

SYSTEMS AND METHODS FOR PREDICTING KIDNEY CONDITIONS

Technical Field

5 The present disclosure generally relates to systems and methods for predicting kidney conditions for patients. More specifically, the present disclosure describes various embodiments of computer systems and computerized methods for training machine learning models and predicting kidney conditions using machine learning models.

10 Background

Many people worldwide suffer from kidney-related problems such as chronic kidney disease (CKD). A significant proportion of patients with CKD end up with crash or unplanned dialysis. A patient may be labelled as having a crash dialysis if there is little
15 to no care by a nephrologist beforehand, and a patient may be labelled as having an unplanned dialysis if it starts during hospitalization. Some studies suggest that the prevalence of crash and unplanned dialysis ranges between 40% and 60%.

It is difficult to detect rapid decline of kidney function in the early stages of CKD as
20 symptoms may not be readily apparent. Healthcare providers would need to make on-the-spot decisions on which patients require treatment and which ones do not. In these early stages, healthcare providers cannot adequately diagnose the rapid decline of kidney function until it is too late. This delay in detection slows patient care with a nephrologist significantly, ultimately leading to kidney failure. This not only disrupts the
25 patient's life but also places a significant operational and economic burden on the healthcare economic system.

Therefore, in order to address or alleviate at least one of the aforementioned problems and/or disadvantages, there is a need to provide improved systems and methods for
30 predicting kidney conditions.

Summary

According to a first aspect of the present disclosure, there is a computer system and computerized method for predicting a kidney condition for a patient using a machine learning model. The method comprises:

obtaining a current patient input dataset associated with the patient, the

current patient input dataset comprising:

age;

gender; and

a baseline estimated glomerular filtration rate (eGFR);

generating a patient input vector representation using the current patient input dataset and an input linear layer of the machine learning model;

obtaining a historical eGFR input dataset associated with the patient, the historical eGFR input dataset comprising a plurality of historical eGFRs of the patient, the historical eGFRs preceding the baseline eGFR;

generating an eGFR vector representation using the historical eGFR input dataset and a first input machine learning module of the machine learning model;

concatenating the patient input and eGFR vector representations into an output vector representation; and

generating a prediction output using the output vector representation and an output machine learning module of the machine learning model, the prediction output comprising a risk assessment of the patient contracting the kidney condition.

According to a second aspect of the present disclosure, there is a computer system and computerized method for training a machine learning model for predicting kidney conditions for patients. The method comprises:

obtaining current patient training datasets associated with a population of patients, each current patient training dataset associated with a respective patient in the population and comprising:

age;

gender; and

a baseline estimated glomerular filtration rate (eGFR);

generating, for each current patient training dataset, a patient input vector representation using the respective current patient training dataset and an input linear layer;

obtaining historical eGFR training datasets associated with the population, each historical eGFR training dataset associated with a respective patient in the population and comprising a plurality of historical eGFRs of the respective patient, the historical eGFRs preceding the baseline eGFR;

generating, for each historical eGFR training dataset, an eGFR vector representation using the respective historical eGFR training dataset and a first input machine learning module of the machine learning model;

concatenating, for each patient, the respective patient input and eGFR vector representations into a respective output vector representation; and

inputting the respective output vector representation and medical data of each patient into an output machine learning module of the machine learning model to thereby train the machine learning model, the medical data comprising details of the respective patient contracting the kidney condition,

wherein the trained machine learning model is configured for generating a prediction output for a new patient, the prediction output comprising a risk assessment of the new patient contracting the kidney condition.

Systems and methods for predicting kidney conditions according to the present disclosure are thus disclosed herein. Various features, aspects, and advantages of the present disclosure will become more apparent from the following detailed description of the embodiments of the present disclosure, by way of non-limiting examples only, along with the accompanying drawings.

Brief Description of the Drawings

Figures 1A and 1B are illustrations of a computer system comprising a machine learning model for predicting kidney conditions.

Figures 2A to 2E are illustrations of the machine learning model.

Figure 3 is a flowchart illustration of a method for predicting kidney conditions using the machine learning model.

Figure 4 is a flowchart illustration of a method for training the machine learning model.

5

Figure 5 is a table showing key characteristics of patients used in training the machine learning models.

Figures 6A and 6B are tables of performance results of the machine learning models in predicting kidney conditions.

10

Figures 7A and 7B are tables of additional performance results of the machine learning models in predicting kidney conditions.

Figures 8A to 8C are illustrations of a prediction report.

15

Detailed Description

For purposes of brevity and clarity, descriptions of embodiments of the present disclosure are directed to systems and methods for predicting kidney conditions, in accordance with the drawings. While aspects of the present disclosure will be described in conjunction with the embodiments provided herein, it will be understood that they are not intended to limit the present disclosure to these embodiments. On the contrary, the present disclosure is intended to cover alternatives, modifications and equivalents to the embodiments described herein, which are included within the scope of the present disclosure as defined by the appended claims. Furthermore, in the following detailed description, specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be recognised by an individual having ordinary skill in the art, i.e. a skilled person, that the present disclosure may be practiced without specific details, and/or with multiple details arising from combinations of aspects of particular embodiments. In a number of instances, well-known systems, methods, procedures, and components have not been described

20

25

30

in detail so as to not unnecessarily obscure aspects of the embodiments of the present disclosure.

In embodiments of the present disclosure, depiction of a given element or consideration or use of a particular element number in a particular figure or a reference thereto in corresponding descriptive material can encompass the same, an equivalent, or an analogous element or element number identified in another figure or descriptive material associated therewith.

References to “an embodiment / example”, “another embodiment / example”, “some embodiments / examples”, “some other embodiments / examples”, and so on, indicate that the embodiment(s) / example(s) so described may include a particular feature, structure, characteristic, property, element, or limitation, but that not every embodiment / example necessarily includes that particular feature, structure, characteristic, property, element or limitation. Furthermore, repeated use of the phrase “in an embodiment / example” or “in another embodiment / example” does not necessarily refer to the same embodiment / example.

The terms “comprising”, “including”, “having”, and the like do not exclude the presence of other features / elements / steps than those listed in an embodiment. Recitation of certain features / elements / steps in mutually different embodiments does not indicate that a combination of these features / elements / steps cannot be used in an embodiment.

As used herein, the terms “a” and “an” are defined as one or more than one. The use of “/” in a figure or associated text is understood to mean “and/or” unless otherwise indicated. The term “set” is defined as a non-empty finite organisation of elements that mathematically exhibits a cardinality of at least one (e.g. a set as defined herein can correspond to a unit, singlet, or single-element set, or a multiple-element set), in accordance with known mathematical definitions.

Representative or exemplary embodiments of the present disclosure describe a computer system 100 for predicting kidney conditions for patients. The kidney

conditions may be associated with progression of CKD, such as to determine the rate of decline in kidney function or the last stage of CKD. In particular, the computer system 100 predicts rapid progression of kidney disease and initiation of renal replacement therapy, especially for patients in stages 3 to 5 of CKD who have not yet started maintenance dialysis or received a kidney transplant.

The computer system 100 includes at least one processor 102, at least one storage device 104, at least one memory 106, and various other modules or components. The modules and components thereof are configured for performing various operations or steps of one or more computer-implemented or computerized methods, as described further below, and are configured as part of the processor 102. Such operations or steps are performed in response to non-transitory instructions operative or executed by the processor 102. The storage device 104 and/or memory 106 is used to store instructions and perhaps data which are read during program execution. The storage device 104 and/or memory 106 may be referred to as computer-readable storage media and/or non-transitory computer-readable media. Non-transitory computer-readable media include all computer-readable media, with the sole exception being a transitory propagating signal per se.

The computer system 100 may be based on a centralized model, decentralized model, or hybrid model. Some non-limiting examples of the computer system 100 include computers, laptops, mini-computers, mainframe computers, any non-transient and tangible machines that can execute a machine-readable code, cloud-based servers, distributed server networks, and a network of computers. In many embodiments, the computer system 100 includes a server 108 which is a physical or cloud data processing system on which a server program runs. The server 108 may be implemented in hardware or software, or a combination thereof.

The computer system 100 is configured for predicting the kidney conditions for patients using one or more machine learning models 200. The computer system 100 is further configured for training one or more machine learning models 200 to perform said predicting of kidney conditions for patients. A machine learning model 200 includes hardware and/or software components or modules that are operable to facilitate

processing using machine learning algorithms or other artificial intelligence-based structures/architectures. By way of non-limiting examples, the machine learning modules include and/or are configured to utilize random forest algorithms, logistic regression functions, and/or neural networks.

5

The computer system 100 includes a patient database 110 storing data associated with patients. The data may be retrieved from electronic health and medical records of the patients, such as from hospitals, clinics, and/or healthcare facilities which the patients may have used. The computer system 100 is communicative with external
10 systems 120 of such hospitals, clinics, and/or healthcare facilities via a communication network 130 to retrieve the patient data and store the patient data on the patient database 110.

The communication network 130 is a medium or environment through which content,
15 notifications, and/or messages are communicated among various components. Suitable security protocols, such as encryption protocols, may be implemented in the communication network for secure communications among the components. Some non-limiting examples of the communication network 130 include a virtual private network (VPN), wireless fidelity (Wi-Fi) network, light fidelity (Li-Fi) network, local area
20 network (LAN), wide area network (WAN), metropolitan area network (MAN), satellite network, Internet, fibre optic network, coaxial cable network, infrared network, radio frequency (RF) network, and any combination thereof. Various components in the communication network 130 may connect to the communication network 130 in accordance with various wired and wireless communication protocols, such as
25 Transmission Control Protocol / Internet Protocol (TCP/IP), User Datagram Protocol (UDP), 2nd to 5th Generation (2G to 5G) communication protocols, Long Term Evolution (LTE) communication protocols, and any combination thereof. Each component to the communication network 130 includes a data communication or transceiver module to communicate and transmit / receive data over the
30 communication network. Some non-limiting examples of a transceiver module include an antenna module, a radio frequency transceiver module, a wireless transceiver module, a Bluetooth transceiver module, an Ethernet port, a Universal Serial Bus (USB)

port, or any other module / component / device configured for transmitting and receiving data.

As shown in Figure 1B, the patient data may include training data 112 from a population of patients that is used for training the machine learning model 200. The patient data may include new patient data 114 from a new patient that is input into the machine learning model 200 for predicting a kidney condition for the new patient. Optionally, the patient data may be pre-processed before inputting to the machine learning model 200. Figures 2A to 2E show schematics of an exemplary machine learning model 200 according to various embodiments of the present disclosure.

The patient data may include, but is not limited to, demographics data, lifestyle data, comorbidity data, physical exam data, biochemical data, and historical eGFR (estimated glomerular filtration rate) data, including various combinations thereof. The demographics data may include data variables such as, but are not limited to, age as at a baseline date and patient gender, including various combinations thereof. The lifestyle data may include data variables such as, but are not limited to, indications of smoking and alcohol consumption by the patient, including various combinations thereof. The comorbidity data may include data variables such as, but are not limited to, indications of the patient having diabetes and hypertension, including various combinations thereof. The physical exam data may include data variables such as, but are not limited to, blood pressure data including systolic blood pressure and diastolic blood pressure, including various combinations thereof. The biochemical data may include data variables such as, but are not limited to, baseline eGFR measured at the baseline date, first eGFR measured after the baseline date, blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP),

including various combinations thereof. The historical eGFR data includes a plurality of historical eGFR data variables preceding the baseline eGFR of the patient.

These data variables may be measured from the patient based on one or more samples obtained from the patient. Further, these samples may be obtained at one or more times from the patient, such as during one or more medical checks or health screenings. The samples may include biofluidic samples such as blood, urine, and saliva samples. Clinicians may order for suitable tests for the patient to measure certain data variables, such as urine chemistry test, blood cell counts, and liver screening.

Prediction Using Machine Learning Models

Further with reference to Figure 3, various embodiments of the present disclosure describe a computerized method 300 for predicting a kidney condition for a patient using the machine learning model 200, such as shown in Figures 2A to 2E. The machine learning model 200 is pre-trained using the training data 112 associated with a population of patients. One or more processors 102 of the computer system 100 are configured to perform various steps of the computerized method 300 in response to non-transitory instructions operative or executed by the processors 102.

New patient data 114 associated with the patient is retrieved from the patient database 110 and the new patient data 114 is used to predict the kidney condition using the machine learning model 200, such as shown in Figure 2A. The new patient data 114 may be fed from the patient database 110, or directly from the external systems 120 of hospitals, clinics, and/or healthcare facilities.

The method 300 includes a step 310 of obtaining a current patient input dataset associated with the patient. The machine learning model 200 includes a patient input module 210 configured to receive the current patient input dataset from the patient database 110. The current patient input dataset is derived from the new patient data 114 and contains the patient's most recent or closest data to the baseline data, and includes the patient's age, gender, and baseline eGFR. The current patient input

dataset may optionally include any one or more of: blood pressure (i.e. systolic and/or diastolic), indication of whether the patient has diabetes and/or hypertension, blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).

Further as shown in Figure 2B, the machine learning model 200 includes an input linear layer 220, also referred to as a fully connected layer, configured to apply a linear transformation through a set of weights matrices to the data variables in the current patient input dataset. The method 300 includes a step 320 of generating a patient input vector representation 222 using the current patient input dataset and the input linear layer 220.

The method 300 includes a step 330 of obtaining a historical eGFR input dataset associated with the patient. The machine learning model 200 includes a historical eGFR input module 230 configured to receive the historical eGFR input dataset from the patient database 110. The historical eGFR input dataset is derived from the new patient data 114 and includes a plurality of historical eGFRs of the patient, the historical eGFRs preceding the baseline eGFR of the patient.

Further as shown in Figure 2C, the machine learning model 200 includes a first input machine learning module 240 configured to process the historical eGFR input dataset. The method 300 includes a step 340 of generating an eGFR vector representation 242 using the historical eGFR input dataset and the first input machine learning module 240. For example, the historical eGFRs were measured within 180 days before the baseline date and fed to the first input machine learning module 240. The first input machine learning module 240 may include a neural network, such as a recurrent

neural network which may include a gated mechanism. For example, the first input machine learning module 240 includes a gated recurrent unit (GRU) or a long short-term memory (LSTM) network. In many embodiments as shown in Figure 2C, the first input machine learning module 240 is a GRU neural network that transforms the historical eGFRs into the eGFR vector representation 242.

A GRU neural network is a type of recurrent neural network that makes use of a loop that allows information to be passed from one step of the network to the next step. Further, a GRU neural network uses two gates – an update gate and a reset gate – that decide what information is allowed through to the output and can be trained to retain information from farther back. This allows relevant information, i.e. the historical eGFRs, to be passed through the GRU neural network to make better predictions.

Further as shown in Figure 2D, the machine learning model 200 includes a vector concatenation module 250 configured to concatenate the patient input vector representation 222 and the eGFR vector representation 242 into an output vector representation 260. The method 300 includes a step 350 of concatenating the patient input vector representation 222 and the eGFR vector representation 242 into the output vector representation 260. The output vector representation 260 is representative of the status of the patient at the baseline date of the patient, i.e. when data variables of the patient are entered into the patient database 110.

Further as shown in Figure 2E, the machine learning model 200 includes an output machine learning module 270 configured to generate a prediction output 280 from the output vector representation 260. The method 300 includes a step 360 of generating the prediction output 280 using the output vector representation 260 and the output machine learning module 270. The output machine learning module 270 may include one or more of a random forest algorithm, an XGBoost algorithm, and a neural network. For example, the output machine learning module 270 includes a feedforward neural network or a multilayer perceptron. In many embodiments as shown in Figure 2E, the output machine learning module 270 includes a multilayer perceptron 272 and a logistic regression function 274 to generate the prediction output 280. The multilayer perceptron 272 is a type of feed forward neural network that has three layers – input

layer, output layer, and hidden layer with a non-linear activation function such as a rectified linear unit (ReLU) activation function.

The prediction output 280 includes a risk assessment of the patient contracting the kidney condition. More specifically, the prediction output 280 includes the risk assessment, such as a probability, of the patient contracting the kidney condition within a predefined future period, such as 3 months, 6 months, 1 year, 2 years, or 5 years.

In some embodiments, the method 300 may include obtaining a second historical input dataset associated with the patient. The machine learning model 200 may include a second historical input module configured to receive the second historical input dataset from the patient database 110. The second historical input dataset is derived from the new patient data 114 and includes a plurality of historical measurements of a biochemical variable of the patient, the historical measurements preceding a baseline measurement of the biochemical variable. The current patient input dataset further includes the baseline measurement of the biochemical variable. The biochemical variable may include any one of: blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).

The machine learning model 200 may include a second input machine learning module configured to process the second historical input dataset. The method 300 may include generating a second vector representation using the second historical input dataset and the second input machine learning module. The vector concatenation module 250 may be configured to concatenate the patient input vector representation 222, the

eGFR vector representation 242, and the second vector representation into the output vector representation 260.

For example, the historical measurements of the biochemical variable were measured within 180 days before the baseline date and fed to the second input machine learning module. The second input machine learning module may include a neural network, such as a recurrent neural network which may include a gated mechanism. For example, the second input machine learning module includes a GRU or a LSTM network. It will be appreciated that various historical data variables of the patient may be used in one or more additional machine learning modules in the machine learning model 200, and the output vector representation 260 is generated accordingly.

Training Machine Learning Models

With reference to Figure 4, various embodiments of the present disclosure describe a computerized method 400 for training the machine learning model 200, such as shown in Figures 2A to 2E, for predicting kidney conditions for patients. One or more processors 102 of the computer system 100 are configured to perform various steps of the computerized method 400 in response to non-transitory instructions operative or executed by the processors 102.

Training data 112 associated with a population of patients is retrieved from the patient database 110 and the training data 112 is used to train the machine learning model 200, such as shown in Figure 2A. The training data 112 may be fed from the patient database 110, or directly from the external systems 120 of hospitals, clinics, and/or healthcare facilities.

The method 400 includes a step 410 of obtaining current patient training datasets associated with the population of patients, each current patient training dataset associated with a respective patient in the population. The patient input module 210 of the machine learning model 200 is configured to receive the current patient training datasets from the patient database 110. Each current patient training dataset is derived from the training data 112 and contains the respective patient's most recent

or closest data before the baseline data, and includes the respective patient's age, gender, and baseline eGFR. Each current patient training dataset may optionally include any one or more of: blood pressure (i.e. systolic and/or diastolic), indication of whether the patient has diabetes and/or hypertension, blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).

These data variables may be measured from each patient based on one or more samples obtained from each patient. Further, these samples may be obtained at one or more times from each patient, such as during one or more medical checks or health screenings. The samples may include biofluidic samples such as blood, urine, and saliva samples. Clinicians may order for suitable tests for the patients to measure certain data variables, such as urine chemistry test, blood cell counts, and liver screening.

The input linear layer 220 of the machine learning model 200 is configured to apply a linear transformation through a set of weights matrices to the data variables in each current patient training dataset. The method 400 includes a step 420 of generating, for each current patient training dataset, the patient input vector representation 222 using the respective current patient training dataset and the input linear layer 220.

The method 400 includes a step 430 of obtaining historical eGFR training datasets associated with the population of patients, each historical eGFR training dataset associated with a respective patient in the population. The historical eGFR input module 230 of the machine learning model 200 is configured to receive the historical eGFR training datasets from the patient database 110. Each historical eGFR training dataset is derived from the training data 112 and includes a plurality of historical

eGFRs of the respective patient, the historical eGFRs preceding the baseline eGFR of the respective patient.

5 The first input machine learning module 240 of the machine learning model 200 is configured to process the historical eGFR training datasets. The method 400 includes a step 440 of generating, for each historical eGFR training dataset, the eGFR vector representation 242 using the respective historical eGFR training dataset and the first input machine learning module 240. For example, the historical eGFRs were measured within 180 days before the baseline date and fed to the first input machine
10 learning module 240. The first input machine learning module 240 preferably includes a GRU neural network.

The vector concatenation module 250 of the machine learning model 200 is configured to concatenate, for each patient, the respective patient input vector representation 222
15 and the respective eGFR vector representation 242 into a respective output vector representation 260. The method 400 includes a step 450 of concatenating, for each patient, the respective patient input vector representation 222 and the respective eGFR vector representation 242 into the respective output vector representation 260. The respective output vector representation 260 is representative of the patient status
20 at the baseline date of the respective patient.

The method 400 includes a step 460 of inputting the respective output vector representation 260 and medical data of each patient into the output machine learning module 270 to thereby train the machine learning model 200. The medical data
25 includes details of the respective patient contracting the kidney condition, such as contracting CKD stage 3 after the baseline date of the patient. Optionally, the medical data includes the first eGFR succeeding the baseline eGFR, i.e. the first eGFR that is measured after the baseline date of the patient.

30 The output machine learning module 270 preferably includes a multilayer perceptron and a logistic regression function to generate the prediction output 270. Optionally, the machine learning model 200 includes a binary cross-entropy loss function for the output machine learning module 270 to facilitate training.

The machine learning model 200 trained using the training data 112 from the population of patients is thus configured for generating the prediction output 280 for a new patient using the method 300 and new patient data 114 of the new patient. The prediction output 280 includes the risk assessment of the new patient contracting the kidney condition.

In some embodiments, the method 400 includes obtaining second historical training datasets associated with the population of patients, each second historical training dataset associated with a respective patient in the population. The second historical input module of the machine learning model 200 is configured to receive the second historical training datasets from the patient database 110. Each second historical training dataset is derived from the training data 112 and includes a plurality of historical measurements of a biochemical variable of the respective patient, the historical measurements preceding a baseline measurement of the biochemical variable. The respective current patient training dataset further includes the baseline measurement of the biochemical variable. The biochemical variable may include any one of: blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).

The second input machine learning module of the machine learning model 200 may be configured to process the second historical training datasets. The method 400 may include generating, for each second historical training dataset, a second vector representation using the respective second historical input dataset and the second input machine learning module. The vector concatenation module 250 may be configured to concatenate the respective patient input vector representation 222, the

respective eGFR vector representation 242, and the respective second vector representation into the respective output vector representation 260. It will be appreciated that various historical data variables of the patients may be used in one or more additional machine learning modules in the machine learning model 200, and the output vector representation 260 is generated accordingly.

As described above, the prediction output 280 includes the risk assessment of the patient contracting the kidney condition within a predefined future period, such as 3 months, 6 months, 1 year, 2 years, or 5 years. Multiple machine learning models 200 may be trained for the respective predefined future periods. More specifically, each respective machine learning model 200 is trained using training data 112 associated with timeframes corresponding to the respective predefined future period. For example, training data 112 for patients over a timeframe of 3 months is used to train the machine learning model 200 corresponding to the 3-month future period. Moreover, training data 112 from shorter timeframes contribute to training of machine learning models 200 corresponding to longer future periods. Alternatively, a single machine learning model 200 may be trained to generate prediction outputs 280 for multiple future periods. For example, a multi-task learning method may be used to train the machine learning model 200 using training data 112 from multiple timeframes.

In some embodiments, a machine learning model 200 is trained to predict initiation of renal replacement (IR) therapy, which is a therapy that replaces the normal function of the kidneys. More specifically, the IR machine learning model 200 predicts a kidney condition that requires the patient, especially if the patient is in stages 3 to 5 of CKD, to undergo maintenance dialysis or kidney transplant. Weights and/or biases may optionally be determined from the IR model 200 to expedite training of another machine learning model 200, such as one for predicting rapid progression of kidney disease, using the same training data 112.

In some embodiments, a machine learning model 200 is trained to predict rapid progression (RP) of kidney disease. More specifically, the RP machine learning model 200 predicts a kidney condition for patients with a baseline eGFR of more than 15 mL/min/1.73m², the kidney condition associated with one or more of: an eGFR

gradient of at least 5 mL/min/1.73m² per year, a sustained decrease in eGFR of at least 40% over at least 3 months, and sustained eGFR of less than or equal to 15 mL/min/1.73m².

- 5 In some embodiments, the RP machine learning model 200 may include the ReLU activation function for the output machine learning module 270. For example in the method 300 of predicting the kidney condition using the RP model 200, the method 300 may include applying the ReLU activation function to the output vector representation 260 before inputting the output vector representation 260 to the output machine learning module 270. Similarly, in the method 400 of training the RP model 200, the method 400 may include applying the ReLU activation function to each output vector representation 260 before inputting the respective output vector representation 260 to the output machine learning module 270.
- 10
- 15 In some embodiments, the RP model 200 may be trained using weights and/or biases obtained from another pre-trained machine learning model 200, such as the IR model 200, and the same training data 112, i.e. the same current patient training datasets and historical eGFR training datasets. Accordingly, the RP model 200 can be developed by fine-tuning the pre-trained IR model 200 and by utilizing the learned weights and/or biases from the IR model 200. The RP model 200 thus inherits knowledge from the pre-trained IR model 200, thereby facilitating training of the RP model 200. For example, the RP model 200 can be trained more efficiently using a smaller learning rate.
- 20

25 Evaluation of Machine Learning Models

The RP and IR machine learning models 200 were trained and tested using actual data of a population of patients from a first major hospital. The output machine learning module 270 of the RP and IR models 200 included a multilayer perceptron and a logistic regression function, and the RP models 200 further included the ReLU activation function. The first input machine learning modules 240 for the historical eGFRs included a GRU neural network.

30

The data from the population of patients included information electronic medical records, questionnaires, laboratory information systems, etc. There were about 10,000 patients in this population aged between 20 and 90 years old with CKD stage 3 and above, i.e. an eGFR of less than or equal to 60 mL/min/1.73m². Key characteristics of this population dataset are shown in Figure 5.

The population dataset was divided into a development dataset (about 80% of the patients) and a validation dataset (about 20% of the patients). The development dataset was divided into a training dataset (about 80% of the patients) and a testing dataset (about 20% of the patients). The training dataset was used for training the machine learning models 200 to capture the information contained in the training dataset. The testing dataset was used for evaluating the performance of the machine learning models 200 and to establish correlations between inputs and outputs. The validation dataset was used to assess the generalizability of the machine learning models 200.

The datasets were further divided in the following ways to train the machine learning models 200 for specific prediction outputs 280:

- The datasets were categorized by age with each age group spanning 10 years.
- The datasets were categorized by gender.
- The datasets were categorized into patients with and without outcomes that happened within the respective timeframes after the baseline date.

Some data variables in the population dataset were measured within 180 days before the baseline date, some were measured within 90 days after the baseline date, and some measurements were missing. The closest measurement to the baseline date was chosen as the baseline for each data variable. One or more data variables with missing measurements may be imputed using missing data algorithms. For example, a data variable may be replaced with a predefined measurements for that data variable, such as with the corresponding median value from the population. However, if the dataset for a patient had more than 30% missing data, the patient's dataset was discarded. Categorical variables, such as gender, lifestyle, and comorbidities, were

transformed into binary values ('1' for true and '0' for false), and any missing value were replaced with a default value of '0' indicating its absence.

Generalizability of the machine learning models 200 was assessed using the first major hospital's validation dataset and another validation dataset of about 1,000 patients from a second major hospital. The evaluation metrics include AUC (area under receiver operating characteristic curve), sensitivity, specificity, precision, PRC (area under precision-recall curve), and NRI (net reclassification improvement). AUC was the primary performance criterion, where $AUC = 0.5$ indicates no discrimination, $0.7 \leq AUC < 0.8$ indicates acceptable discrimination, $0.8 \leq AUC < 0.9$ indicates excellent discrimination, and $AUC \geq 0.9$ indicates outstanding discrimination. The cut-off thresholds were based on Youden's J statistic (Youden's index) from the training dataset, wherein the cut-off point for having an acceptable Youden's index was 50%. Figures 6A and 6B show the AUC performance results for the RP and IR models 200, respectively, for prediction outputs 280 in respective future time periods of 1 year, 2 years, and 5 years.

Risk stratification performance of the machine learning models 200 was evaluated for both major hospitals' populations. The prediction outputs 280 for the respective future periods were divided into low-risk and high-risk groups based on predefined cut-off thresholds from the first major hospital. The evaluation metrics include sensitivity, specificity, positive predictive value (PPV) for the high-risk groups, and negative predictive value (NPV) for the low-risk groups. Figures 7A and 7B show the risk stratification performance results for the RP and IR models 200, respectively, for populations from the first major hospital (labelled 'first') and the second major hospital (labelled 'second').

As described above, the method 300 generates a prediction output 280 for a patient using a trained machine learning model 200, such as the RP or IR model, wherein the prediction output 280 includes the risk assessment of the patient contracting the kidney condition, such as one corresponding to the respective RP or IR model. The method 300 may further include generating a prediction report 290 for visualization of the risk assessment, such as shown in Figure 8A. The prediction report 290 may

include clinical guidance that is recommended for comprehensive care management of the patient. The guidance is preferably pursuant to the KDIGO Clinical Practice Guideline for the Evaluation and Management of CKD. The guidance may include advice on, but not limited to, preparation for renal replacement therapy, monitoring frequency of medical condition, nephrology referral, blood pressure management and target, cardiovascular risk reduction, blood sugar control, diet and lifestyle.

The prediction report 290 may include a risk classification for the patient, such as shown in Figure 8A, wherein the patient is classified into one among a plurality of the risk classifications based on predefined risk thresholds. The prediction report 290 may include a risk score indicating the risk of rapid progression of kidney disease or renal replacement therapy, as appropriate, for the patient within a future period of such as 1 year. The patient is then classified as low-risk, medium-risk, or high-risk according to the risk score and the respective risk thresholds. For example as shown in Figure 8A, the patient has a high risk of rapid progression of kidney diseases in 1 year, and a medium risk of initiation of renal replacement therapy in 1 year.

The risk score correlates with the percentile of the predicted probability in the population dataset used to test the RP and IR models 200. For example, if the predicted probability of 0.3 was the lowest 45th percentile in population dataset, then the risk score will be 45. The risk stratification between the low-risk, medium-risk, and high-risk groups was done based on the risk thresholds or risk score cut-offs related to the clinical outcomes for the RP and IR models 200. The risk thresholds were derived from the first major hospital's population dataset and focused on maintaining a good balance between sensitivity and precision so as to capture the majority of high-risk patients without a large number of false positives.

The risk thresholds may be adjusted for different hospitals and other healthcare facilities depending on their specific requirements. For example, the risk thresholds can be defined based only on sensitivity to capture more high-risk patients. For example, the risk thresholds can be defined based only on precision to minimize the number of false positives. The adjustment can be done by taking retrospective data from the target hospital or healthcare facility and applying the above-described

validation methods to determine the appropriate risk thresholds. Alternatively or additionally, the adjustment can be done by using the existing risk thresholds from the first major hospital and then using prospective on-site data collected from the target hospital or healthcare facility for about 6 months to 1 year to adjust the risk thresholds.

5

In some embodiments, the prediction report 290 includes explanations generated using an explainable artificial intelligence. For example, the explainable artificial intelligence comprises a Shapley Additive Explanations (SHAP) method and/or a Local Interpretable Model-Agnostic Explanations (LIME) method. These explanations help to make the prediction results more valuable and actionable for the clinicians, such as to explain a specific prediction result.

10

In some embodiments, the prediction report 290 includes rankings of data variables in the current patient input dataset, wherein the data variables are ranked based on contribution to the prediction of the kidney condition. The explainable artificial intelligence may be used to determine the data variables that contribute most to the prediction of the kidney condition. For example, Figure 8B shows the top 5 contributing data variables or risk factors for an exemplary patient (male, 60 years old, baseline eGFR = 50 mL/min/1.73m²). The length of the bars represents the contribution of the data variable in predicting the patient's risk. The bars to the left of '0' represent that these data variables drive the risk score lower while the bars to the right of '0' represent that these data variables drive the risk score higher. This means that for this patient, the baseline eGFR, blood urea nitrogen, and haemoglobin are currently in his favour, but the serum phosphate and age are increasing his risk. Since the age cannot be changed, the patient's clinician may prioritise controlling the patient's serum phosphate levels to manage his risk.

15

20

25

The prediction report 290 may include a hypothetical risk assessment of the patient contracting the kidney condition when one or more of the ranked data variables are modified. Following the above example, the prediction report 290 provides information on how the risk score would change when the serum phosphate level is changed. For example as shown in Figure 8C, reducing the serum phosphate value from 6 mg/dL

30

to 4 mg/dL modifies the patient's predicted 2-year risk from 50.6% (current probability) to 30.8% (hypothetical probability).

This visualization of the lower risk in the hypothetical risk assessment can help clinicians to visualize the top contributing factors for the risk predictions and help clinicians to plan appropriate care for patients. The visualization can educate patients of their current risks and show them how improving certain factors can modify their risk predictions. Patients may understand their conditions with better clarity and may be more willing to comply with treatments if they can visualize the change in the risk prediction together with the top contributing factors for the same risk prediction.

As described above, the computer system 100 is communicative, via the communication network 130, with the external systems 120 of hospitals, clinics, and/or healthcare facilities. Patient data including training data 112 and new patient data 114 is retrieved from the external systems 120 and stored in the patient database 110. In some embodiments, the patient data is retrieved in response to user triggers. In some embodiments, the patient data is retrieved according to a predefined schedule, such as on a daily basis for the patients visiting the hospital that day. In some embodiments, the patient data is retrieved each time new patient data 114 is entered into the computer systems.

The method 300 may be performed for predicting the kidney condition for a new patient when the patient database 110 retrieves the new patient data 114 from the computer systems. The method 300 generates the prediction outputs 280 and prediction reports 290 which may be stored on a database, such as the patient database 110, of the computer system 100 for subsequent retrieval and viewing. In some embodiments, the prediction outputs 280 and prediction reports 290 are communicated to the external systems 120 of these hospitals and healthcare facilities via the communication network 130. For example, the prediction outputs 280 and prediction reports 290 are communicated in response to user triggers, according to a predefined schedule, or when the method 300 has been completed for a new patient. Optionally, the prediction reports 290 are accessible by patients and clinicians via a secure online portal.

In some embodiments, the computer system 100 is integrated with the external systems 120 of the hospitals and healthcare facilities. This would allow the computer system 100 to directly access the electronic medical records and other software programs in these external systems 120 to perform machine learning model training and predictions of kidney conditions for their patients.

Use of the machine learning models 200 described in various embodiments herein improves prediction of kidney conditions for patients. Specific models 200, such as the RP and IR ones, can be trained to target specific kidney conditions like rapid progression of kidney disease and initiation of renal replacement therapy, especially for patients in CKD stages 3 to 5. The machine learning models 200 are trained using two distinct data inputs – current baseline data of the patient and historical data of the patient, particularly the historical eGFR data. This advantageously enables the models 200 to take into the account the natural longitudinal variability of the patient's health over time to thereby generate more reliable predictions of the patient's risk of contracting the specific kidney condition.

In the foregoing detailed description, embodiments of the present disclosure in relation to systems and methods for predicting kidney conditions are described with reference to the provided figures. The description of the various embodiments herein is not intended to call out or be limited only to specific or particular representations of the present disclosure, but merely to illustrate non-limiting examples of the present disclosure. The present disclosure serves to address at least one of the mentioned problems and issues associated with the prior art. Although only some embodiments of the present disclosure are disclosed herein, it will be apparent to a person having ordinary skill in the art in view of this disclosure that a variety of changes and/or modifications can be made to the disclosed embodiments without departing from the scope of the present disclosure. Therefore, the scope of the disclosure as well as the scope of the following claims is not limited to embodiments described herein.

Claims

1. A computerized method for predicting a kidney condition for a patient using a machine learning model, the method comprising:
 - 5 obtaining a current patient input dataset associated with the patient, the current patient input dataset comprising:
 - age;
 - gender; and
 - a baseline estimated glomerular filtration rate (eGFR);
 - 10 generating a patient input vector representation using the current patient input dataset and an input linear layer of the machine learning model;
 - obtaining a historical eGFR input dataset associated with the patient, the historical eGFR input dataset comprising a plurality of historical eGFRs of the patient, the historical eGFRs preceding the baseline eGFR;
 - 15 generating an eGFR vector representation using the historical eGFR input dataset and a first input machine learning module of the machine learning model;
 - concatenating the patient input and eGFR vector representations into an output vector representation; and
 - 20 generating a prediction output using the output vector representation and an output machine learning module of the machine learning model, the prediction output comprising a risk assessment of the patient contracting the kidney condition.
- 25 2. The method according to claim 1, wherein the first input machine learning module and the output machine learning module comprise neural networks.
3. The method according to claim 1 or 2, wherein the first input machine learning module comprises a gated recurrent unit neural network or a long short-term memory (LSTM) network, and the output machine learning module comprises a multilayer
30 perceptron.

4. The method according to any one of claims 1 to 3, wherein the current patient input dataset comprises any one or more of: blood pressure, indication of whether the patient has diabetes and/or hypertension, blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).
5. The method according to any one of claims 1 to 4, further comprising:
- obtaining a second historical input dataset associated with the patient, the second historical input dataset comprising a plurality of historical measurements of a biochemical variable of the patient, the historical measurements preceding a baseline measurement of the biochemical variable, wherein the current patient input dataset comprises the baseline measurement of the biological variable, the biochemical variable comprising any one of: blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP); and
- the method optionally comprising:
- generating a second vector representation using the second historical input dataset and a second input machine learning module of the machine learning model; and

concatenating the patient input, eGFR, and second vector representations into the output vector representation.

6. The method according to any one of claims 1 to 5, wherein the kidney condition is associated with:

an eGFR gradient of at least 5 mL/min/1.73m² per year;

a decrease in eGFR of at least 40% over at least 3 months; or

sustained eGFR less than or equal to 15 mL/min/1.73m², and

wherein the machine learning model is optionally trained using weights and/or biases obtained from another pre-trained machine learning model.

7. The method according to any one of claims 1 to 5, wherein the kidney condition requires the patient to undergo maintenance dialysis or kidney transplant, and wherein the method optionally comprises determining weights and/or biases from the machine learning model to expedite training of another machine learning model.

8. The method according to any one of claims 1 to 7, further comprising generating a prediction report for visualization of the risk assessment, wherein the prediction report optionally comprises a risk classification for the patient where the patient is classified into one among a plurality of the risk classifications based on predefined risk thresholds.

9. The method according to claim 8, wherein the prediction report comprises explanations generated using an explainable artificial intelligence, and wherein the explainable artificial intelligence optionally comprises a Shapley Additive Explanations (SHAP) method and/or a Local Interpretable Model-Agnostic Explanations (LIME) method.

10. The method according to claim 8 or 9, wherein the prediction report comprises rankings of data variables in the current patient input dataset, the data variables ranked based on contribution to the prediction of the kidney condition, and wherein the prediction report optionally comprises a hypothetical risk assessment of the patient

contracting the kidney condition when one or more of the ranked data variables are modified.

11. A computer system comprising:

- 5 a machine learning model; and
- a processor configured for executing computer-readable instructions to predict a kidney condition for a patient using the machine learning model, wherein the machine learning model comprises:
 - a patient input module configured for obtaining a current patient
 - 10 input dataset associated with the patient, the current patient input dataset comprising:
 - age;
 - gender; and
 - a baseline estimated glomerular filtration rate (eGFR);
 - 15 an input linear layer configured for generating a patient input vector representation using the current patient input dataset;
 - a historical eGFR input module configured for obtaining a historical eGFR input dataset associated with the patient, the historical eGFR input dataset comprising a plurality of historical eGFRs of the
 - 20 patient, the historical eGFRs preceding the baseline eGFR;
 - a first input machine learning module configured for generating an eGFR vector representation using the historical eGFR input dataset;
 - a vector concatenation module configured for concatenating the patient input and eGFR vector representations into an output vector
 - 25 representation; and
 - an output machine learning module configured for generating a prediction output using the output vector representation, the prediction output comprising a risk assessment of the patient contracting the kidney condition.

30

12. A computerized method for training a machine learning model for predicting kidney conditions for patients, the method comprising:

obtaining current patient training datasets associated with a population of patients, each current patient training dataset associated with a respective patient in the population and comprising:

age;

5 gender; and

a baseline estimated glomerular filtration rate (eGFR);

generating, for each current patient training dataset, a patient input vector representation using the respective current patient training dataset and an input linear layer;

10 obtaining historical eGFR training datasets associated with the population, each historical eGFR training dataset associated with a respective patient in the population and comprising a plurality of historical eGFRs of the respective patient, the historical eGFRs preceding the baseline eGFR;

generating, for each historical eGFR training dataset, an eGFR vector representation using the respective historical eGFR training dataset and a first input machine learning module of the machine learning model;

15 concatenating, for each patient, the respective patient input and eGFR vector representations into a respective output vector representation; and

inputting the respective output vector representation and medical data of each patient into an output machine learning module of the machine learning model to thereby train the machine learning model, the medical data comprising details of the respective patient contracting the kidney condition,

20 wherein the trained machine learning model is configured for generating a prediction output for a new patient, the prediction output comprising a risk assessment of the new patient contracting the kidney condition.

13. The method according to claim 12, wherein the first input machine learning module and the output machine learning module comprise neural networks.

30 14. The method according to claim 12 or 13, wherein the first input machine learning module comprises a gated recurrent unit neural network or a long short-term memory (LSTM) network, and wherein the output machine learning module comprises a feedforward neural network or a multilayer perceptron.

15. The method according to any one of claims 12 to 14, wherein the medical data further comprises a first eGFR succeeding the baseline eGFR.

16. The method according to any one of claims 12 to 15, wherein each current patient training dataset comprises any one or more of: blood pressure, indication of whether the patient has diabetes and/or hypertension, blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone (iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP).

17. The method according to any one of claims 12 to 16, further comprising:

obtaining second historical training datasets associated with the population of patients, each second historical training dataset associated with a respective patient in the population and comprising a plurality of historical measurements of a biochemical variable of the respective patient, the historical measurements preceding a baseline measurement of the biological variable,

wherein the respective current patient training dataset comprises the baseline measurement of the biological variable, the biological variable comprising any one of: blood urea nitrogen, serum sodium, fasting blood glucose, total cholesterol, urine creatinine, urine protein, urine protein creatinine ratio (PCR), urine albumin-creatinine ratio (uACR), serum calcium, serum phosphate, serum creatinine, serum chloride, serum potassium, serum bicarbonate, serum albumin, serum haemoglobin, C-reactive protein, total lipids, low-density lipoprotein (LDL), triglycerides, eGFR, serum urea, haemoglobin A1C, urine phase contrast, serum uric acid, 24-hour urine volume, urea, serum iron, ferritin, transferrin saturation, haematocrit, intact parathyroid hormone

(iPTH), aspartate transferase (AST), alanine transaminase (ALT), and alkaline phosphatase (ALP); and

the method optionally comprising:

generating, for each second historical training dataset, a second vector representation using the respective second historical input dataset and a

second input machine learning module of the machine learning model; and concatenating the respective patient input, eGFR, and second vector representations into the respective output vector representation.

18. The method according to any one of claims 12 to 17, wherein the kidney condition is associated with:

an eGFR gradient of at least 5 mL/min/1.73m² per year;

a decrease in eGFR of at least 40% over at least 3 months; or

sustained eGFR less than or equal to 15 mL/min/1.73m², and

wherein the method optionally comprises training the machine learning model using weights and/or biases obtained from another machine learning model that has been trained using the same patient training datasets and historical eGFR training datasets.

19. The method according to any one of claims 12 to 17, wherein the kidney condition requires the respective patient to undergo maintenance dialysis or kidney transplant, and wherein the method optionally comprises determining weights and/or biases from the machine learning model to expedite training of another machine learning model with the same current patient training datasets and historical eGFR training datasets.

20. A computer system comprising:

a machine learning model;

a processor configured for executing computer-readable instructions to train the machine learning model for predicting kidney conditions for patients, wherein the machine learning model comprises:

a patient input module configured for obtaining current patient training datasets associated with a population of patients, each current

patient training dataset associated with a respective patient in the population and comprising:

age;

gender; and

5 a baseline estimated glomerular filtration rate (eGFR);

an input linear layer configured for generating, for each current patient training dataset, a patient input vector representation using the respective current patient training dataset and an input linear layer;

10 a historical eGFR input module configured for obtaining historical eGFR training datasets associated with the population, each historical eGFR training dataset associated with a respective patient in the population and comprising a plurality of historical eGFRs of the respective patient, the historical eGFRs preceding the baseline eGFR;

15 a first input machine learning module configured for generating, for each historical eGFR training dataset, an eGFR vector representation using the respective historical eGFR training dataset;

a vector concatenation module configured for concatenating, for each patient, the respective patient input and eGFR vector representations into a respective output vector representation; and

20 an output machine learning module configured for receiving the respective output vector representation and medical data of each patient to thereby train the machine learning model, the medical data comprising details of the respective patient contracting the kidney condition,

25 wherein the trained machine learning model is configured for generating a prediction output for a new patient, the prediction output comprising a risk assessment of the new patient contracting the kidney condition.

30