minimum Arrival and Departure Model

The minimum arrival and departure model fixes the modelled N-population in the start year and compensates in subsequent years either by having non-zero departure statistics (if $N > K$) or by importing new people ($K > N$).

From equation (1.27):

$$if\ N(a,y)\left(1-p_{\Omega}(a)\right) > K(a+1,y+1)$$

$$\left(1-p_{\delta}(a,y)\right) = \frac{K(a+1,y+1)}{N(a,y)\left(1-p_{\Omega}(a)\right)} \quad (a>1)$$

$$\Rightarrow$$

$$N(a+1,y+1) = N(a,y)\left(1-p_{\Omega}(a)\right)\left(1-p_{\delta}(a,y)\right) = K(a+1,y+1) \quad (a>1) \tag{1.29}$$

$$if\ N(a,y)\left(1-p_{\Omega}(a)\right) < K(a+1,y+1)$$

$$N_{Arr}(a,y) = K(a+1,y+1) - N(a,y)\left(1-p_{\Omega}(a)\right) \quad (a>1)$$

$$\Rightarrow$$

$$N(a+1,y+1) = N(a,y)\left(1-p_{\Omega}(a)\right) + N_{Arr}(a,y) = K(a+1,y+1) \tag{1.30}$$

The implementation of this model can be arranged using multiple populations – one population for each year of the simulation. The first population consists of the baseline model that matches the *N*- and *K*-populations in the start year; subsequent populations contain the corrections (the arrivals, if any in that year). When arrivals enter the simulated population, they have a start year corresponding to this population's start year. They usually will have been modelled from birth in the appropriate risk and disease environment. Arrivals are ordinary members of the modelled population – they simply enter the population at times after the simulation start time. Arrivals carry with them a population identifier. The numbers of males and females and their ages are known for all populations. Within the microsimulation multiple populations are sampled at a rate proportional to their population size.

Risk Factor Module

The distribution of risk factors in the population is estimated using regression analysis stratified by both sex S = {male, female} and age group A = {0–9, 10–19, ..., 70–79, 80+}. The fitted trends are extrapolated to forecast the distribution of each risk factor category in the future. For each sex-and-age-group stratum, the set of cross-sectional, time-dependent, discrete distributions $D = \{p_k(t)|k = 1, \dots N;\ t > 0\}$ is used to manufacture risk factor trends for individual members of the population. We model the urine–albumin and eGFR risk factors continuously.

In the case of a continuous risk factor, for each discrete distribution D there is a continuous counterpart. Let β denote the risk factor value in the continuous scale and let $f(\beta|A,S,t)$ be the probability density function of β for age group A and sex S at time t. Then:

$$p_k(t|A,S) = \int_{\beta \in k} f(\beta|A,S,t)\, d\beta. \tag{1.31}$$

Equations (1.2) and (1.31) both refer to the same quantity. However, equation (1.31) uses the definition of the probability density function to express the age- and sex-specific percentage of individuals in risk factor category k at time t. Equation (1.2) gives an estimate of this quantity using equation (1.1) for all $k = 0, \dots, N$. The cumulative distribution function of β is:

$$F(\beta|A,S,t) = \int_0^{\beta} f(\beta|A,S,t)\, d\beta. \tag{1.32}$$

At time t, a person with sex S belonging to the age group A is said to be on the p–th percentile of this distribution if $F(\beta|A,S,t) = p/100$. Given the cross-sectional information from the set of distributions D, it is possible to simulate longitudinal trajectories by forming pseudo-cohorts within the population. A key requirement for these sets of longitudinal trajectories is that they reproduce the cross-sectional distribution of risk factor categories for any year with available data. The method adopted here is based on the assumption that a person's risk factor value changes throughout their lives in such a way that they always have the same associated percentile rank. As they age, individuals move from one age group to another and their risk factor value changes so that they have the same percentile rank but of a different risk factor distribution. In a nutshell, we assume (in accordance with research on the long-term success rate in dieting) that relatively high-risk people will remain relatively high risk and that relatively low-risk people will remain relatively low risk. Crucially, it meets the important condition that the cross-sectional risk factor distributions obtained by simulation match the risk factor distributions of the observed data.

The above procedure can be explained using the example of the albuminuria distributions. The albuminuria distributions are known for the population stratified by sex and age for all years of the simulation (by extrapolation of fitted model, see equation (1.1)). A person who is in age group A and

who grows 10 years older will at some time move into the next age group A' and will have an albuminuria that was described first by the distribution $f(\beta|A,S,t)$ and then at the later time t' by the distribution $f(\beta|A',S,t')$. If the albuminuria of that individual is on the p-th percentile of the albuminuria distribution, their albuminuria will change from β to β' so that:

$$\beta = F^{-1}\left(\frac{p}{100}|A,S,t\right) \tag{1.33}$$

$$\beta' = F^{-1}\left(\frac{p}{100}|A',S,t'\right) \Rightarrow \beta' = F^{-1}\left(F(\beta \mid A,S,t)|A',S,t'\right) \tag{1.34}$$

Where F^{-1} is the inverse of the cumulative distribution function of β, which we model with a continuous uniform, normal, or lognormal distribution (depending on the risk factor) within the risk factor categories. Equation (1.34) guarantees that the transformation taking the random variable β to β' ensures the correct cross-sectional distribution at time t'.

The microsimulation first generates individuals from the risk factor distributions of the set D and, once generated, grows the individual's risk factors in a way that is also determined by the set D. It is possible to implement equation (1.34) as a suitably fast algorithm.

Disease Module

Disease modelling relies heavily on the sets of incidence, mortality, survival, relative risk, and prevalence statistics. The microsimulation uses risk-dependent incidence statistics, and these are inferred from the relative risk statistics and the distribution of the risk factor within the population. In the simulation, individuals are assigned a risk factor trajectory giving their personal risk factor history for each year of their lives. Their probability of getting a particular risk factor-related disease in a particular year will depend on their risk factor state in that year. The necessary equations are given below. The microsimulation model has the ability to model discrete multiple stages of a disease.

Once a person has a fatal disease (or diseases) their probability of survival will be controlled by a combination of the disease survival statistics and the probabilities of dying from other causes. Disease survival statistics are modelled as age- and gender-dependent exponential distributions.

Relative Risks

The reported incidence risks for any disease do not make reference to any underlying risk factor. The microsimulation requires this dependence to be manifested.

The risk factor dependence of disease incidence must be inferred from the distribution of the risk factor in the population (here denoted as π); suppose that α is a risk factor state of some risk factor A and denote by $p_A(d|\alpha,a,s)$ the incidence probability for the disease d given the risk state, α, the person's age, a, and gender, s. The relative risk ρ_A is defined by equation (1.35).

$$\begin{aligned} p_{\mathrm{A}}\left(d\left|\alpha,a,s\right.\right) &= \rho_{\mathrm{A}|d}\left(\alpha\left|a,s\right.\right) p_{\mathrm{A}}\left(d\left|\alpha_0,a,s\right.\right) \\ \rho_{\mathrm{A}|d}\left(\alpha_0\left|a,s\right.\right) &\equiv 1 \end{aligned} \tag{1.35}$$

where α_0 is the zero-risk state.

The incidence probabilities, as reported, can be expressed in terms of the equation:

$$
\begin{aligned}
p(d|a,s) &= \sum_{\alpha} p_{\mathrm{A}}(d|\alpha,a,s)\pi_A(\alpha|a,s) \\
&= p_A(d|\alpha_0,a,s)\sum_{\alpha}\rho_{\mathrm{A}|d}(\alpha|a,s)\pi_A(\alpha|a,s)
\end{aligned}
\tag{1.36}
$$

Combining these equations allows the conditional incidence probabilities to be written in terms of known quantities:

$$
p(d|\alpha,a,s) = \rho_{A|d}(\alpha|a,s)\frac{p(d|a,s)}{\sum_{\beta}\rho_{A|d}(\beta\,|\,a,s)\pi_A(\alpha|a,s)} \tag{1.37}
$$

Previous to any series of Monte Carlo trials, the microsimulation programme pre-processes the set of diseases and stores the calibrated incidence statistics $p_A(d|\square_0, a, s)$. For each scenario, the incidence statistics are calibrated against the baseline trends.

Approximating Missing Data Points

Published disease statistics are frequently incomplete and occasionally inconsistent. The microsimulation programme makes use of several supporting methods to check and, as necessary, to estimate missing disease statistics.

Model Output Module

Cross-sectional outputs (epidemiological and economic) per 100,000 of the population are computed for each year of the simulation.

A range of different epidemiological outputs are produced by the model including:

- incidence rates
- cumulative incidence rates
- prevalence rates
- premature mortality
- QALY
- costs.

Some of these outputs are standard and do not require further explanation. The QALY can be discounted if required and this can be defined by the user at the start of a modelling project. The discounting rate each year ($Discount(year)$) was calculated as shown in equation (1.38).

$$
Discount(year) = \frac{1}{(1+R)^{year-year_{start}}} \tag{1.38}
$$

Where $year_{start}$ refers to the start year of the modelling, which is 2020 in this study, and R is the annual discount rate.

Confidence Intervals

The confidence intervals (CIs) that accompany the sets of output data represent the accuracy of the microsimulation (stochastic or aleatoric uncertainty) as opposed to the confidence of the input data itself (parameter uncertainty). Errors around the input data were not available.

To estimate Monte Carlo error, we first calculate a variance for all estimates using the binomial variance formula $\sigma^2 = np(1-p)$, where n is the total number of trials (individuals modelled in a given year), and p is the proportion of individuals out of the total within the given group (e.g., risk factor group, disease group, demographic group, or a combination thereof). CIs were calculated by multiplying variance by 1.96 (2 decimal places). For downstream analyses, CIs were aggregated using the formula $\sqrt{\sum_{i=1}^{n} CI_i}$ (i.e., the square root of the sum of squared CIs).

Appendix S4 *Inside CKD* Scientific Steering Committee

Country/region	**Key external expert**	**Affiliation**
Australia	Prof. Steven Chadban	Royal Prince Alfred Hospital, Sydney, Australia
Belgium	Prof. Michel Jadoul	Cliniques Universitaires Saint-Luc, Université catholique de Louvain, Brussels, Belgium
Brazil	Prof. Marcelo Costa Batista	Hospital Israelita Albert Einstein, São Paulo, Brazil
Canada	Dr Navdeep Tangri	University of Manitoba, Winnipeg, Canada
China	Prof. Guisen Li	Sichuan Academy of Medical Sciences, Sichuan Provincial People's Hospital, Chengdu, China
Colombia	Prof. José Javier Arango Álvarez	Universidad del Quindío, Quindío, Colombia
France	Prof. Jean-Michel Halimi	Service de Néphrologie-HTA, Dialyses, Transplantation Rénale, CHU Tours, Tours, France
Germany	Prof. Kai-Uwe Eckardt	Department of Nephrology and Medical Intensive Care, Charité Universitätsmedizin Berlin, Berlin, Germany
India	Prof. Vivekanand Jha	George Institute for Global Health India, New Delhi, India
Israel	Prof. Avraham Karasik	Maccabi Institute for Research and Innovation, Tel-Aviv, Israel
Israel	Prof. Gil Chernin	Kaplan Medical Center, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel
Italy	Prof. Francesco Saverio Mennini	CEIS-EEHTA, Faculty of Economics, University of Rome Tor Vergata, Rome, Italy
Italy	Prof. Luca De Nicola	Department of Advanced Medical and Surgical Sciences, University of Campania Luigi Vanvitelli, Naples, Italy
Japan	Prof. Eiichiro Kanda	Kawasaki Medical University, Okayama, Japan
Mexico	Prof. José Ricardo Correa-Rotter	Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Mexico City, MEXICO
Philippines, Singapore, and Thailand	Assoc. Prof. Jason Choo Chon Jun	Singapore General Hospital, Singapore

Saudi Arabia	Prof. Saeed M. G. Al-Ghamdi	King Abdulaziz University Hospital and King Faisal Specialist Hospital and Research Centre, Jeddah, Saudi Arabia
South Korea	Prof. Kook-Hwan Oh	Seoul National University College of Medicine, Seoul, South Korea
Spain	Prof. Juan Francisco Navarro-González	Research Unit and Nephrology Service, University Hospital Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain
Sweden	Prof. Johan Ärnlöv	Department of Neurobiology, Care Sciences and Society, Division of Family Medicine and Primary Care, Karolinska Institute, Stockholm, Sweden
Taiwan	Prof. Mai-Szu Wu	Division of Nephrology, Taipei Medical University, Taipei, Taiwan
Turkey	Prof. Mustafa ARICI	Hacettepe University Faculty of Medicine, Department of Nephrology, Ankara, Turkey
United Arab Emirates	Prof. Stephen Holt	SEHA Kidney Care, Abu Dhabi, UAE
United Kingdom	Dr Albert Power	North Bristol NHS Trust, Bristol, UK
United States of America	Prof. Glenn Chertow	Stanford University School of Medicine, California, USA
United States of America	Prof. Jay Wish	Indiana University School of Medicine, Indianapolis, USA

CKD chronic kidney disease, *NHS* National Health Service