

1 **Estimating and predicting kidney function decline in the general population**

2

3 Running title: Estimating and Predicting Kidney Function Decline

4

5 Masao Iwagami, PhD^{1,2,3*}; Kazunori Odani, MSc^{4*}; Tomoki Saito, BSc⁴

6 ¹Department of Health Services Research, Institute of Medicine, University of Tsukuba,

7 Ibaraki, Japan

8 ²Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical

9 Medicine, London, UK

10 ³International Institute for Integrative Sleep Medicine (IIIS), University of Tsukuba, Ibaraki,

11 Japan

12 ⁴JMDC Inc., Tokyo, Japan

13 *These authors contributed equally

14

15 **Corresponding authors:**

16 Masao Iwagami, MD, MPH, MSc, PhD

17 Department of Health Services Research, Institute of Medicine, University of Tsukuba, 1-1-1

18 Tennodai, Tsukuba, Ibaraki, 305-8575, Japan

19 Tel: +81-29-853-8849

20 Fax: +81-29-853-8849

21 Email: iwagami-ky@umin.ac.jp

22

23 Tomoki Saito, BSc

24 JMDC Inc.

25 Sumitomo Shiba Daimon Building 12F, 2-5-5 Shiba Daimon, Minato-ku, Tokyo, 108-0072,

26 Japan

27 Tel: +81-3-5733-5010

28 Fax: +81-3-5733-5010

29 Email: tosaito@jmhc.co.jp

30

31 **Funding:** none (self-funded)

32

33 **Keywords:** chronic kidney disease; estimated glomerular filtration rate; creatinine; machine

34 learning; prediction; health checkup

35 **Abstract**

36 **Introduction:** We aimed to estimate the rate of kidney function decline over 10 years in the
37 general population and develop a machine learning model to predict it.

38 **Methods:** We used the JMDC database from 2012 to 2021, which includes company
39 employees and their family members in Japan, where annual health checks are mandated for
40 people aged 40–74 years. We estimated the slope (average change) of estimated glomerular
41 filtration rate (eGFR) over a period of 10 years. Then, using the annual health-check results
42 and prescription claims for the first five years from 2012 to 2016 as predictor variables, we
43 developed an XGBoost model, evaluated its prediction performance with the root mean
44 squared error (RMSE), R^2 , and area under the receiver operating characteristic curve
45 (AUROC) for rapid decliners (defined as the slope < -3 ml/min/1.73 m²/year) using 5-fold cross
46 validation, and compared these indicators with those of the linear regression model using only
47 eGFR data from 2012 to 2016.

48 **Results:** We included 126 424 individuals (mean age, 45.2 years; male, 82.4%; mean eGFR,
49 79.0 ml/min/1.73 m² in 2016). The mean slope was -0.89 (standard deviation, 0.96)
50 ml/min/1.73 m²/year. The predictive performance of the XGBoost model (RMSE, 0.78; R^2 ,
51 0.35; and AUROC, 0.89) was better than that of the linear regression model using only eGFR
52 data (RMSE, 1.94; R^2 , -3.03; and AUROC, 0.79).

53 **Conclusion:** Application of machine learning to annual health-check and claims data could

54 predict the rate of kidney function decline, whereas the linear regression model using only

55 eGFR data did not work.

56

57 **Keywords:** chronic kidney disease; estimated glomerular filtration rate; creatinine; machine

58 learning; prediction; health checkup

59 **Introduction**

60 Chronic kidney disease (CKD) is a large burden in the society as it is associated with
61 increased risk of cardiovascular and non-cardiovascular diseases, as well as the health care
62 costs, especially if patients require renal replacement therapy (RRT).¹⁻⁴ Most people in the
63 general population have normal kidney function (i.e., glomerular filtration rate [GFR]) at
64 birth, whereas the GFR naturally decreases with age, with faster decline among people with
65 risk factors such as diabetes.⁵ An old study estimated that the rate of GFR decline was -0.75
66 ml/min/year in the generally healthy population.⁶ Since then, there have been a number of
67 studies estimating the rate of kidney function decline,⁷ but their study periods are often short
68 and only a few studies targeting the general population without CKD are reported.

69 A number of clinical trials and observational studies have set the study endpoints as
70 the time to dialysis initiation or the time to a 30% or 40% drop in estimated GFR (eGFR),
71 mostly among patients at high risk for these events, such as those with late stage CKD.^{8,9}
72 However, the incidence of these outcomes is low in the early stage of CKD or in the general
73 population.⁹ Meanwhile, the rate of kidney function decline or slope (average change) of
74 eGFR can be calculated for individuals and could be a surrogate endpoint for clinical trials,
75 even in the early stage of CKD and in the general population.¹⁰⁻¹⁴

76 In observational studies, prediction models have been developed for the time to
77 dialysis initiation¹⁵⁻¹⁷ or the time to a 30% or 40% drop in eGFR,¹⁸⁻²⁰ showing good

78 discrimination ability and/or calibration. However, to the best of our knowledge, no previous
79 study has developed a prediction model for the rate of kidney function decline as a
80 continuous variable. Such prediction model would be useful for stratifying the general
81 population and identifying those with rapid decline in kidney function. To date, no consensus
82 has been reached on the definition of rapid decliners,⁷ and therefore, the prediction of a
83 continuous (rather than a dichotomous) outcome would have a wider application.

84 In Japan, the government introduced a specific health checkup system in 2008,
85 which obliges all insurers to provide annual health checkups for insured persons aged 40–74
86 years.²¹ Notably, under employee insurance, the attendance rates for annual checkups are
87 high, approximately >80% (>90% among men) among company employees.²² Utilizing this
88 situation, we aimed (i) to estimate the rate of kidney function decline in a period of 10 years
89 using data obtained from the JMDC database, a large database of large and middle-scale
90 companies and their family members in Japan and (ii) to develop a prediction model based on
91 annual health checkup data and claims for the first five years. Machine learning has been used
92 to handle a large number of candidate predictor variables and their potential interactions.

93

94 **Methods**

95 *Data source*

96 The details of the JMDC database have been described elsewhere.²³ In brief, the JMDC
97 database was developed by the JMDC Co. This database is a large-scale database covering
98 Japanese health insurance union members, including employees of large- and middle-scale
99 companies and their family members aged <75 years; it includes all claims for outpatient
100 treatment, hospitalization, and prescriptions and dispensations of drugs, as well as the results
101 of annual health checkups. Annual health checkups are required by law for insured persons
102 aged 40–74 years,²¹ whereas those aged <40 years can also undergo annual checkups. Annual
103 health checkups are usually conducted in the facilities of health insurance unions with which
104 the companies are affiliated. The details of the annual health checkups are listed in the
105 “Predictor variables” section below. Serum creatinine measurement is optional but depends
106 on the decision of each health insurance union rather than on the medical conditions of the
107 participants. For the present study, we used the most recent 10-year data from April 2012 to
108 March 2022 (i.e., from 2012 to 2021 financial years).

109 The data used in this study were anonymized and processed anonymously by JMDC,
110 Inc. This study was approved by the Ethics Committee of The Research Institute of
111 Healthcare Data Science (Date of approval, October 30, 2023; Approval number, RI
112 2023003).

113

114 ***Study population***

115 First, in the JMDC database, we identified people with annual health checkup results
116 (including serum creatinine) for five consecutive years, from 2012 to 2016. We excluded
117 patients receiving RRT (identified as Japanese procedure codes J038 for hemodialysis, J042
118 for peritoneal dialysis, and K780 for kidney transplantation) from 2012 to 2016 or those with
119 an eGFR <15 ml/min/1.73 m² in 2016. Among the remaining individuals, we further
120 identified those with annual health checkup results (including serum creatinine) for the latter
121 five consecutive years, from 2017 to 2021. We identified and excluded patients who
122 underwent RRT between 2017 and 2021 because their serum creatinine levels did not reflect
123 their GFRs.

124 Consequently, the study population consisted of people with annual health checkup
125 results (including serum creatinine) for 10 consecutive years, from 2012 to 2021, who did not
126 receive RRT.

127

128 ***Outcome definition***

129 The outcome of interest was the slope (average change) of eGFR during the 10 years from
130 2012 to 2021, which was estimated using unadjusted linear regression. eGFR was calculated
131 using the following Japanese estimation formula²⁴:

132
$$\text{eGFR} = 194 \times \text{Cr}^{-1.094} \times \text{Age}^{-0.287} (\times 0.739 \text{ for women})$$

133

134 ***Predictor variables***

135 We used the annual health checkup results for the first five years, from 2012 to 2016. As
136 demonstrated prior,²³ the mandatory annual health checkups in Japan generally include both
137 objective and subjective (self-reported) findings. Objective findings include body mass index
138 (BMI), abdominal circumference, systolic blood pressure (sBP), diastolic blood pressure
139 (dBP), triglyceride (TG), high density lipoprotein (HDL) cholesterol, low density lipoprotein
140 (LDL) cholesterol, total cholesterol, aspartate aminotransferase (ALT), alanine
141 aminotransferase (ALT), gamma glutamyl transpeptidase (γ -GTP), fasting blood sugar,
142 casual blood sugar, hemoglobin A1c (HbA1c) according to the National Glycohemoglobin
143 Standardization Program, hematocrit, hemoglobin content, erythrocyte count, serum uric acid,
144 urinary sugar (dipstick test), and uric protein (dipstick test). Among subjective (self-reported)
145 findings,²³ we used the information pertaining to current smoking status (yes or no), drinking
146 habits (every day, sometimes, or rarely/none), and exercise habit (yes or no for ≥ 2
147 times/week for ≥ 30 min in the past year).

148 In addition, using the prescription records in the medical claims, we identified the
149 presence or absence in the use of lipid-lowering agents (any), statins, antidiabetic drugs (any),
150 sodium-glucose transport protein 2 (SGLT2) inhibitors, antiplatelet drugs, antihypertensive
151 drugs (any), and angiotensin converting enzyme inhibitors (ACEI) or angiotensin II receptor

152 blockers (ARB), which are recorded as the Anatomical Therapeutic Chemical (ATC)

153 Classification codes (**Supplementary Table S1**).

154 **Table 1** displays the list of predictor variables and their distributions (mean and
155 standard deviation [SD] for continuous variables and number and percentage for categorical
156 variables) in 2016, whereas **Supplementary Table S2** shows all predictor variables from
157 2012 to 2016 that were used for prediction. In addition, the slope (average change) of eGFR
158 during the five years was used for prediction.

159

160 *Statistical analysis*

161 First, we showed the distribution of outcome variable (i.e., the slope of eGFR during the 10
162 years) and estimated the mean and standard deviation, overall and by age group (<40, 40–49,
163 50–59, and ≥ 60 years), sex, and Kidney Disease Improving Global Outcomes (KDIGO) GFR
164 stages (eGFR ≥ 90 , 60–89, 45–59, 30–44, and 15–29 ml/min/1.73 m²) in 2016.

165 For the model development, we used the XGBoost regression model²⁵, because it is
166 generally known to show high predictive performance in the case of table data. The
167 implementation was based on the “xgboost” package (version: 1.7.5) of Python. For the
168 hyperparameters, *eta* (step size shrinkage used in the update to prevent overfitting) was set to
169 0.05, *subsample* (i.e., subsample ratio of the training instances) was set to 0.9, and
170 *colsample_bytree* (i.e., subsampling of columns) was set to 0.8. With a grid search, the

171 *max_depth* (i.e., maximum depth of a tree) and *min_child_weight* (i.e., minimum sum of
172 instance weights [Hessian] needed in a child) were set to 4 and 16, respectively. *n_estimators*
173 was set by early stopping. The other hyperparameters were set to be default values
174 (“XGBoost-link”). We input the annual health checkup data and prescription data for 2016
175 into the model as they are. For data from 2012 to 2015, we input the subjective (self-reported)
176 findings and prescription data as they are, whereas we calculated and used the difference in
177 values of objective findings between each year and 2016 for each individual. Missing values
178 were input into the XGBoost model as they are.

179 For model validation, using the 5-fold cross validation, we evaluated the root mean
180 squared error (RMSE) and R^2 as the prediction performance:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

181 (y_i : actual value, \hat{y}_i : predicted value, \bar{y} : average of actual values)

182

183 In addition, considering that the identification of rapid decliners (defined in the present study
184 as the slope < -3 ml/min/ 1.73 m²/year¹⁴) is clinically important, we estimated the area under
185 the receiver operating characteristic curve (AUROC) for the rapid decliners using 5-fold

186 cross validation. For cross-validation, we merged the validation data of each fold to estimate
187 the RMSE, R^2 , and AUROC of the entire training dataset.

188 For comparison, we also estimated these indicators (i.e., RMSE, R^2 , and AUROC) in
189 a crude linear regression model based only on the eGFR values from 2012 to 2016. We aimed
190 to understand the extent to which our XGBoost model was superior to the simplest linear
191 regression model.

192 To interpret the model, we used the Shapley Addictive explanation (SHAP) function
193 in the XGBoost model.²⁶ We estimated the SHAP feature importance, which is the mean of
194 the absolute SHAP values (i.e., the contribution of each feature to the outcome). For the
195 predictor variables with higher feature importance, we depicted SHAP dependence plots to
196 determine the impact of the increase or decrease in each feature on the predicted value.

197 All data management and statistical analyses were performed using Python (version
198 3.8.10).

199

200 **Results**

201 In the JMDC database, we identified 183 485 people who underwent annual health checkups
202 for five consecutive years from 2012 to 2016 (**Supplementary Figure S1**). We then excluded
203 107 patients who underwent RRT and 28 patients with eGFR <15 ml/min/1.73 m². Of the
204 remaining 183 350 people, we excluded 56 926 without annual health checkup results

205 (including serum creatinine) for five consecutive years from 2017 to 2021, including 109
206 patients who received RRT during this period. The remaining 126 424 individuals with
207 consecutive measurements of serum creatinine levels from 2012 to 2021 were included in the
208 subsequent analyses. A comparison of the predictor variables between the two groups (i.e.,
209 126 424 and 56 926 people with and without annual health checkup results from 2017 to
210 2021) suggested that older people and women were more likely to retire and lose employee
211 insurance to quit from the JMDC database, whereas the annual health checkup results from
212 2012 to 2016 were not very different between the two groups (**Supplementary Table S2**).
213 The mean age of the 126 424 people was 45.2 (SD, 8.6) years, 104 219 (82.4%) were male,
214 and the mean eGFR was 79.0 (SD, 13.4) ml/min/1.73 m². According to the KDIGO GFR
215 categories, 23 850 (18.9%), 95 516 (75.6%), 6 781 (5.4%), 247 (0.2%), and 30 (0.02%)
216 accounted for stages G1, G2, G3a, G3b, and G4 (eGFR ≥90, 60–89, 45–59, 30–44, and
217 15–29 ml/min/1.73 m²), respectively.

218 **Figure 1** shows the distribution of eGFR slope for the 10 years, with a mean value
219 of -0.89 (SD, 0.96) ml/min/1.73 m²/year. Among the 126 424 people, 2 511 (2.0%) were
220 rapid decliners, with <-3 ml/min/1.73 m²/year. The distributions of eGFR slope by age group,
221 sex, and KDIGO GFR stage are shown in **Supplementary Table S3**.

222 **Figure 2** shows scatter plots of the predicted and actual values in each model,
223 suggesting that the prediction was better in the developed XGBoost model than in the linear

224 regression model using only eGFR data from 2012 to 2016. The RMSE and R^2 of the
225 developed XGBoost model were 0.78 and 0.35, respectively, while those of the linear
226 regression model using only eGFR data were 1.94 and -3.03, respectively. To discriminate
227 the rapid decliners, the AUROC of the XGBoost model was 0.89, whereas that of the linear
228 regression model using only eGFR data was 0.79. The ROC curves are shown in **Figure 3**.

229 **Figure 4** shows the ranking of the predictor variables with higher feature importance
230 in the developed XGBoost model. The eGFR-related features, particularly the eGFR slope
231 from 2012 to 2016 and the eGFR in 2016, ranked high. Several parameters measured in the
232 annual health checkups, including hematocrit, HbA1c, γ -GTP, BMI, HDL cholesterol, serum
233 uric acid, and dBp, were also ranked. **Figure 5** shows the SHAP dependence plots of the
234 selected predictor variables with more important features (the difference in values between
235 2012–2015 and 2016 is not shown because its interpretation was difficult). The model learned
236 to reduce the predicted value (i.e., slope of eGFR for the 10 years) when the slope in the first
237 five years was smaller, and when the eGFR in 2016 was higher in the range above
238 approximately 60 ml/min/1.73 m². The figures also suggest negative associations with
239 Hb1Ac (in the range of above approximately 6.5%), serum uric acid, and dBp (in the range of
240 above approximately 80 mmHg) and positive associations with hematocrit, γ -GTP, and HDL
241 cholesterol. There was a U-shaped association with BMI. Men were more likely to show
242 lower predictive values than women.

243

244 **Discussion**

245 Using the JMDC claims database covering the general population, we estimated the rate of
246 kidney function decline over 10 years and developed a machine learning model (XGBoost
247 model) to predict this decline based on annual health checkups and prescription data for the
248 first five years. The predictive performance of the model was good or moderate (RMSE,
249 0.78; R^2 , 0.35; and AUROC for the rapid decliners, 0.89), whereas the linear regression
250 model using only eGFR values did not work (RMSE, 1.94; R^2 , -3.03; and AUROC for the
251 rapid decliners, 0.79). The top features of the developed model were dominated by
252 eGFR-related features, whereas known risk factors for kidney function decline, such as
253 HbA1c, contributed to the prediction.

254 Among the study population consisting of large- or medium-sized company
255 employees and their family members, most of whom had normal range of kidney function,
256 the mean slope of eGFR for the 10 years was -0.89 (SD, 0.96) ml/min/1.73 m²/year. This was
257 similar to an old US study estimating the mean decrease in creatinine clearance to be -0.75
258 ml/min/year among normal volunteers.⁶ In a recent study conducted in a single health checkup
259 center of Japan, excluding patients with any comorbidities, the mean eGFR decline rate was
260 -1.07 (SD 0.42) ml/min/1.73 m²/year.²⁷ Those with higher eGFR at baseline were generally

261 more likely to show faster decline, which is in line with the present study (**Supplementary**
262 **Table S3**).

263 To the best of our knowledge, this is the first study to predict the rate of kidney
264 function decline as a continuous variable. First, we found it difficult to predict the rate of
265 kidney function decline over 10 years from only eGFR values for the first five years, with an
266 R^2 of -3.03, which is even worse than a prediction assuming the average value for all
267 individuals (i.e., R^2 of 0). In other words, the correlation between the slope for the 10 years
268 and that for the first five years was weak. This may be partly due to fluctuations in measured
269 serum creatinine over time, although mandatory annual health checkups in Japanese
270 companies are usually conducted at the same place every year and when participants are not
271 sick. Another reason could be that the rate of kidney function decline is also affected by
272 patient demographics; known risk factors for CKD or ESRD, such as diabetes and
273 hypertension; lifestyle factors, such as smoking and exercise; and drugs protecting the
274 kidneys, such as ACEI/ARB and SGLT2 inhibitors. Therefore, we additionally used these
275 variables as predictors and observed a large increase in R^2 to 0.35. However, the predictive
276 ability was far from perfect (R^2 of 1.00). The remaining possibility is that there may be
277 measurement errors in predictor variables or unknown factors predicting the rate of kidney
278 function decline. Meanwhile, the discrimination ability of the XGBoost model for the rapid

279 decliners (eGFR slope < -3 ml/min/1.73 m²/year) was seemingly very good, with an AUROC
280 of 0.89.

281 The feature importance and SHAP dependence plots in the established XGBoost
282 model were remarkable. First, the coefficient (slope) of eGFR for the first five years
283 correlated with the outcome in the same direction, which is intuitive. Meanwhile, the absolute
284 value of eGFR in 2016 was negatively correlated with the outcome in the range above 60
285 ml/min/1.73 m², meaning that participants with higher eGFR were more likely (i.e.,
286 participants with lower eGFR were less likely) to show faster kidney function decline. This
287 phenomenon was also suggested in a previous Japanese study,²⁷ wherein the authors
288 speculated that a compensatory mechanism might work as kidney function decreases. For
289 features other than eGFR-related features, negative associations with Hb1Ac, serum uric acid,
290 and DBP, as well as positive associations with hematocrit and HDL cholesterol, have been
291 suggested in some previous studies.²⁷⁻³² Meanwhile, a U-shaped association with BMI,
292 suggesting that those with a normal BMI range were most likely to show a faster decline,
293 may conflict with previous overseas studies suggesting that obesity is a risk factor for
294 ESRD.^{33,34} Further studies are warranted to examine whether this finding is specific to the
295 prediction of the rate of kidney function decline in the general population or whether this is
296 ascribed to the difficulty in estimating accurate eGFR from the existing formula²⁴ in obese
297 patients.

298 The strength of this study is that we used consecutive 10-year annual health checkup
299 data for the general adult population. A systematic review of kidney disease progression⁷
300 indicated that following the same population for a long time is practically difficult, and only
301 one study was reported to achieve the mean follow-up of over 10 years.³⁵ The mandatory
302 health checkup system for people aged 40–74 years in Japan made the present study feasible.

303 However, this study has some limitations. The database consists of large- and
304 medium-sized company employees and their family members; therefore, their socioeconomic
305 status is expected to be higher than the average in Japan. Accordingly, their health-related
306 behaviors (e.g., smoking and drinking) may be better than those of other Japanese citizens,
307 whereas they may be exposed to stress specific to company employees (e.g., sedentary
308 lifestyles). Therefore, it is unknown whether and to what extent the findings of the present
309 study can be generalized to other citizens in Japan as well as to those living in foreign
310 countries. Second, loss to follow-up could cause selection bias, especially if unhealthy people
311 are more likely to be lost to follow-up.⁷ In the present study, older people and women were
312 more likely to quit the JMDC database. We speculate that the main reasons for the loss to
313 follow-up were social (e.g., retirement at age 60–65, retirement due to pregnancy and
314 childbirth) and not directly associated with the health status of study participants, but we
315 could not confirm the exact reasons. Thus, the effect of the loss to follow-up on our study
316 results is unknown, although the follow-up rate in the present study is better than those in

317 previous studies with long follow-up periods.²⁷⁻³² Third, we obtained the annual health
318 checkup results from health insurance associations instead of the laboratories to measure
319 blood samples, including serum creatinine. Although we believe that creatinine was measured
320 using an internationally standardized enzymatic method (traceable to isotope dilution mass
321 spectrometry) during the study period, creatinine measurements might not be perfectly
322 standardized across laboratories in Japan. However, the influence of this issue in estimating
323 the rate of kidney function decline seems to be small because blood samples from the same
324 individual are expected to be sent to the same laboratories every year. Finally, as discussed
325 above, there may be measurement errors in some predictor variables, especially in
326 self-reported variables such as smoking, drinking, and exercise habits. Furthermore, there
327 may be unknown, and therefore, unmeasured factors predicting the rate of kidney function
328 decline. Further studies are warranted to identify novel risk factors for the rapid decline in
329 kidney function and reassess the performance of the prediction model.

330 In conclusion, using a large database of company employees and their family
331 members in Japan, we estimated the rate of kidney function decline over 10 years and
332 developed a machine learning prediction model based on annual health checkup results and
333 claims for the first five years. The model showed a good or moderate predictive ability,
334 whereas the linear regression model using only eGFR data did not.

335

336 **Disclosures**

337 O.K. and S.T. are employees of JMDC Inc. M.I. previously received honoraria from JMDC
338 Inc. for conference presentations and academic consultations, but does not receive any fee for
339 the present study.

340

341 **Acknowledgments**

342 We thank Editage (www.editage.com) for English language editing.

343

344 **Author contributions**

345 M.I., O.K., and S.T. planned the study. O.K. and S.T. collected and analyzed the data. M.I.
346 wrote the manuscript. O.K. and S.T. prepared the tables, figures, and supplementary materials.
347 All the authors have reviewed the final version of the manuscript.

348

349 **Data availability**

350 The data used in this study were licensed by JMDC Inc. Proposals and requests for data access
351 should be directed to the corresponding authors via email.

352 **References**

- 353 1. Go AS, Chertow GM, Fan D, et al. Chronic kidney disease and the risks of death,
354 cardiovascular events, and hospitalization. *N Engl J Med*. 2004;351(13):1296–1305.
355 doi:10.1056/NEJMoa041031
- 356 2. Iwagami M, Caplin B, Smeeth L, et al. Chronic kidney disease and cause-specific
357 hospitalisation: a matched cohort study using primary and secondary care patient data.
358 *Br J Gen Pract*. 2018;68(673):e512–e523. doi:10.3399/bjgp18X697973
- 359 3. Sakoi N, Mori Y, Tsugawa Y, et al. Early-stage chronic kidney disease and related
360 health care spending. *JAMA Netw Open*. 2024;7(1):e2351518.
361 doi:10.1001/jamanetworkopen.2023.51518
- 362 4. Khan SS, Kazmi WH, Abichandani R, et al. Health care utilization among patients
363 with chronic kidney disease. *Kidney Int*. 2002;62(1):229–236.
364 doi:10.1046/j.1523-1755.2002.00432.x
- 365 5. Fujii M, Ohno Y, Ikeda A, et al. Current status of the rapid decline in renal function
366 due to diabetes mellitus and its associated factors: analysis using the National
367 Database of Health Checkups in Japan. *Hypertens Res*. 2023;46(5):1075–1089.
368 doi:10.1038/s41440-023-01185-2

- 369 6. Lindeman RD, Tobin J, Shock NW. Longitudinal studies on the rate of decline in
370 renal function with age. *J Am Geriatr Soc*. 1985;33(4):278–285.
371 doi:10.1111/j.1532-5415.1985.tb07117.x
- 372 7. Cleary F, Prieto-Merino D, Nitsch D. A systematic review of statistical methodology
373 used to evaluate progression of chronic kidney disease using electronic healthcare
374 records. *PLOS ONE*. 2022;17(7):e0264167. doi:10.1371/journal.pone.0264167
- 375 8. Baigent C, Herrington WG, Coresh J, et al. Challenges in conducting clinical trials in
376 nephrology: conclusions from a Kidney Disease-Improving Global Outcomes
377 (KDIGO) Controversies Conference. *Kidney Int*. 2017;92(2):297–305.
378 doi:10.1016/j.kint.2017.04.019
- 379 9. Carrero JJ, Fu EL, Vestergaard SV, et al. Defining measures of kidney function in
380 observational studies using routine health care data: methodological and reporting
381 considerations. *Kidney Int*. 2023;103(1):53–69. doi:10.1016/j.kint.2022.09.020
- 382 10. Itano S, Kanda E, Nagasu H, et al. eGFR slope as a surrogate endpoint for clinical
383 study in early stage of chronic kidney disease: from The Japan Chronic Kidney
384 Disease Database. *Clin Exp Nephrol*. 2023;27(10):847–856.
385 doi:10.1007/s10157-023-02376-4
- 386 11. Levey AS, Gansevoort RT, Coresh J, et al. Change in albuminuria and GFR as end
387 points for clinical trials in early stages of CKD: a scientific workshop sponsored by

- 388 the National Kidney Foundation in collaboration with the US Food and Drug
389 Administration and European Medicines Agency. *Am J Kidney Dis.*
390 2020;75(1):84–104. doi:10.1053/j.ajkd.2019.06.009
- 391 12. Grams ME, Sang Y, Ballew SH, et al. Evaluating glomerular filtration rate slope as a
392 surrogate end point for ESKD in clinical trials: an individual participant meta-analysis
393 of observational data. *J Am Soc Nephrol.* 2019;30(9):1746–1755.
394 doi:10.1681/ASN.2019010008
- 395 13. Inker LA, Collier W, Greene T, et al. A meta-analysis of GFR slope as a surrogate
396 endpoint for kidney failure. *Nat Med.* 2023;29(7):1867–1876.
397 doi:10.1038/s41591-023-02418-0
- 398 14. Rifkin DE, Shlipak MG, Katz R, et al. Rapid kidney function decline and mortality
399 risk in older adults. *Arch Intern Med.* 2008;168(20):2212–2218.
400 doi:10.1001/archinte.168.20.2212
- 401 15. Tangri N, Grams ME, Levey AS, et al. Multinational assessment of accuracy of
402 equations for predicting risk of kidney failure: a meta-analysis. *JAMA.*
403 2016;315(2):164–174. doi:10.1001/jama.2015.18202
- 404 16. Tangri N, Stevens LA, Griffith J, et al. A predictive model for progression of chronic
405 kidney disease to kidney failure. *JAMA.* 2011;305(15):1553–1559.
406 doi:10.1001/jama.2011.451

- 407 17. Tsai MK, Gao W, Chien KL, et al. A prediction model with lifestyle factors improves
408 the predictive ability for renal replacement therapy: a cohort of 442 714 Asian adults.
409 *Clin Kidney J.* 2022;15(10):1896–1907. doi:10.1093/ckj/sfac119
- 410 18. Grams ME, Brunskill NJ, Ballew SH, et al. Development and validation of prediction
411 models of adverse kidney outcomes in the population with and without diabetes.
412 *Diabetes Care.* 2022;45(9):2055–2063. doi:10.2337/dc22-0698
- 413 19. Aoki J, Kaya C, Khalid O, et al. CKD progression prediction in a diverse US
414 population: a machine-learning model. *Kidney Med.* 2023;5(9):100692.
415 doi:10.1016/j.xkme.2023.100692
- 416 20. Chan L, Nadkarni GN, Fleming F, et al. Derivation and validation of a machine
417 learning risk score using biomarker and electronic patient data to predict progression
418 of diabetic kidney disease. *Diabetologia.* 2021;64(7):1504–1515.
419 doi:10.1007/s00125-021-05444-0
- 420 21. OECD. *OECD Reviews of Public Health: Japan: A Healthier Tomorrow.* OECD
421 Publishing; 2019.
- 422 22. Ministry of Health, Labour and Welfare. Implementation status of specific health
423 checkups and specific health guidance in 2021.
424 https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/newpage_00043.html.

- 425 23. Nagai K, Tanaka T, Kodaira N, et al. Data resource profile: JMDC claims database
426 sourced from health insurance societies. *J Gen Fam Med*. 2021;22(3):118–127.
427 doi:10.1002/jgf2.422
- 428 24. Matsuo S, Imai E, Horio M, et al. Revised equations for estimated GFR from serum
429 creatinine in Japan. *Am J Kidney Dis*. 2009;53(6):982–992.
430 doi:10.1053/j.ajkd.2008.12.034
- 431 25. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the*
432 *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*
433 *Mining*; 2016:785–794. doi:10.1145/2939672.2939785
- 434 26. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for
435 tree ensembles. doi:10.48550/arXiv.1802.03888
- 436 27. Baba M, Shimbo T, Horio M, et al. Longitudinal study of the decline in renal function
437 in healthy subjects. *PLOS ONE*. 2015;10(6):e0129036.
438 doi:10.1371/journal.pone.0129036
- 439 28. Masrouri S, Alijanzadeh D, Amiri M, et al. Predictors of decline in kidney function in
440 the general population: a decade of follow-up from the Tehran Lipid and glucose
441 Study. *Ann Med*. 2023;55(1):2216020. doi:10.1080/07853890.2023.2216020

- 442 29. Jaques DA, Vollenweider P, Bochud M, et al. Aging and hypertension in kidney
443 function decline: a 10 year population-based study. *Front Cardiovasc Med*.
444 2022;9:1035313. doi:10.3389/fcvm.2022.1035313
- 445 30. Cohen E, Nardi Y, Krause I, et al. A longitudinal assessment of the natural rate of
446 decline in renal function with age. *J Nephrol*. 2014;27(6):635–641.
447 doi:10.1007/s40620-014-0077-9
- 448 31. Tsai CW, Ting IW, Yeh HC, et al. Longitudinal change in estimated GFR among
449 CKD patients: a 10-year follow-up study of an integrated kidney disease care program
450 in Taiwan. *PLOS ONE*. 2017;12(4):e0173843. doi:10.1371/journal.pone.0173843
- 451 32. Imai E, Horio M, Yamagata K, et al. Slower decline of glomerular filtration rate in the
452 Japanese general population: a longitudinal 10-year follow-up study. *Hypertens Res*.
453 2008;31(3):433–441. doi:10.1291/hypres.31.433
- 454 33. Hsu CY, McCulloch CE, Iribarren C, et al. Body mass index and risk for end-stage
455 renal disease. *Ann Intern Med*. 2006;144(1):21–28.
456 doi:10.7326/0003-4819-144-1-200601030-00006
- 457 34. Lew QJ, Jafar TH, Talaei M, et al. Increased body mass index is a risk factor for
458 end-stage renal disease in the Chinese Singapore population. *Kidney Int*.
459 2017;92(4):979–987. doi:10.1016/j.kint.2017.03.019

460 35. Abdelhafiz AH, Tan E, Levett C, et al. Natural history and predictors of faster
461 glomerular filtration rate decline in a referred population of older patients with type 2
462 diabetes mellitus. *Hosp Pract (1995)*. 2012;40(4):49–55.
463 doi:10.3810/hp.2012.10.1003

464

465 **Figure legends**

466 **Figure 1. Distribution of the slope (average change) of kidney function decline over 10**
467 **years among the study participants (n=126 464)**

468 eGFR = estimated glomerular filtration rate.

469

470 **Figure 2. Scatter plots of predicted and actual values (A) in the XGBoost model and (B)**
471 **in the linear regression model using only eGFR data from 2012 to 2016**

472 eGFR = estimated glomerular filtration rate.

473

474 **Figure 3. Receiver operating characteristic curves for the rapid decliners (<-3**
475 **ml/min/1.73 m²/year)**

476 eGFR = estimated glomerular filtration rate.

477

478 **Figure 4. Ranking of predictor variables with higher feature importance in the XGBoost**

479 **model**

480 eGFR = estimated glomerular filtration rate, HbA1c = hemoglobin A1c, γ -GTP = gamma

481 glutamyl transpeptidase, BMI = body mass index, HDL = high density lipoprotein.

482

483 **Figure 5. Shapley Addictive explanation (SHAP) dependence plots of predictor**

484 **variables with higher important features in the XGBoost model**

485 eGFR = estimated glomerular filtration rate, HbA1c = hemoglobin A1c, γ -GTP = gamma

486 glutamyl transpeptidase, BMI = body mass index, HDL = high density lipoprotein.

487

488 **Supplementary materials**

489 **Supplementary Table S1. Anatomical Therapeutic Chemical Classification codes to**

490 **define each drug**

491

492 **Supplementary Table S2. List of predictor variables from 2012 to 2016 and**

493 **distributions by the status of data availability from 2017 to 2021**

494

495 **Supplementary Figure S1. Study flow chart**

496

497 **Supplementary Table S3. The mean (standard deviation) slope of eGFR (ml/min/1.73**

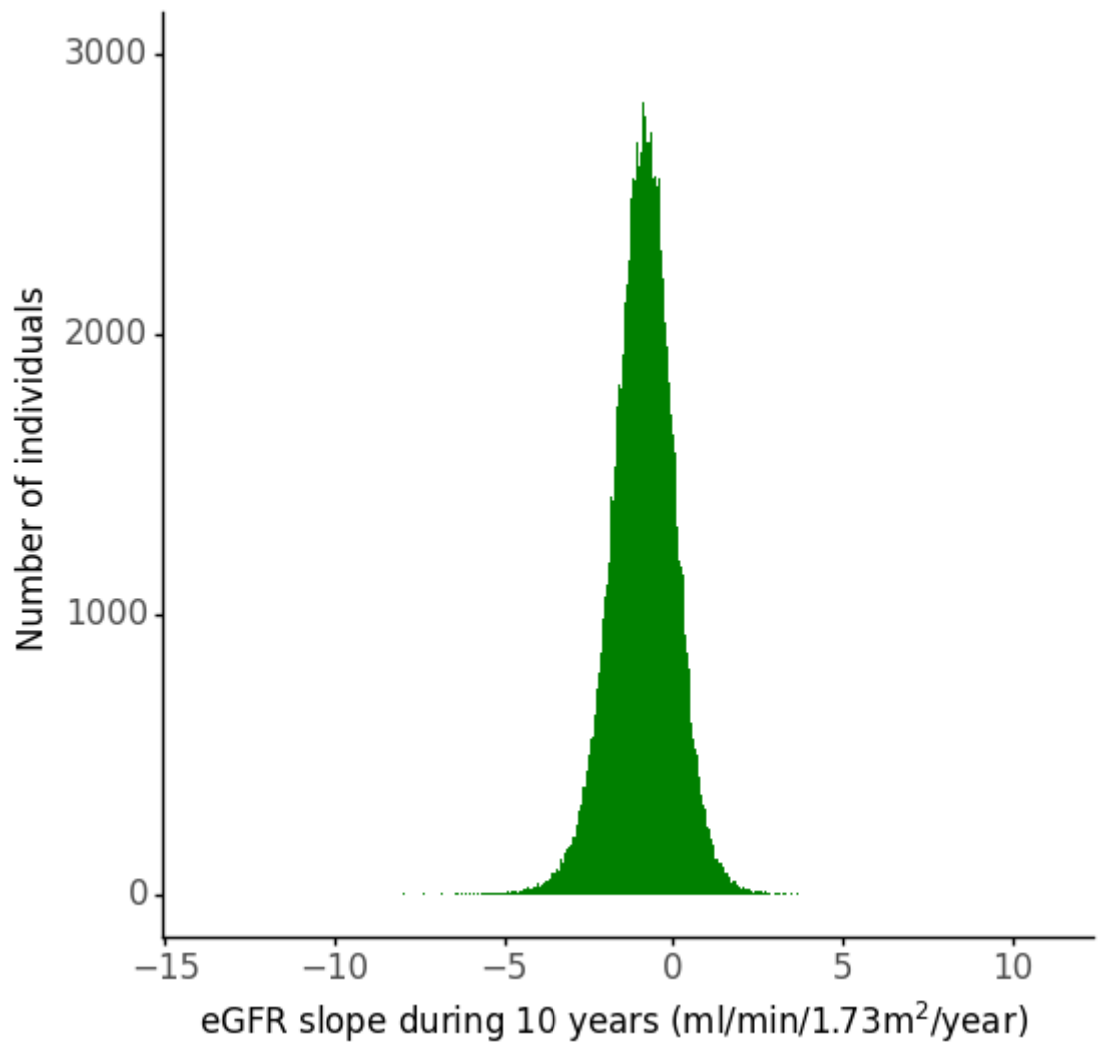
498 **m²/year) for the 10 years by age, sex, and KDIGO GFR stages in 2016**

499 **Table 1. List of predictor variables in 2016 and distributions (n=126 424)**

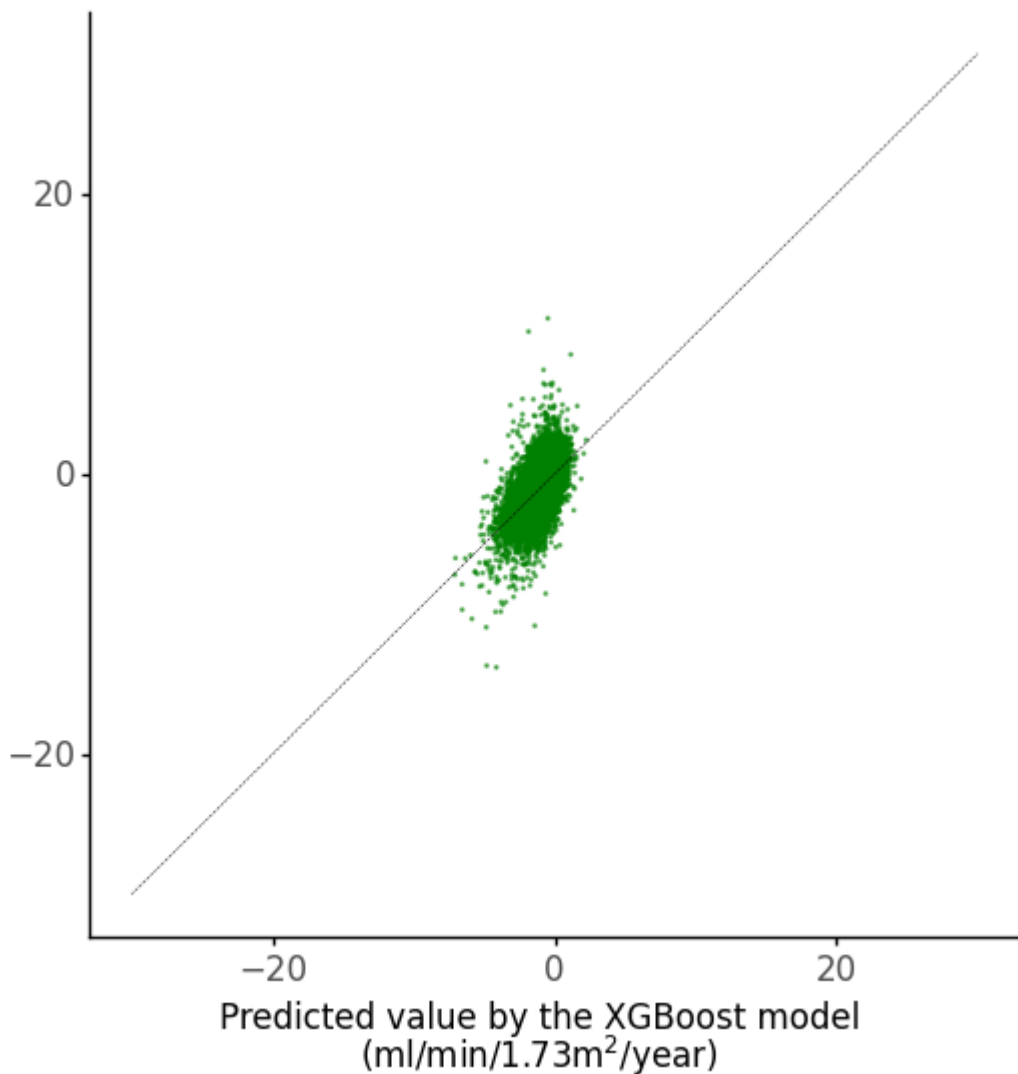
Variables	Distribution	Missing
Age (years)	Mean 45.2, SD 8.6	0 (0%)
Sex	Men: 104 219 (82.4%), Women: 22 205 (17.6%)	0 (0%)
Results of annual health checkups in 2016		
eGFR (ml/min/1.73 m ²)	Mean 79.0, SD 13.4	0 (0%)
Body mass index (kg/m ²)	Mean 23.4, SD 3.6	18 (0.01%)
Abdominal circumference (cm)	Mean 82.6, SD 9.6	10 435 (8.3%)
Systolic blood pressure (mmHg)	Mean 121.6, SD 14.4	20 (0.02%)
Diastolic blood pressure (mmHg)	Mean 75.6, SD 11.1	20 (0.02%)
Triglycerides (mg/dl)	Mean 115.7, SD 91.6	31 (0.02%)
HDL cholesterol (mg/dl)	Mean 60.7, SD 16.1	17 (0.01%)
LDL cholesterol (mg/dl)	Mean 120.7, SD 29.7	29 (0.02%)
Total cholesterol (mg/dl)	Mean 201.8, SD 33.6	76 961 (60.9%)
Aspartate aminotransferase (U/l)	Mean 22.4, SD 10.1	14 (0.01%)
Alanine aminotransferase (U/l)	Mean 25.3, SD 18.8	14 (0.01%)
Gamma glutamyl transpeptidase (U/l)	Mean 39.8, SD 43.4	30 (0.02%)
Fasting blood sugar (mg/dl)	Mean 94.4, SD 16.7	27 532 (21.8%)
Casual blood sugar (mg/dl)	Mean 98.5, SD 25.9	116 912 (92.5%)
Hemoglobin A1c (NGSP) (%)	Mean 5.5, SD 0.6	37 179 (29.4%)
Hematocrit (%)	Mean 45.1, SD 3.8	6 895 (5.5%)
Hemoglobin (g/dl)	Mean 14.8, SD 1.3	790 (0.6%)
Red blood cells (10 ⁶ /μl)	Mean 488.1, SD 41.1	575 (0.5%)
Serum uric acid (mg/dl)	Mean 5.7, SD 1.3	6 429 (5.1%)
Urinary sugar (dipstick test)	-: 120 612 (95.4%), +/-: 628 (0.5%), +: 738 (0.6%), ++: 477 (0.4%), +++: 906 (0.7%)	3 063 (2.4%)
Urinary protein (dipstick test)	-: 112 157 (88.7%), +/-: 7 961 (6.3%), +: 2 437 (1.9%), ++: 656 (0.5%), +++: 166 (0.1%)	3 047 (2.4%)
Smoking status	Yes: 36 760 (29.1%), No: 81 194 (64.2%)	8 470 (6.7%)
Drinking habits	Every day: 31 809 (25.2%), Sometimes: 37 867 (30.0%), Rarely/none: 47 576 (37.6%)	9 172 (7.3%)

Exercise (≥ 2 /week and ≥ 30 minutes in the past year)	Yes: 22 549 (17.8%), No: 86 889 (68.7%)	16 986 (13.4%)
Prescriptions in 2016		
Lipid-lowering agents (any)	Yes: 12 979 (10.3%), No: 113 445 (89.7%)	0 (0%)
Statins	Yes: 10 262 (8.1%), No: 116 162 (91.9%)	0 (0%)
Antidiabetic drugs (any)	Yes: 4 428 (3.5%), No: 121 996 (96.5%)	0 (0%)
Sodium-glucose cotransporter-2 inhibitors	Yes: 1 106 (0.9%), No: 125 318 (99.1%)	0 (0%)
Antihypertensive drugs (any)	Yes: 14 885 (11.8%), No: 111 539 (88.2%)	0 (0%)
ACEI or ARB	Yes: 10 038 (7.9%), No: 116 386 (92.1%)	0 (0%)
Antiplatelet drugs	Yes: 2 948 (2.3%), No: 123 476 (97.7%)	0 (0%)

500 SD, standard deviation; eGFR, estimated glomerular filtration rate; HDL, high density
501 lipoprotein; LDL, low density lipoprotein; NGSP, National Glycohemoglobin
502 Standardization Program; ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin
503 receptor blocker.



(A) XGBoost model



(B) Linear regression model using only eGFR data from 2012 to 2016

