Estimating and predicting kidney function decline in the general population

Running title: Estimating and Predicting Kidney Function Decline

Masao Iwagami, PhD\textsuperscript{1,2,3*}; Kazunori Odani, MSc\textsuperscript{4*}; Tomoki Saito, BSc\textsuperscript{4}

\textsuperscript{1}Department of Health Services Research, Institute of Medicine, University of Tsukuba, Ibaraki, Japan

\textsuperscript{2}Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

\textsuperscript{3}International Institute for Integrative Sleep Medicine (IIIS), University of Tsukuba, Ibaraki, Japan

\textsuperscript{4}JMDC Inc., Tokyo, Japan

*These authors contributed equally

Corresponding authors:

Masao Iwagami, MD, MPH, MSc, PhD

Department of Health Services Research, Institute of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8575, Japan

Tel: +81-29-853-8849

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Fax: +81-29-853-8849
Email: iwagami-tky@umin.ac.jp

Tomoki Saito, BSc

JMDC Inc.

Sumitomo Shiba Daimon Building 12F, 2-5-5 Shiba Daimon, Minato-ku, Tokyo, 108-0072, Japan
Tel: +81-3-5733-5010
Fax: +81-3-5733-5010
Email: tosaito@jmdc.co.jp

Funding: none (self-funded)

Keywords: chronic kidney disease; estimated glomerular filtration rate; creatinine; machine learning; prediction; health checkup
Abstract

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

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

Results: We included 126,424 individuals (mean age, 45.2 years; male, 82.4%; mean eGFR, 79.0 ml/min/1.73 m$^2$ in 2016). The mean slope was -0.89 (standard deviation, 0.96) ml/min/1.73 m$^2$/year. The predictive performance of the XGBoost model (RMSE, 0.78; $R^2$, 0.35; and AUROC, 0.89) was better than that of the linear regression model using only eGFR data (RMSE, 1.94; $R^2$, -3.03; and AUROC, 0.79).
Conclusion: Application of machine learning to annual health-check and claims data could predict the rate of kidney function decline, whereas the linear regression model using only eGFR data did not work.

Keywords: chronic kidney disease; estimated glomerular filtration rate; creatinine; machine learning; prediction; health checkup
Introduction

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

A number of clinical trials and observational studies have set the study endpoints as the time to dialysis initiation or the time to a 30% or 40% drop in estimated GFR (eGFR), mostly among patients at high risk for these events, such as those with late stage CKD. However, the incidence of these outcomes is low in the early stage of CKD or in the general population. Meanwhile, the rate of kidney function decline or slope (average change) of eGFR can be calculated for individuals and could be a surrogate endpoint for clinical trials, even in the early stage of CKD and in the general population. In observational studies, prediction models have been developed for the time to dialysis initiation or the time to a 30% or 40% drop in eGFR, showing good...
discrimination ability and/or calibration. However, to the best of our knowledge, no previous study has developed a prediction model for the rate of kidney function decline as a continuous variable. Such prediction model would be useful for stratifying the general population and identifying those with rapid decline in kidney function. To date, no consensus has been reached on the definition of rapid decliners, and therefore, the prediction of a continuous (rather than a dichotomous) outcome would have a wider application.

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

**Methods**

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

The data used in this study were anonymized and processed anonymously by JMDC, Inc. This study was approved by the Ethics Committee of The Research Institute of Healthcare Data Science (Date of approval, October 30, 2023; Approval number, RI2023003).

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

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

**Outcome definition**

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

\[ eGFR = 194 \times Cr^{-1.094} \times Age^{-0.287} \times 0.739 \text{ (for women)} \]
Predictor variables

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

In addition, using the prescription records in the medical claims, we identified the presence or absence in the use of lipid-lowering agents (any), statins, antidiabetic drugs (any), sodium-glucose transport protein 2 (SGLT2) inhibitors, antiplatelet drugs, antihypertensive drugs (any), and angiotensin converting enzyme inhibitors (ACEI) or angiotensin II receptor
blockers (ARB), which are recorded as the Anatomical Therapeutic Chemical (ATC) Classification codes (Supplementary Table S1).

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

Statistical analysis

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

For the model development, we used the XGBoost regression model\textsuperscript{25}, because it is generally known to show high predictive performance in the case of table data. The implementation was based on the “xgboost” package (version: 1.7.5) of Python. For the hyperparameters, \textit{eta} (step size shrinkage used in the update to prevent overfitting) was set to 0.05, \textit{subsample} (i.e., subsample ratio of the training instances) was set to 0.9, and \textit{colsample_bytree} (i.e., subsampling of columns) was set to 0.8. With a grid search, the
171 \textit{max_depth} (i.e., maximum depth of a tree) and \textit{min_child_weight} (i.e., minimum sum of instance weights [Hessian] needed in a child) were set to 4 and 16, respectively. \textit{n_estimators} was set by early stopping. The other hyperparameters were set to be default values (“XGBoost-link”). We input the annual health checkup data and prescription data for 2016 into the model as they are. For data from 2012 to 2015, we input the subjective (self-reported) findings and prescription data as they are, whereas we calculated and used the difference in values of objective findings between each year and 2016 for each individual. Missing values were input into the XGBoost model as they are.

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

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

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

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

In addition, considering that the identification of rapid decliners (defined in the present study as the slope $<-3$ ml/min/1.73 m$^2$/year$^{14}$) is clinically important, we estimated the area under the receiver operating characteristic curve (AUROC) for the rapid decliners using 5-fold
cross validation. For cross-validation, we merged the validation data of each fold to estimate
the RMSE, $R^2$, and AUROC of the entire training dataset.

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

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

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

Results

In the JMDC database, we identified 183,485 people who underwent annual health checkups
for five consecutive years from 2012 to 2016 (**Supplementary Figure S1**). We then excluded
107 patients who underwent RRT and 28 patients with eGFR $<15$ ml/min/1.73 m$^2$. Of the
remaining 183,350 people, we excluded 56,926 without annual health checkup results
including serum creatinine) for five consecutive years from 2017 to 2021, including 109 patients who received RRT during this period. The remaining 126,424 individuals with consecutive measurements of serum creatinine levels from 2012 to 2021 were included in the subsequent analyses. A comparison of the predictor variables between the two groups (i.e., 126,424 and 56,926 people with and without annual health checkup results from 2017 to 2021) suggested that older people and women were more likely to retire and lose employee insurance to quit from the JMDC database, whereas the annual health checkup results from 2012 to 2016 were not very different between the two groups (Supplementary Table S2). The mean age of the 126,424 people was 45.2 (SD, 8.6) years, 104,219 (82.4%) were male, and the mean eGFR was 79.0 (SD, 13.4) ml/min/1.73 m². According to the KDIGO GFR categories, 23,850 (18.9%), 95,516 (75.6%), 6,781 (5.4%), 247 (0.2%), and 30 (0.02%) accounted for stages G1, G2, G3a, G3b, and G4 (eGFR ≥90, 60–89, 45–59, 30–44, and 15–29 ml/min/1.73 m²), respectively.

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

Figure 2 shows scatter plots of the predicted and actual values in each model, suggesting that the prediction was better in the developed XGBoost model than in the linear
reduction model using only eGFR data from 2012 to 2016. The RMSE and $R^2$ of the
developed XGBoost model were 0.78 and 0.35, respectively, while those of the linear
regression model using only eGFR data were 1.94 and -3.03, respectively. To discriminate
the rapid decliners, the AUROC of the XGBoost model was 0.89, whereas that of the linear
regression model using only eGFR data was 0.79. The ROC curves are shown in Figure 3.

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

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

Among the study population consisting of large- or medium-sized company employees and their family members, most of whom had normal range of kidney function, the mean slope of eGFR for the 10 years was -0.89 (SD, 0.96) ml/min/1.73 m$^2$/year. This was similar to an old US study estimating the mean decrease in creatinine clearance to be -0.75 ml/min/year among normal volunteers. In a recent study conducted in a single health checkup center of Japan, excluding patients with any comorbidities, the mean eGFR decline rate was -1.07 (SD 0.42) ml/min/1.73 m$^2$/year. Those with higher eGFR at baseline were generally
more likely to show faster decline, which is in line with the present study ([Supplementary Table S3](#)).

To the best of our knowledge, this is the first study to predict the rate of kidney function decline as a continuous variable. First, we found it difficult to predict the rate of kidney function decline over 10 years from only eGFR values for the first five years, with an $R^2$ of -3.03, which is even worse than a prediction assuming the average value for all individuals (i.e., $R^2$ of 0). In other words, the correlation between the slope for the 10 years and that for the first five years was weak. This may be partly due to fluctuations in measured serum creatinine over time, although mandatory annual health checkups in Japanese companies are usually conducted at the same place every year and when participants are not sick. Another reason could be that the rate of kidney function decline is also affected by patient demographics; known risk factors for CKD or ESRD, such as diabetes and hypertension; lifestyle factors, such as smoking and exercise; and drugs protecting the kidneys, such as ACEI/ARB and SGLT2 inhibitors. Therefore, we additionally used these variables as predictors and observed a large increase in $R^2$ to 0.35. However, the predictive ability was far from perfect ($R^2$ of 1.00). The remaining possibility is that there may be measurement errors in predictor variables or unknown factors predicting the rate of kidney function decline. Meanwhile, the discrimination ability of the XGBoost model for the rapid
decliners (eGFR slope < -3 ml/min/1.73 m²/year) was seemingly very good, with an AUROC of 0.89.

The feature importance and SHAP dependence plots in the established XGBoost model were remarkable. First, the coefficient (slope) of eGFR for the first five years correlated with the outcome in the same direction, which is intuitive. Meanwhile, the absolute value of eGFR in 2016 was negatively correlated with the outcome in the range above 60 ml/min/1.73 m², meaning that participants with higher eGFR were more likely (i.e., participants with lower eGFR were less likely) to show faster kidney function decline. This phenomenon was also suggested in a previous Japanese study, wherein the authors speculated that a compensatory mechanism might work as kidney function decreases. For features other than eGFR-related features, negative associations with Hb1Ac, serum uric acid, and dBP, as well as positive associations with hematocrit and HDL cholesterol, have been suggested in some previous studies. Meanwhile, a U-shaped association with BMI, suggesting that those with a normal BMI range were most likely to show a faster decline, may conflict with previous overseas studies suggesting that obesity is a risk factor for ESRD. Further studies are warranted to examine whether this finding is specific to the prediction of the rate of kidney function decline in the general population or whether this is ascribed to the difficulty in estimating accurate eGFR from the existing formula in obese patients.
The strength of this study is that we used consecutive 10-year annual health checkup data for the general adult population. A systematic review of kidney disease progression indicated that following the same population for a long time is practically difficult, and only one study was reported to achieve the mean follow-up of over 10 years. The mandatory health checkup system for people aged 40–74 years in Japan made the present study feasible.

However, this study has some limitations. The database consists of large- and medium-sized company employees and their family members; therefore, their socioeconomic status is expected to be higher than the average in Japan. Accordingly, their health-related behaviors (e.g., smoking and drinking) may be better than those of other Japanese citizens, whereas they may be exposed to stress specific to company employees (e.g., sedentary lifestyles). Therefore, it is unknown whether and to what extent the findings of the present study can be generalized to other citizens in Japan as well as to those living in foreign countries. Second, loss to follow-up could cause selection bias, especially if unhealthy people are more likely to be lost to follow-up. In the present study, older people and women were more likely to quit the JMDC database. We speculate that the main reasons for the loss to follow-up were social (e.g., retirement at age 60–65, retirement due to pregnancy and childbirth) and not directly associated with the health status of study participants, but we could not confirm the exact reasons. Thus, the effect of the loss to follow-up on our study results is unknown, although the follow-up rate in the present study is better than those in
previous studies with long follow-up periods. Third, we obtained the annual health checkup results from health insurance associations instead of the laboratories to measure blood samples, including serum creatinine. Although we believe that creatinine was measured using an internationally standardized enzymatic method (traceable to isotope dilution mass spectrometry) during the study period, creatinine measurements might not be perfectly standardized across laboratories in Japan. However, the influence of this issue in estimating the rate of kidney function decline seems to be small because blood samples from the same individual are expected to be sent to the same laboratories every year. Finally, as discussed above, there may be measurement errors in some predictor variables, especially in self-reported variables such as smoking, drinking, and exercise habits. Furthermore, there may be unknown, and therefore, unmeasured factors predicting the rate of kidney function decline. Further studies are warranted to identify novel risk factors for the rapid decline in kidney function and reassess the performance of the prediction model.

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

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

Acknowledgments

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

Author contributions

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

Data availability

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


11. Levey AS, Gansevoort RT, Coresh J, et al. Change in albuminuria and GFR as end points for clinical trials in early stages of CKD: a scientific workshop sponsored by
the National Kidney Foundation in collaboration with the US Food and Drug Administration and European Medicines Agency. *Am J Kidney Dis.*

2020;75(1):84–104. doi:10.1053/j.ajkd.2019.06.009


doi:10.1007/s40620-014-0077-9


doi:10.7326/0003-4819-144-1-200601030-00006


doi:10.3810/hp.2012.10.1003

Figure legends

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

eGFR = estimated glomerular filtration rate.

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

eGFR = estimated glomerular filtration rate.

**Figure 3.** Receiver operating characteristic curves for the rapid decliners (<-3 ml/min/1.73 m²/year)
eGFR = estimated glomerular filtration rate.
Figure 4. Ranking of predictor variables with higher feature importance in the XGBoost model

eGFR = estimated glomerular filtration rate, HbA1c = hemoglobin A1c, γ-GTP = gamma glutamyl transpeptidase, BMI = body mass index, HDL = high density lipoprotein.

Figure 5. Shapley Addictive explanation (SHAP) dependence plots of predictor variables with higher important features in the XGBoost model

eGFR = estimated glomerular filtration rate, HbA1c = hemoglobin A1c, γ-GTP = gamma glutamyl transpeptidase, BMI = body mass index, HDL = high density lipoprotein.

Supplementary materials

Supplementary Table S1. Anatomical Therapeutic Chemical Classification codes to define each drug

Supplementary Table S2. List of predictor variables from 2012 to 2016 and distributions by the status of data availability from 2017 to 2021

Supplementary Figure S1. Study flow chart
Supplementary Table S3. The mean (standard deviation) slope of eGFR (ml/min/1.73 m²/year) for the 10 years by age, sex, and KDIGO GFR stages in 2016
<table>
<thead>
<tr>
<th>Variables</th>
<th>Distribution</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Mean 45.2, SD 8.6</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Sex</td>
<td>Men: 104 219 (82.4%), Women: 2205 (17.6%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

**Results of annual health checkups in 2016**

- **eGFR (ml/min/1.73 m²)**: Mean 79.0, SD 13.4, 0 (0%)
- **Body mass index (kg/m²)**: 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 (10⁶/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%)
- **Urinary protein (dipstick test)**: -: 112 157 (88.7%), +/-: 7 961 (6.3%), +: 2 437 (1.9%), ++: 656 (0.5%), +++: 166 (0.1%)
- **Smoking status**: Yes: 36 760 (29.1%), No: 81 194 (64.2%)
- **Drinking habits**: Every day: 31 809 (25.2%), Sometimes: 37 867 (30.0%), Rarely/none: 47 576 (37.6%)
<table>
<thead>
<tr>
<th>Exercise (≥2/week and ≥30 minutes in the past year)</th>
<th>Yes: 22 549 (17.8%), No: 86 889 (68.7%)</th>
<th>16 986 (13.4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prescriptions in 2016</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid-lowering agents (any)</td>
<td>Yes: 12 979 (10.3%), No: 113 445 (89.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Statins</td>
<td>Yes: 10 262 (8.1%), No: 116 162 (91.9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Antidiabetic drugs (any)</td>
<td>Yes: 4 428 (3.5%), No: 121 996 (96.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Sodium-glucose cotransporter-2 inhibitors</td>
<td>Yes: 1 106 (0.9%), No: 125 318 (99.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Antihypertensive drugs (any)</td>
<td>Yes: 14 885 (11.8%), No: 111 539 (88.2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ACEI or ARB</td>
<td>Yes: 10 038 (7.9%), No: 116 386 (92.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Antiplatelet drugs</td>
<td>Yes: 2 948 (2.3%), No: 123 476 (97.7%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

SD, standard deviation; eGFR, estimated glomerular filtration rate; HDL, high density lipoprotein; LDL, low density lipoprotein; NGSP, National Glycohemoglobin Standardization Program; ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker.
(A) XGBoost model

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