Supplementary Material

Supplementary Table 1: ICD codes used to classify patients.

Medical Condition	ICD-9 code	ICD-10 code
Type 2 Diabetes	250.x0, 250.x2	E11
Type 1 Diabetes	250.x1, 250.x3	E10, O24.0
Kidney Transplant	V42.0	Z94.0
Chronic Kidney Disease	585.9	N18.9

Supplementary Table 2: Hyperparameters tuned for each MLA model.

Random forest (RF)	○ Max_depth = 12, the maximum depth of each tree ○ N_estimators = 580, the number of trees in the ensemble ○ Criterion (one of Gini or Entropy) = Entropy, the criteria used to evaluate utility of splits ○ Max_features (one of Log2 or Sqrt) = Sqrt, the number of features to consider when looking for the best split ○ Class_weight (one of balanced or balanced subsample) = balanced, the weight given to the positive and negative classes when calculating the training loss ○ Min_samples_split = 208, the minimum number of examples found in a node to be able to split the node further

Gradient boosted tree model XGBoost (XGB)	○ Eta = 0.20, learning rate ○ Gamma = 4.44, loss reduction require to split a node in the tree ○ Max_depth = 11, the maximum depth in each tree ○ Lambda = 15.2, the strength of L2-regularization ○ Scale_pos_weight = 21.3, the weight for each positive example when calculating the training loss ○ Min_child_weight = 202.7, the minimum sum of the instance weight for a node to be able to split ○ Use Preprocessing? = True, whether or not to impute and scale the data before input to the model

Supplementary Table 3. Demographic information and comorbidities for the hold-out test dataset.

	Test Demographics	No CKD	Any Stage CKD	Stage ≥3 CKD	Stage ≥4 CKD
Age					
	18-29 yo	121 (2.2%)	19 (0.9%)	7 (0.6%)	140 (1.8%)
	30-39 yo	381 (6.9%)	56 (2.6%)	14 (1.1%)	437 (5.7%)
	40-49 yo	1,058 (19.1%)	227 (10.7%)	94 (7.5%)	1,285 (16.8%)
	50-59 yo	1,687 (30.5%)	478 (22.6%)	236 (18.9%)	2,165 (28.3%)
	60-69 yo	1,415 (25.6%)	621 (29.3%)	407 (32.6%)	2,036 (26.6%)
	≥70 yo	875 (15.8%)	718 (33.9%)	492 (39.4%)	1,593 (20.8%)
Sex					
	Female	3,059 (55.2%)	1,073 (50.6%)	619 (49.5%)	4,132 (54.0%)
	Male	2,478 (44.8%)	1,046 (49.4%)	631 (50.5%)	3,524 (46.0%)
Ethnicity					
	Caucasian	4,330 (78.2%)	1,714 (80.9%)	1,022 (81.8%)	6,044 (78.9%)
	African American	624 (11.3%)	282 (13.3%)	169 (13.5%)	906 (11.8%)
	Asian	138 (2.5%)	30 (1.4%)	11 (0.9%)	168 (2.2%)

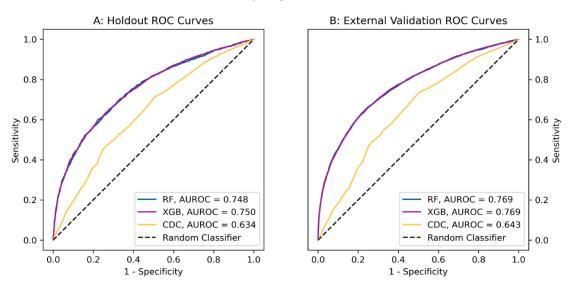|	Other/Unknown Race	445 (8.0%)	93 (4.4%)	48 (3.8%)	538 (7.0%)
Comorbidities					
	(Presence of) Coronary Artery Bypass Graft	68 (0.5%)	52 (1.0%)	34 (1.2%)	120 (0.7%)
	Acute Kidney Injury	44 (0.3%)	41 (0.8%)	30 (1.0%)	85 (0.5%)
	Anemia	544 (4.3%)	245 (4.9%)	151 (5.1%)	789 (4.5%)
	Arrhythmia	614 (4.8%)	302 (6.0%)	179 (6.1%)	916 (5.2%)
	Cancer/Solid Tumor	412 (3.2%)	191 (3.8%)	122 (4.2%)	603 (3.4%)
	Cardiovascular Disease (CVD)	713 (5.6%)	392 (7.8%)	249 (8.5%)	1,105 (6.2%)
	Congestive Heart Failure (CHF)	200 (1.6%)	129 (2.6%)	91 (3.1%)	329 (1.9%)
	Dyslipidemia	2,604 (20.5%)	918 (18.3%)	522 (17.8%)	3,522 (19.9%)
	Heart Valve Disease	225 (1.8%)	118 (2.4%)	74 (2.5%)	343 (1.9%)
	Hyperglycemia	2,137 (16.8%)	571 (11.4%)	304 (10.4%)	2,708 (15.3%)
	Hypertension	2,584 (20.3%)	996 (19.9%)	567 (19.3%)	3,580 (20.2%)
	Obesity	1,282 (10.1%)	328 (6.5%)	153 (5.2%)	1,610 (9.1%)
	Peripheral Vascular Disease (PVD)	133 (1.0%)	83 (1.7%)	56 (1.9%)	216 (1.2%)
	Previous Percutaneous transluminal coronary angioplasty	92 (0.7%)	50 (1.0%)	30 (1.0%)	142 (0.8%)
	Prior MI	159 (1.3%)	84 (1.7%)	50 (1.7%)	243 (1.4%)
	Proteinuria	43 (0.3%)	30 (0.6%)	18 (0.6%)	73 (0.4%)
	Stroke/Transient Ischemic Attack	218 (1.7%)	117 (2.3%)	77 (2.6%)	335 (1.9%)
	Type I Diabetes	183 (1.4%)	98 (2.0%)	57 (1.9%)	281 (1.6%)

Supplementary Table 4. Demographic information and comorbidities for the external test dataset.

External Demographics	**No CKD**	**Any Stage CKD**	**Stage ≥3 CKD**	**Stage ≥4 CKD**
Age				
18-29 yo	358 (2.3%)	66 (0.9%)	18 (0.4%)	10 (0.8%)
30-39 yo	1,122 (7.2%)	236 (3.1%)	99 (2.0%)	34 (2.6%)
40-49 yo	2,906 (18.7%)	700 (9.3%)	336 (6.7%)	112 (8.4%)

	50-59 yo	4,819 (31.0%)	1,683 (22.3%)	972 (19.5%)	271 (20.4%)
	60-69 yo	3,669 (23.6%)	2,114 (28.0%)	1,451 (29.1%)	363 (27.3%)
	≥70 yo	2,657 (17.1%)	2,743 (36.4%)	2,111 (42.3%)	539 (40.6%)
Sex					
	Female	8,381 (54.0%)	3,843 (51.0%)	2,578 (51.7%)	672 (50.6%)
	Male	7,150 (46.0%)	3,699 (49.0%)	2,409 (48.3%)	657 (49.4%)
Ethnicity					
	Caucasian	12,061 (77.7%)	5,890 (78.1%)	3,921 (78.6%)	956 (71.9%)
	African American	2,652 (17.1%)	1,418 (18.8%)	920 (18.4%)	323 (24.3%)
	Asian	238 (1.5%)	72 (1.0%)	44 (0.9%)	14 (1.1%)
	Other/Unknown Race	580 (3.7%)	162 (2.1%)	102 (2.0%)	36 (2.7%)
Comorbidities					
	(Presence of) Coronary Artery Bypass Graft	250 (0.8%)	183 (1.2%)	128 (1.2%)	28 (1.1%)
	Acute Kidney Injury	141 (0.4%)	128 (0.8%)	86 (0.8%)	16 (0.6%)
	Anemia	1,583 (4.8%)	825 (5.3%)	581 (5.6%)	151 (6.1%)
	Arrhythmia	1,660 (5.0%)	1,049 (6.7%)	742 (7.2%)	193 (7.8%)
	Cancer/Solid Tumor	1,500 (4.5%)	801 (5.1%)	555 (5.4%)	123 (5.0%)
	Cardiovascular Disease (CVD)	2,208 (6.6%)	1,510 (9.6%)	1,065 (10.3%)	266 (10.7%)
	Congestive Heart Failure (CHF)	513 (1.5%)	515 (3.3%)	388 (3.7%)	133 (5.4%)
	Dyslipidemia	6,456 (19.4%)	2,502 (16.0%)	1,548 (15.0%)	330 (13.3%)
	Heart Valve Disease	704 (2.1%)	450 (2.9%)	317 (3.1%)	69 (2.8%)
	Hyperglycemia	5,306 (16.0%)	1,423 (9.1%)	813 (7.9%)	130 (5.2%)
	Hypertension	6,444 (19.4%)	2,885 (18.4%)	1,882 (18.2%)	425 (17.1%)
	Obesity	2,942 (8.8%)	930 (5.9%)	541 (5.2%)	105 (4.2%)
	Peripheral Vascular Disease (PVD)	396 (1.2%)	334 (2.1%)	246 (2.4%)	73 (2.9%)
	Previous Percutaneous transluminal coronary angioplasty	301 (0.9%)	165 (1.1%)	123 (1.2%)	23 (0.9%)
	Prior MI	393 (1.2%)	245 (1.6%)	169 (1.6%)	36 (1.5%)
	Proteinuria	122 (0.4%)	101 (0.6%)	59 (0.6%)	17 (0.7%)
	Stroke/Transient Ischemic Attack	671 (2.0%)	439 (2.8%)	306 (3.0%)	78 (3.1%)

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

BMJ Open Diab Res Care

	Type I Diabetes	527 (1.6%)	306 (2.0%)	190 (1.8%)	71 (2.9%)

Supplementary Figure 1. Area under the receiver operating characteristic curve (AUROC) plots of machine learning (ML) models random forest (RF) and gradient boosted tree (XGB), and Centers for Disease Control and Prevention (CDC) CKD scoring system for a) hold-out dataset and b) external validation dataset for prediction of any-stage DKD in the five years following T2DM diagnosis. A random classifier was used as the baseline

.

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

BMJ Open Diab Res Care

Supplementary Figure 2. Area under the receiver operating characteristic curve (AUROC) plots of machine learning (ML) models random forest (RF) and gradient boosted tree (XGB), and Centers for Disease Control and Prevention (CDC) CKD scoring system for a) hold-out dataset and b) external validation dataset for prediction of DKD stages 4-5 in the five years following T2DM diagnosis. A random classifier was used as the baseline.