Chronic Kidney Disease
Phenotype Algorithm Pseudo Code
Columbia University
12/6/2017 (V4.2)

Contacts:
- Ning "Sunny" Shang (ns3026@cumc.columbia.edu)
- George Hripcsak (gh13@cumc.columbia.edu)
- Krzysztof Kiryluk (kk473@cumc.columbia.edu)

Update (V4.2 Compared to V4.1-08/17/2017)
- More detailed documentation of the pseudocode.

Minor Update (V4.1 Compared to V4-06/09/2017):
- Remove 50200 (Renal biopsy; percutaneous, by trocar or needle) from dialysis procedure CPT4 codes.

Updates (V4 Compared to V3B-December, 2016):
- Revised codes for dialysis, transplant and other kidney diseases
- Added sepsis, prenatal kidney injury, volume depletion and shock along with AKI for considering acute conditions that might confounding CKD affect on eGFR
- Extended the time window for finding qualified the latest urine test
- Further sub-classified CKD controls as CKD G1A1 controls (normal eGFR and negative urine test) and CKD G1 controls (normal eGFR and no urine test available)

Updates (V3B Compared to V3-August, 2016):
- Updated Columbia UPCR Ordinal Classifier and Columbia UA Ordinal Classifier
- Added query template name for coding variables in terminology dictionary

I. Background and Significance:
- **Chronic kidney disease (CKD)** is defined as an abnormality of kidney structure or function present for longer than 3 months. CKD can occur as a result of heterogeneous disorders affecting the kidney. In the United States, an estimated 13.6% of adults have CKD. Notably, adjusted mortality rates are higher for patients with CKD than those without, and rates increase with CKD stage. The purpose of this algorithm is to enable accurate CKD diagnosis and staging based on EHR data. The information on CKD stage/severity can be used to select appropriate cases for inclusion in genetic and epidemiologic studies. This information can also be used to design specific EHR-based interventions or clinical decision support tools for the diagnosis and management of CKD.
- Our algorithm generally follows the National Kidney Foundation's (NKF) Kidney Disease Outcomes Quality Initiative (KDOQI) CKD staging recommendations

(http://www2.kidney.org/professionals/KDOQI/guidelines_ckd/toc.htm), as well as the Kidney Disease: Improving Global Outcomes (KDIGO) 2012 Clinical Practice Guideline for the Evaluation and Management of CKD. Specifically, the NFK KDOQI guideline formalizes the definition and staging of CKD, while the KDIGO guideline provides an enhanced classification framework for CKD with a direct relationship to the prognosis and management of progression and complications of CKD. Overall, two measures of kidney disease severity are used to perform CKD staging: estimated glomerular filtration rate (G-staging) and albuminuria (A-staging).

- **Estimated Glomerular Filtration Rate (eGFR)** is widely accepted as the best overall index of kidney function. The estimation of GFR is performed using established equations based on serum creatinine measurement (we select the CKD-EPI as the most robust eGFR equation). The G-staging of CKD is based on specific cut-offs of eGFR. However, the eGFR equation is valid only under steady state conditions in the absence of renal replacement therapy or acute kidney injury.
- **Albuminuria** represents another marker of kidney damage (increased glomerular permeability) and refers to abnormal loss of albumin in the urine. Albuminuria is defined by daily urine albumin excretion of $\geq$30 mg/24 hours, which is approximately equivalent to spot urine albumin-to-creatinine ratio (UACR) $\geq$30 mg/g. KDIGO defines three A-stages of albuminuria: A1 defined by UACR < 30 mg/g (Normal), A2 defined by UACR 30-300 mg/g (‘‘microalbuminuria’’ or ‘‘moderately increased albuminuria’’) and A3 defined as UACR > 300 mg/g (‘‘macroalbuminuria’’ or ‘‘severely increased albuminuria’’).
- We note several important challenges that we attempt to address in the implementation of computable CKD phenotype:
  1) Challenges with G-staging: for adults, serum Cr, age, gender, and race are needed to determine eGFR; for children, concurrent height information is also needed. The race data may be incomplete in the EHR, and “White” race is typically used by default in case self-reported race is not available. The concurrent height data is also frequently incomplete in children, necessitating data extrapolation from the available measurements. Additionally, estimating equations for GFR are presently not validated in other racial/ethnic groups (e.g. Asians, Hispanics, Native Americans).
  2) Challenges with A-staging: a wide range of urine tests are being used to estimate the degree of proteinuria, and not all of the tests have a direct correspondence to the KDIGO-defined albuminuria thresholds. We use a machine learning approach to predict specific A-stage based on different types of urine tests, ranging from semi-quantitative urine dipstick, two common types of the urinalysis (UA) test, urine protein-to-creatinine ratio (UPCR), and 24-hour urine collection for protein (P24).
  3) Challenges with the Longitudinal Aspects of CKD Definition: although CKD is defined as an abnormality of kidney structure or function of longer than 3 months duration, repeat serum creatinine and/or urine albumin tests are frequently not available. Defining CKD based on a single time point serum

creatinine is problematic, since fluctuations in serum creatinine levels may occur under various physiological conditions (e.g. volume depletion, shock), as well as in the setting of acute kidney injury (AKI).

II. Development:

- CKD case and controls are developed using structured EHR data using domain knowledge and standards
- Validation is performed against expert opinion

III. Algorithm Definitions (flowchart Figure 1)

1. CKD Control (No CKD)

1.1 Definition 1: CKD G1-Control

- No diagnosis and procedure of kidney transplant or dialysis.
- Have eGFR measures.
- Most recent eGFR does not co-occur with the diagnosis of acute conditions (AKI, prerenal kidney injury, sepsis, volume depletion, shock). eGFR co-occur with the diagnosis of acute conditions is defined as that the acute condition happens before or after 1 month (31 days) of the eGFR.
- Most recent eGFR >= 90.
- No diagnosis of CKD or other kidney disease.
- No urine protein test recorded from 24 months (730 days) before the most recent eGFR until present.

1.2 Definition 2: CKD G1A1-Control

- No diagnosis and procedure of kidney transplant or dialysis.
- Have eGFR measures.
- Most recent eGFR does not co-occur with the diagnosis of acute conditions (AKI, prerenal kidney injury, sepsis, volume depletion, shock). eGFR co-occur with the diagnosis of acute conditions is defined as that the acute condition happens before or after 1 month (31 days) of the eGFR.
- Most recent eGFR >= 90.
- No diagnosis of CKD or other kidney disease.
- Have urine protein test recorded from 24 months (730 days) before the most recent eGFR until present.
- No proteinuria: the latest A staging (based on the latest urine test from 24 months before the most recent eGFR until present) is A1. The latest urine test for deriving A staging should be from 24 months (730 days) before the most recent eGFR until present.

2. CKD Case

2.1 Definition 1: ESRD with transplant

- One or more diagnosis or procedure of kidney transplant.

2.2 Definition 2: ESRD on dialysis

- One or more diagnosis or procedure of dialysis.

- The diagnosis or procedure of dialysis does not co-occur with the diagnosis of AKI. The co-occurrence is defined as AKI diagnosis happens before or after 1-month (31 days) of the diagnosis or procedure of dialysis.

2.3 Definition 3: NKF CKD Stage 1

- No diagnosis or procedure of kidney transplant or dialysis.
- Have eGFR measures.
- The most recent eGFR does not co-occur with the diagnosis of acute conditions (AKI, prerenal kidney injury, sepsis, volume depletion, shock). The co-occurrence is defined as that the acute condition happens before or after 1 month (31 days) of the eGFR.
- The most recent eGFR >=90.
- One of the following two criteria
    - One or more existing diagnosis of chronic kidney disease or other kidney disease.
    - Presence of proteinuria: the latest and its previous A-staging are A2 or A3. The latest A staging is derived from the latest urine test which should be conducted from 24 months (730 days) before the most recent eGFR until present. The previous A-staging is derived from a previous urine test, which was conducted more than 3 months prior to the most recent test.

2.4 Definition 4: NKF CKD Stage 2 to 5

- No diagnosis or procedure of kidney transplant or dialysis.
- Have eGFR measures.
- The most recent eGFR does not co-occur with the diagnosis of acute conditions (AKI, prerenal kidney injury, sepsis, volume depletion, shock). The co-occurrence is defined as that the acute condition happens before or after 1 month (31 days) of the eGFR.
- The most recent eGFR <90
- One of the two criteria
    - One or more diagnosis of CKD or other kidney disease.
    - One or more eGFR < 90 more than 3 months prior to the most recent eGFR
- NKF CKD staging based on the most recent eGFR (Figure 2, [1])
    - Stage 2: the most recent eGFR >= 60 and <90
    - Stage 3a: the most recent eGFR >= 45 and <60
    - Stage 3b: the most recent eGFR >= 30 and <45
    - Stage 4: the most recent eGFR >= 15 and <30
    - Stage 5: most recent eGFR <15

3. CKD Unknown/Indeterminate

3.1 Indeterminate due to dialysis co-occur with AKI

3.2 Indeterminate due to no eGFR

3.3 Indeterminate due to the most recent eGFR co-occur with acute conditions

3.4 Indeterminate due to no diagnosis of CKD or other type of kidney disease and no second qualified eGFR

3.5 Indeterminate due to no qualified previous A-staging for defining G1A1-control

IV. CKD Implementation

1. CKD algorithm related variables (coding see Table 1 and file CKDalgorithm_V3B_coding.txt)

1.1 Patient demographics: age, gender, race, ethnicity, and height.
1.2 Patient diagnosis: kidney transplant, dialysis, acute kidney injury (AKI), prerenal kidney injury, sepsis, volume depletion, shock, chronic kidney disease (CKD), and other types of kidney disease.
1.3 Patient procedure: kidney transplant, dialysis
1.4 Patient lab test: serum creatinine, 24-hour urine protein (P24), urine protein-to-creatinine ratio (UPCR), 24-hour urine albumin (A24), urine albumin-to-creatinine ratio (UACR), and urinalysis (UA) [protein and specific gravity]

2. Retrieve serum creatinine (Cr) and calculate estimated glomerular filtration rate (eGFR)

2.1 Get Serum Cr and demographic information (birthday, race, gender).
2.2 Get height information.
2.2.1 The height should be the closet measure to the serum Cr among all height values that are measured in the 3 months before or after the serum Cr is measured.
2.2.2 Because of incomplete height information, there may not have height measured in the 3 months range of the serum Cr measured. In this scenario, if there are height measurements recorded before and after the serum Cr measured. Then height is extrapolated from pre- and post- Height measurements:

$$\frac{HtVal_{pre} - HtVal_{post}}{HtVal_{pre} - HtVal} = \frac{HtVDateInDay_{pre} - HtDateInDay_{post}}{HtVDateInDay_{pre} - HtDateInDay}$$

$$\Longrightarrow HtVal$$

$$= HtVal_{pre} + \frac{(HtVal_{pre} - HtVal_{post})(HtVDateInDay_{pre} - HtDateInDay)}{HtVDateInDay_{pre} - HtDateInDay_{post}}$$

3. Calculate eGFR from serum Cr:

3.1 For age at the Cr measure >=18, use CKD-EPI formula [2]:

$$eGFR = 141 * \min(Scr/\kappa, 1)^{\alpha} * \max(Scr/\kappa, 1)^{-1.209} * 0.993^{Age} * (1.018\ if\ female) * (1.159\ if\ black)$$

Unit: eGFR in mL/min/1.73 m^2, serum creatinine in mg/dL.
κ is 0.7 for females and 0.9 for males; α is -0.329 for females and -0.411 for males.

3.2 For age at the Cr measure <18, use Bedside Schwartz equation[3,4]:

$$eGFR = (0.41 * Height)/Creatinine$$

Unit: eGFR in mL/min/1.73 m^2, Height in cm, and creatinine in mg/dL.

4. Deriving albuminuria categories (A-staging)

4.1 Collect microalbuminuria urine tests: 24-hour urine protein (P24), urine protein-to-creatinine ratio (UPCR), 24-hour urine albumin (A24), urine albumin-to-creatinine ratio (UACR), spot urine albumin, spot urine creatinine, spot urine protein, urine analysis protein (dipstick urine protein, UA protein, part of routine urinalysis).

- For dipstick urine protein, urine specific gravity test needs to be collected if it is tested at the same time with the dipstick urine protein.
- If UACR is not documented in the EHR, but spot urine albumin and spot urine creatinine are documented. Please use spot urine albumin / spot urine creatinine from same sample to derive UACR. If the sample ID is not available, the same sample is defined as the lab order for spot urine albumin and spot urine creatinine is within 60 minutes of each other.
- If UPCR is not documented in the EHR, but spot urine protein and spot urine creatinine are documented. Please use spot urine protein / spot urine creatinine from same sample to derive UPCR. If the sample ID is not available, the same sample is defined as the lab order for spot urine protein and spot urine creatinine is within 60 minutes of each other.

4.2 Convert lab units and check categorical value range

- P24: mg/24hr
- UPCR: mg/g Cr
- A24: mg/24hr
- UACR: mg/g Cr
- Spot Urine Albumin: mg/dL
- Spot Urine Creatinine: mg/dL
- Spot Urine Protein: mg/dL
- Urine Specific Gravity: no unit
- UA protein can be one of the two following categorical values:
    - UA Protein: Negative, Trace, 1+, 2+, 3+, 4+
    - UA Protein: Negative, Trace, 10, 30, 100, 300, >=300 (some may have >600)

4.3 Albuminuria categories (A-staging) description and range (Figure 3)

4.3.1 UACR and A24 [1]

- A1: UACR < 30 mg/g Cr or A24 < 30 mg/24hr
- A2: UACR >= 30 mg/g Cr AND <= 300 mg/g Cr or A24 >= 30 mg/24hr AND <= 300 mg/24hr
- A3: UACR > 300 mg/g Cr or A24 > 300 mg/24hr

4.3.2 Columbia UPCR Ordinal Classifier

- If UPCR or P24 is 0, then A-staging is classified as A1
- If UPCR or P24 is not 0

- P(A1)=exp(13.136-2.497*log(UPCR)) / (1+exp(13.136-2.497*log(UPCR)))
- P(A1,A2)=exp(17.993-2.666*log(UPCR)) / (1+exp(17.993-2.666*log(UPCR)))
- P(A2) = P(A1,A2) - P(A1)
- P(A3) = 1- P(A1) - P(A2)
- A stage = MAX (P(A1), P(A2), P(A3))

4.3.3 Columbia UA Protein Ordinal Classifier if UA protein data range is (Negative, Trace, 1+, 2+, 3+, 4+)

- If both UA protein and SG are available
  - UA protein is NEGATIVE,
    - P(A1) = exp(-141.736+140.813*SG)/(1+exp(-141.736+140.813*SG))
    - P(A1,A2) = exp(-200.777+203.011*SG)/(1+exp(-200.777+203.011*SG))
  - UA protein is Trace,
    - P(A1) = exp(-143.142+140.813*SG) / (1+exp(-143.142+140.813*SG))
    - P(A1,A2) = exp(-202.959+203.011*SG) / (1+exp(-202.959+203.011*SG))
  - UA protein is 1+
    - P(A1) = exp(-145.145+140.813*SG) / (1+exp(-145.145+140.813*SG))
    - P(A1,A2) = exp(-204.642+203.011*SG) / (1+exp(-204.642+203.011*SG))
  - UA protein is 2+ or more
    - P(A1) = exp(-148.117+140.813*SG) / (1+exp(-148.117+140.813*SG))
    - P(A1,A2) = exp(-208.287+203.011*SG) / (1+exp(-208.287+203.011*SG))
  - P(A2) = P(A1,A2) - P(A1)
  - P(A3) = 1- P(A1) - P(A2)
  - A stage = MAX (P(A1), P(A2), P(A3))
- If only UA protein is available (i.e. the corresponding SG is missing)
  - UA protein of Negative corresponds to A1
  - UA protein of Trace corresponds to A1
  - UA protein of 1+ corresponds to A2
  - UA protein of 2+ or more corresponds to A3

4.3.4 Columbia UA Protein Ordinal Classifier if UA protein data range is (Negative, Trace, 10, 30, 100, 300, >=300)

- If both UA protein and SG are available
  - UA protein is NEGATIVE
    - P(A1)=exp(-129.764+129.454*SG)/(1+exp(-129.764+129.454*SG))
    - P(A1,A2)=exp(-198.543+201.365*SG)/(1+exp(-198.543+201.365*SG))
  - UA protein is Trace
    - P(A1)= exp(-143.109+140.777*SG)/(1+exp(-143.109+ 140.777*SG))

        - P(A1,A2)= exp(-218.272+231.444*SG)/(1+exp(-218.272+231.444*SG))
    - UA protein is 10
        - P(A1)=exp(-131.101+129.454*SG)/(1+exp(-131.101+129.454*SG))
        - P(A1,A2)=exp(-200.683+201.365*SG)/(1+exp(-200.683+201.365*SG))
    - UA protein is 30
        - P(A1)=exp(-133.02+129.454*SG)/(1+exp(-133.02+129.454*SG))
        - P(A1,A2)=exp(-203.25+201.365*SG)/(1+exp(-203.25+201.365*SG))
    - UA protein is 100 or more
        - P(A1)=exp(-136.286+129.454*SG)/(1+exp(-136.286+129.454*SG))
        - P(A1,A2)=exp(-206.478+201.365*SG)/(1+exp(-206.478+201.365*SG))
    - P(A2) = P(A1,A2) - P(A1)
    - P(A3) = 1- P(A1) - P(A2)
    - A stage = MAX (P(A1), P(A2), P(A3))
- If only UA protein is available (i.e. the corresponding SG is missing)
    - UA protein of Negative corresponds to A1
    - UA protein of Trace corresponds to A1
    - UA protein of 10 corresponds to A1
    - UA protein of 30 corresponds to A2
    - UA protein of 100 or more corresponds to A3

5. Implementation parameterized and modularized query (see CKDalgorithm_V4_queryTemplate.sql)

The implementation is written in parameterized and modularized query. The parameterization aspect deals with different terminologies and source data schemas by using parameter placeholders for codes and schema element names and then injecting corresponding terminology codes or schema/table/field names at execution time. The Modularization aspect is to build complex queries from simple query building blocks, separating concerns in different blocks.

The CKD query template (in sql server) includes both CKD algorithm implementation (Block 1 to Block 12) and co-variable data dictionary extraction (Block 13 to Block 21). The first block (Block 1) directly interacts with the Source database for data extraction and then stores data in a temporary table (#tmp). The following blocks only interact and extract data from the temporary table. So each institution needs to customize Block 1 on database schema and instantiate the coding parameters for the implementation. The query template intends to assist portable implementation. If you have any questions, please contact the algorithm developers for assistance.

V. Collection of covariates

1. Demographic, phenotyping results and comorbidity
    - Age
    - Sex
    - Race

- Ethnicity
- Case_Control_Unknown_Status
- Case_Control_Unknown_Category
- G-staging
- A-staging
- Type 2 Diabetes
- Hypertension

2. Serum creatinine and corresponding eGFR
3. Urine protein test results
4. Relevant procedure information (dialysis, transplant)
5. Relevant diagnosis information (AKI, prerenal kidney injury, sepsis, volume depletion, shock, CKD, dialysis and other kidney disease)

Detailed information please check CKD_dd_V4.xls file.

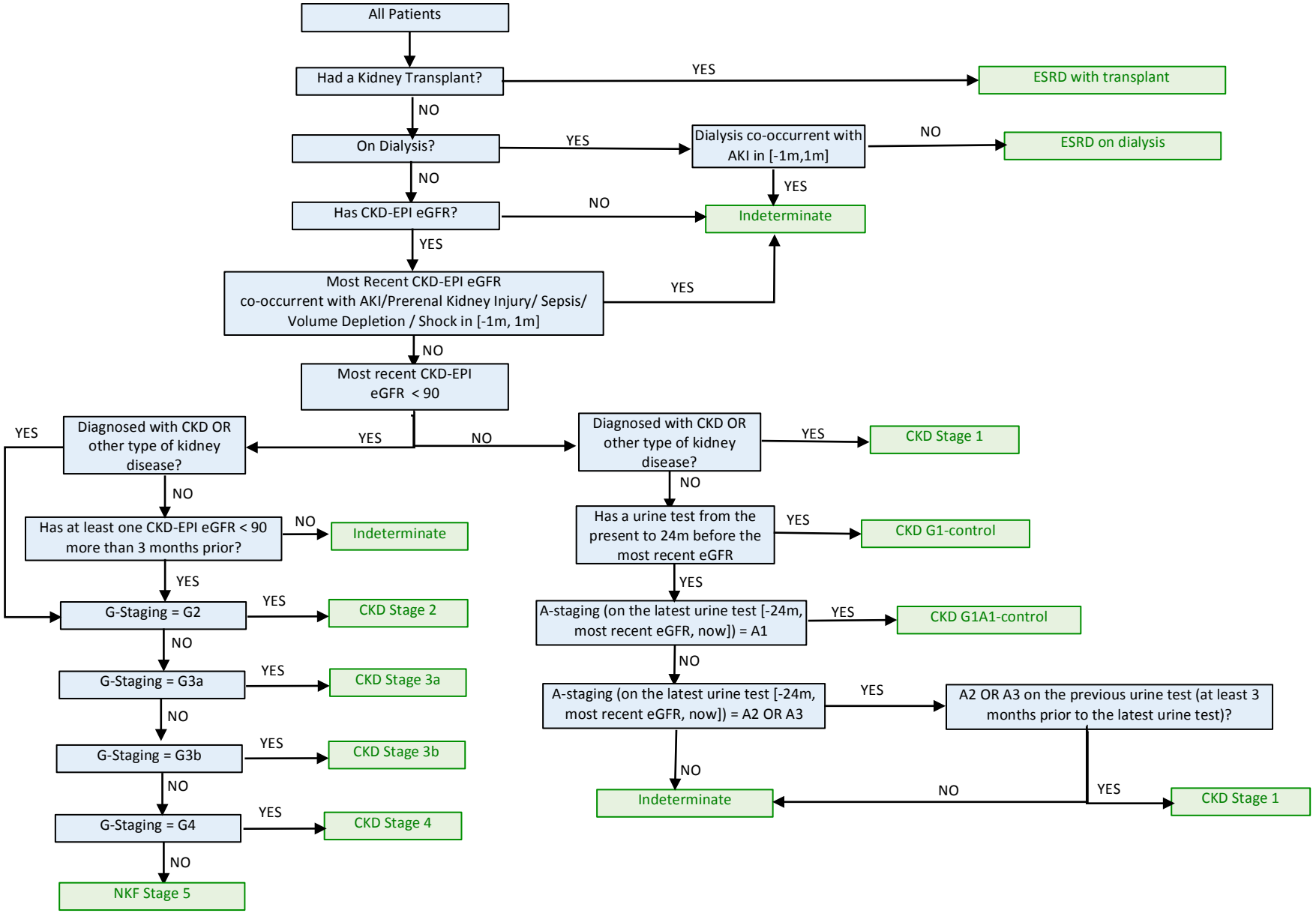

Figure 1 CKD Phenotyping Algorithm Definition V4 Flowchart

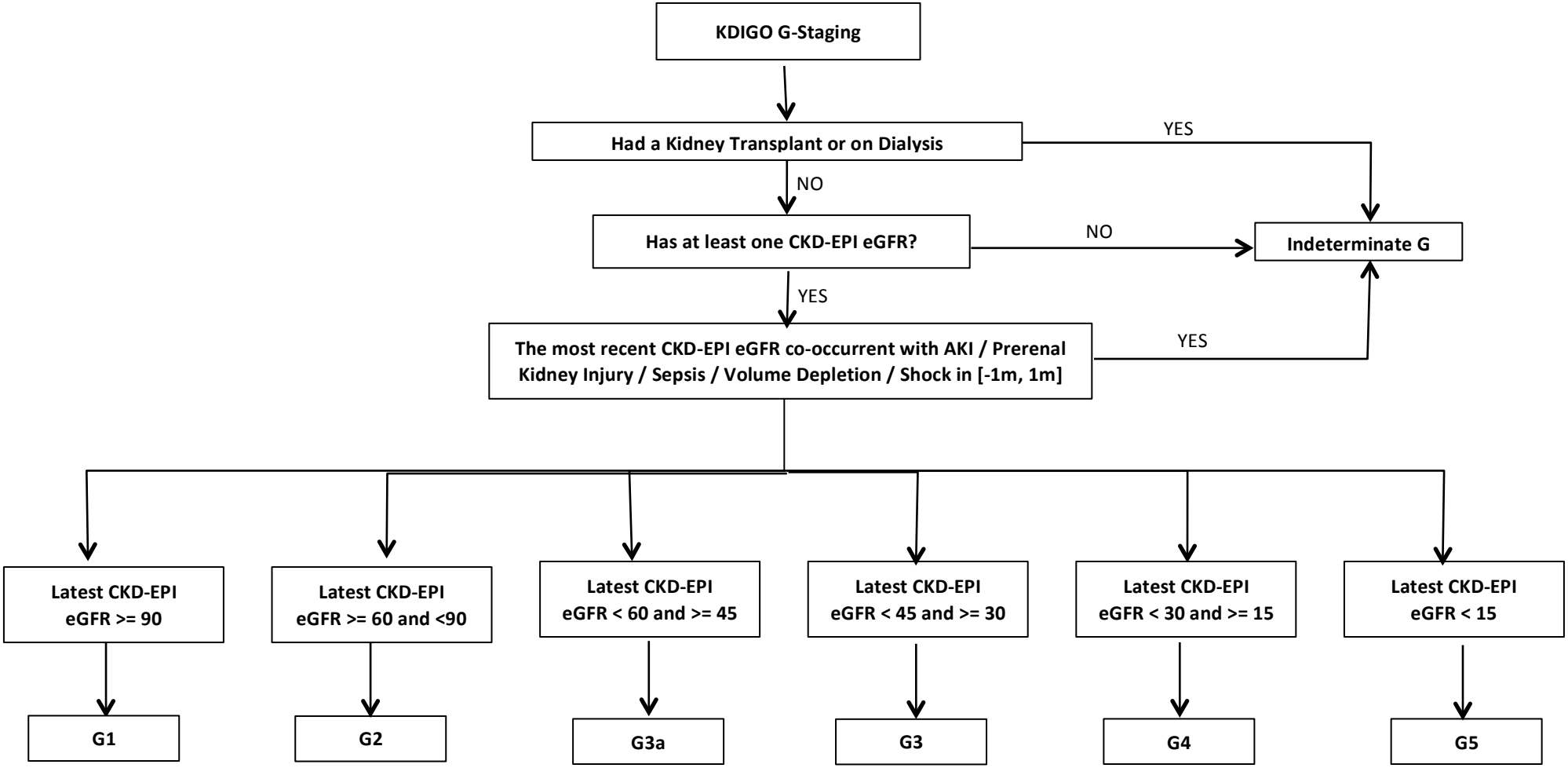

Figure 2 GFR categories (G-Staging) description and range

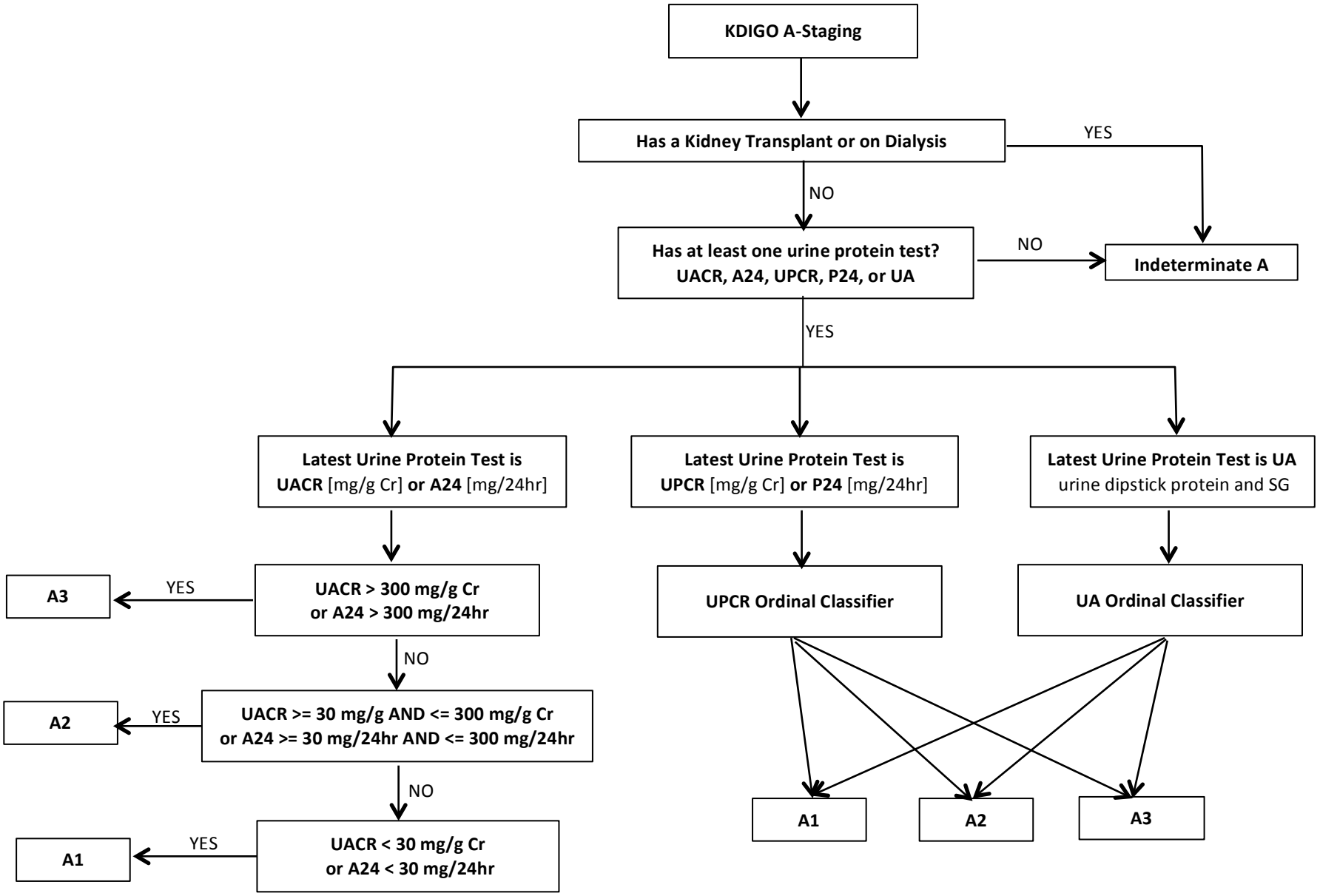

Figure 3 Albuminuria categories (A-Staging) description and range

Table 1 Codes used in the algorithm*

Data Element	Coding Variable (used in CKDalgorithm_V4_coding.txt)
Gender	
Race	
Ethnicity	
Height	lab.height.loinc.cm
Acute Kidney Injury (AKI) diagnosis	dx.aki.icd9, dx.aki.icd10, dx.aki.snomed
Prerenal Kidney Injury diagnosis	dx.prerenalInjury.icd9, dx.prerenalInjury.icd10
Sepsis diagnosis	dx.sepsis.icd9, dx.sepsis.icd10, dx.sepsis.snomed
Volume Depletion diagnosis	dx.volumeDepletion.icd9, dx.volumeDepletion.icd10, dx.volumeDepletion.snomed
Shock diagnosis	dx.shock.icd9, dx.shock.icd10, dx.shock.snomed
Chronic kidney disease (CKD) diagnosis	dx.ckd.icd9, dx.ckd.icd10, dx.ckd.snomed
Other kidney disease diagnosis	dx.otherKidneyDis.icd9, dx.otherKidneyDis.icd10, dx.otherKidneyDis.snomed
Dialysis diagnosis	dx.dialysis.icd9, dx.dialysis.icd10, dx.dialysis.snomed,
Dialysis procedure	proc.dialysis.cpt4, proc.dialysis.Icd9, proc.dialysis.icd10
Kidney transplant diagnosis	dx.kidneyTransplant.icd9, dx.kidneyTransplant.icd10, dx.kidneyTransplant.snomed,
Kidney transplant procedure	proc.kidneyTransplant.cpt4, proc.kidneyTransplant.Icd9, proc.kidneyTransplant.icd10
24-hour urine albumin (A24) test	lab.A24.loinc
24-hour urine protein (P24) test	lab.P24.loinc
Serum creatinine test	lab.serumCreatinine.loinc
Urine albumin-to-creatinine ratio (UACR) test	lab.Uacr.loinc
Urine protein-to-creatinine ratio (UPCR) test	lab.Upcr.loinc
Urinalysis (UA) protein test	lab.uaProtein.loinc
Urine specific gravity test	lab.specificGravity.loinc
Spot urine albumin test	lab.spotUrineAlbumin.loinc
Spot urine protein test	lab.spotUrineProtein.loinc
Spot urine creatinine test	lab.spotUrineCr.loinc

*Detailed codes are provided in CKDalgorithm_V4.2_coding.txt. This file is a machine-readable file. In addition to concept code, concept name, vocabulary and source code (since ICD10 codes are converted from ICD9 codes, so the source code lists the ICD9 codes where these ICD10 codes are mapped from), it also includes query template parameter name. The query template parameter name is the coding variable names that are used in the CKD parameterized and modularized query.

References:

1 Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney inter* 2013;**Suppl.**:1–150.

2 Levey AS, Stevens LA, Schmid CH, *et al.* A New Equation to Estimate Glomerular Filtration Rate. *Ann Intern Med* 2009;**150**:604–12.

3 Schwartz GJ, Muñoz A, Schneider MF, *et al.* New Equations to Estimate GFR in Children with CKD. *J Am Soc Nephrol* 2009;**20**:629–37. doi:10.1681/ASN.2008030287

4 Schwartz GJ, Work DF. Measurement and Estimation of GFR in Children and Adolescents. *Clin J Am Soc Nephrol* 2009;**4**:1832–43. doi:10.2215/CJN.01640309