Cohort-derived machine learning models for individual prediction of chronic kidney disease in people living with HIV: a prospective multicentre cohort study

Jan A. Roth,[1,2*] Gorjan Radevski,[3*] Catia Marzolini,[1] Andri Rauch,[4] Huldrych F. Günthard,[5,6] Roger D. Kouyos,[5,6] Christoph A. Fux,[7] Alexandra U. Scherrer,[5,6] Alexandra Calmy,[8] Matthias Cavassini,[9] Christian R. Kahlert,[10,11] Enos Bernasconi,[12] Jasmina Bogojeska,[3**] Manuel Battegay,[1**] and the *Swiss HIV Cohort Study (SHCS)*

*Shared first authors.

**Shared last authors.

[1]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, 4031 Basel, Switzerland.

[2]Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, 4031 Basel, Switzerland.

[3]IBM Research — Zurich, 8803 Rüschlikon, Switzerland.

[4]University Clinic of Infectious Diseases, University Hospital Berne, University of Berne, 3010 Berne, Switzerland.

[5]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, 8091 Zurich, Switzerland.

[6]Institute of Medical Virology, University of Zurich, 8057 Zurich, Switzerland.

[7]Clinic for Infectious Diseases and Hospital Hygiene, Kantonsspital Aarau, 5001 Aarau, Switzerland.

[8]Division of Infectious Diseases, University Hospital Geneva, University of Geneva, 1205 Geneva, Switzerland.

[9]Division of Infectious Diseases, University Hospital Lausanne, University of Lausanne, 1011 Lausanne, Switzerland.

[10]Division of Infectious Diseases and Hospital Epidemiology, Cantonal Hospital St. Gallen, 9007 St. Gallen, Switzerland.

[11]Division of Infectious Diseases and Hospital Epidemiology, Children's Hospital of Eastern Switzerland, 9006 St. Gallen, Switzerland.

[12]Division of Infectious Diseases, Regional Hospital Lugano, 6900 Lugano, Switzerland.

Corresponding author (final publication): Prof. Dr Manuel Battegay, MD, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland. Tel: +41 61 328 60 72, Fax: +41 61 265 31 98. E-mail address: manuel.battegay@usb.ch

Alternate Corresponding Author: Jan A. Roth, MD, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland. Tel: +41 (0)61 556 55 85. E-mail address: janadam.roth@usb.ch

SHORT SUMMARY

In people living with HIV who participate in the Swiss HIV Cohort Study, we observed state-of-the-art performances in forecasting individual onsets of chronic kidney disease with different machine learning algorithms.

ABSTRACT

Background: It is unclear whether data-driven machine learning models, which are trained on large epidemiological cohorts, may improve prediction of co-morbidities in people living with HIV.

Methods: In this proof-of-concept study, we included people living with HIV of the prospective Swiss HIV Cohort Study with a first estimated glomerular filtration rate (eGFR) >60 ml/min/1.73 m^2 after January 1, 2002. Our primary outcome was chronic kidney disease (CKD) ─ defined as confirmed decrease in eGFR ≤60 ml/min/1.73 m^2 over three months apart. We split the cohort data into a training set (80%), validation set (10%), and test set (10%) ─ stratified for CKD status and follow-up length.

Results: Of 12,761 eligible individuals (median baseline eGFR, 103 ml/min/1.73 m^2), 1,192 (9%) developed a CKD after a median of eight years. We used 64 static and 502 time-changing variables: Across prediction horizons and algorithms and in contrast to expert-based standard models, most machine learning models achieved state-of-the-art predictive performances with areas under the receiver operating characteristic curve and precision recall curve ranging from 0.926 to 0.996 and from 0.631 to 0.956, respectively.

Conclusions: In people living with HIV, we observed state-of-the-art performances in forecasting individual CKD onsets with different machine learning algorithms.

Keywords: chronic kidney disease; digital epidemiology; HIV; machine learning; prediction.

INTRODUCTION

With the advent of combined antiretroviral therapy, HIV-related morbidity and mortality have continuously decreased ─ with people living with HIV having nowadays, under optimal conditions, an almost identical life expectancy to the general population [1-4]. As HIV infection has become a chronic condition, accurate prediction of primarily non-HIV-related co-morbidities such as chronic kidney disease (CKD) have gained importance in the individualised care of people living with HIV [5].

As the occurrence of CKD and of other non-HIV-related chronic conditions may be influenced by hundreds of potentially interacting, static and time-changing factors across the healthcare continuum, data-rich and well-curated HIV cohorts may offer ideal conditions to develop machine learning models and to validate their usefulness to optimise personalised prevention and treatment strategies in people living with HIV. Cohort-based machine learning is an evolving field in digital epidemiology, which has the potential to improve decision support and underlying prediction models [6, 7]. Previous prediction models of CKD and of other multifactorial conditions may be limited, as it is challenging to account for complex interactions and to analyse high-dimensional datasets (i.e. data collections with a multitude of variables) with standard regression models. Conversely, some machine learning prediction models have limited generalisability to other settings with intransparent predictions for single individuals [8].

In the present proof-of-concept study conducted in people living with HIV, we aimed to evaluate different machine learning algorithms and modeling strategies for individual CKD prediction in order to exemplify whether machine learning models can be readily trained in a high-dimensional cohort setting. The resulting machine learning prediction models of CKD onsets may become part of an integrated decision support tool for shared decision-making and personalisation of prevention and treatment strategies in people living with HIV. In a wider context, our investigation may be helpful for current large-scale cohorts to assess the feasability and challenges with cohort-based machine learning prediction.

METHODS

Swiss HIV Cohort Study

The Swiss HIV Cohort Study (SHCS; www.shcs.ch) is a nationwide, prospective multicentre cohort study with semi-annual visits and blood collections ─ having enrolled >20,000 HIV-infected adults who live in Switzerland [9]. The SHCS is representative of the HIV epidemic in Switzerland [9]. A standardised protocol is used in the SHCS for data collection: Socio-demographic and clinical data are recorded at study entry and various laboratory tests are routinely performed at registration. At each follow-up visit, extensive laboratory, clinical and treatment information is recorded. Additional interim laboratory and clinical evaluations are recorded, if available. The SHCS is registered on the longitudinal study platform of the Swiss National Science Foundation (www.snf.ch/en/funding/programmes/longitudinal-studies).

For the training of pragmatic and individualised machine learning models, most SHCS variables have been used, but potentially identifying variables (including living/working situations), information on sexual behaviour, variables recorded only within a short period, genetic/-omics data, and some metadata (e.g. name of study nurse) were omitted as defined *a priori* in the study group and as discussed with a national representative of people living with HIV. Where applicable, we followed ‘The Strengthening the Reporting of Observational

Studies in Epidemiology' and the 'Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis' statement when reporting our study results [10, 11]; furthermore, we used the reporting criteria developed by Luo *et al.* (2016) [12].

Study population and definitions

After January 1, 2002, when calibrated creatinine measurements were incorporated in the SHCS, we included HIV-infected individuals aged ≥18 years with a baseline estimated glomerular filtration rate (eGFR) >60 ml/min/1.73 m^2 ─ independent of antiretroviral treatment regimens/status ─ and at least three calibrated serum creatinine measurements before October 10, 2018. Individuals with a baseline eGFR ≤60 ml/min/1.73 m^2, less than three creatinine measurements, and/or less than three months of follow-up were excluded.

We defined the baseline as the first creatinine measurement after January 1, 2002. We followed individuals from baseline until occurrence of CKD or the last recorded creatinine measurement, whichever came first. However, we used horizons of three to twelve months for machine learning prediction of CKD onset.

We defined CKD, our *a priori* primary outcome, as a *confirmed* (over three months apart) decrease in eGFR ≤60 ml/min/1.73 m^2, in line with the 'Kidney Diseases ─ Improving Global Outcomes' algorithm and previous large-scale investigations on CKD in people living with HIV [5, 13]. As a measure of kidney function, we calculated the eGFR using the well-established 'Chronic Kidney Disease Epidemiology Collaboration' equation, which had been validated extensively in people living with HIV [14-17].

All participants in the SHCS provided informed consent and the study was approved by the ethical committees of the respective participating centers (EKNZ project No. 2017─02252). We report deviations from the study protocol in the appendix (page 3).

Predictive modeling

We trained a set of data-driven machine learning models (full models) to predict CKD events within prespecified prediction horizons ─ representing a classification problem, which relied on both static and irregularly sampled time and event series data. We applied the following five machine learning algorithms for CKD prediction with single patient visits as unit of observation and parameter tuning (selection) on the validation set:

(i) *Elastic net* is a regularised, linear logistic regression method that includes both the lasso (L_1) and the ridge (L_2) penalty via a linear combination [18]. It optimises the following objective:

$$\max_{\beta,\lambda,\nu} \log \sum_{i=1}^{N} \log p(y_i|x_i, \beta_i) + \lambda||\boldsymbol{\beta}||^2 + \nu||\boldsymbol{\beta}||_1$$

where $\{(x_1, y_1), (x_3, y_2), \ldots, (x_N, y_N)\}$ is the training dataset, and β, λ and ν are the model parameters.

(ii) *Random forest models* [19] average a collection of decorrelated classification or regression trees, in which a prespecified number of trees are fitted ─ each on a separate bootstrap sample drawn with replacement from the training data. We describe the details of the algorithm in appendix table 1.

(iii) *Gradient boosting machine* [20] is an ensemble approach that iteratively adds simple models to the ensemble such that in each iteration a new model is trained with respect to the updated error of the ensemble learned in the previous iteration. We describe the details of the respective training algorithm in appendix table 2.

(iv) *Multilayer perceptron* [21] is a non-linear machine learning approach — representing a feedforward neural network with at least three fully connected layers. We used the rectified linear unit:

$f(x) = \max(0, x)$ as activation function.

(v) *Recurrent neural networks* (RNNs) are artificial neural networks that use a directed graph to model the connections between the nodes and are thus directly applicable to temporal sequence data. We used the 'Long Short Term Memory' (LSTM) architecture [22]. We describe the details of the respective training algorithm in appendix table 3.

For comparison with data-driven machine learning models, we have manually built logistic regression models (short models) for the different prediction horizons – in analogy to the well-established full risk score model by Mocroft et al. for prediction of CKD in people living with HIV.[13] We used the following predictors: HIV exposure through intravenous drug use (yes, no, or unknown), hepatitis C coinfection (yes or no), birth year, estimated glomerular filtration rate until day of prediction (normalized scale; modelled as described for the data-driven machine learning models), sex (male or female), CD4 count until day of prediction (normalized scale; modelled as described for the data-driven machine learning models), hypertension (yes, no, or unknown), prior cardiovascular diseases (yes or no), and diabetes mellitus (yes or no). Our manually built logistic regression models use the last two most recent measurements of the considered variables along with the summary statistics of all their previous measurements.

Dataset representation

To train our machine learning models, we extracted the anonymised study data from the SHCS main database ─ comprising a vast collection of static and time-changing (dynamic) variables, which were often irregularly measured as part of the clinical routine. The RNN-based methods process sequences of inputs and can thus use the visit sequence directly. For the remaining machine learning methods, the input information for each individual is a concatenation of the information from the two last (most recent) hospital visits and the corresponding summary statistics (mean, median, max, standard deviation) from all previous visits. Note that the visit sequence for each patient is derived from the considered observation period determined by the target prediction horizon and the last (most recent) visits refer to these derived sequences. We describe the detailed data representation and missing value imputation strategy in the appendix (page 4).

Model evaluation

To evaluate the performance of the different machine learning approaches and models, we split all study data into three subsets, namely a *training set*, a *validation set* and a *test set*. We created the validation and test sets by randomly sampling (without replacement) 10% of the study population. The sampling was stratified with respect to the follow-up length and CKD status, i.e. 10% of individuals were at first randomly sampled from the group of individuals that have developed CKD and then 10% were randomly sampled from the group of the individuals that did not develop CKD. The remaining 80% of the individuals comprised the training set.

We applied each of the described machine learning methods to predict CKD events as a set of adjusted hyperparameters to deliver accurate predictions on unseen data. We performed the model selection/hyperparameter tuning process on the validation set. Finally, we evaluated the predictive performance of the best-performing model for each considered approach on the test set (reported in the results section). We considered four different evaluation scenarios, each with a different prediction horizon, namely 90, 180, 270, and 365 days. The prediction horizon specifies how many days in advance we aimed to predict the occurrence of CKD where the time of diagnosis is determined by the second eGFR measurement of the CKD definition used.

Performance measures

Due to the large CKD imbalance in our dataset (i.e. most individuals did not develop CKD), the classification accuracy was not suitable to measure the models' performance. Therefore, we calculated five well-established measures for the class imbalance scenario; namely, the F-score, precision (i.e. positive predictive value), recall (i.e. sensitivity), area under the receiver operating characteristic curve (ROC-AUC), and area under the precision recall curve (PR-AUC). The precision, recall and F-score are defined as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

where TP denotes the true positives, FP denotes the false positives, FN denotes the false negatives and positives refer to the minority class (in our case individuals with CKD onset).

The precision recall curve is a plot of the recall versus the precision for all possible decision thresholds. As the precision and recall focus only on the correct prediction of the minority class (i.e. CKD), the F-Score and the PR-AUC reflect the model's prediction quality for CKD events. The receiver operating characteristic curve is a widely used plot of the false positive rate (the proportion of false positives out of all negatives) versus the true positive rate (the proportion of true positives out of all positives) for all possible decision thresholds. The ROC-AUC thus illustrates the ranking ability in binary classification: A ROC-AUC of, for instance, 0.80 indicates that 80% of the predictions are correctly classified (for pairs of individuals with and without the endpoint). For model selection, we used the F-score for the RNN-based approaches and the log loss for the remaining approaches.

Due to the time-consuming model selection process, we performed all experiments and computed all relevant evaluation metrics for one training, validation and test split. We believe that our results reflect the predictive quality of the considered machine learning models, as our test set was fairly large.

RESULTS

Within the study period, 12,761 individuals were included in the final analysis ─ with 10,209 (80%), 1,276 (10%), and 1,276 (10%) of participants' prospectively collected cohort records contributing to the machine learning model training, validation, and test sets, respectively (figure 1). We describe the main characteristics of the study population in table 1: Overall, 1,192 of 12,761 (9%) individuals developed a CKD within the study period; the median follow-up in individuals with and without CKD was 8 years (interquartile range [IQR], 4 to 12 years) and 9 years (IQR, 4 to 15 years), respectively.

We describe the eGFR distribution of individuals with and without CKD in figure 2: At baseline, eGFR distributions were partly overlapping between individuals with and without a subsequent CKD ─ with increased eGFRs of individuals without subsequent CKD onset across prediction horizons. For individuals with and without subsequent CKD, the overlap in eGFR distributions increased over longer prediction horizons. Overall, at day of prediction, the frequency of subsequent eGFR measurements within 365 days was slightly increased for individuals with a decreased eGFR of ≤60 ml/min/1.73 m^2 compared to individuals with eGFRs >60 ml/min/1.73 m^2 (median, 1.8 measurements per month; IQR, 1.0 to 2.5; versus; median, 1.5 measurements per month; IQR, 0.7 to 2.3).

We used 64 static and 502 dynamic variables for machine learning model development (full models) ─ including 28 demographic variables, 159 variables pertaining to treatment information, 93 laboratory variables, and 286 clinical variables: Across prediction horizons and machine learning algorithms, most models achieved similar predictive performances with ROC-AUCs and PR-AUCs ranging from 0.926 to 0.996 (i.e. 92.6% to 99.6% of predictions are correctly classified for pairs with and without CKD) and from 0.631 to 0.956, respectively (table 2). In regard to ROC-AUCs and PR-AUCs, the machine learning models' classification performance can be considered as excellent and moderate to excellent, respectively: The PR-AUCs were lower than the corresponding ROC-AUCs, as CKD events were relatively rare. For comparison with the full machine learning models, we have manually built logistic regression models (short models) based on well-established predictors (table 2): In most cases, these short models had a worse predictive performance than the full machine learning models for CKD prediction.

For illustration purposes, we describe in figure 3 the variable importance of the highest scoring predictors for the gradient boosting model (prediction horizon, 180 days): Overall, the eGFR information was the most important marker for CKD prediction within 180 days. Across prediction horizons, we describe the gradient boosting models' output and individual key predictors for three complex cases (table 3): Information on predicted outcome probabilities and the individual variable importance can be obtained for all applied machine learning algorithms to increase the interpretability/transparency of machine learning models and to potentially personalise prevention and treatment decisions.

The preparation and structuring of our datasets for machine learning training required one-month full-time work. The RNN-based model selection procedure was computing-intensive and required 20 to 30 hours on a high-performance computing cluster. The corresponding computing time for model selection among the remaining non-linear approaches was in the order of one to two hours each. The final model training was fast for all machine learning methods except for the RNN-based methods, which required approximately 30 minutes. Obtaining individual predictions with a trained model was fast (couple of minutes at most) for all machine learning methods.

DISCUSSION

In this large cohort study, we have developed pragmatic machine learning models to predict CKD onset and derive CKD development probabilities at the point of care in single individuals living with HIV. The respective machine learning models had a rather high predictive performance despite using prediction horizons of three to twelve months, which may decrease the precision (i.e. positive predictive value) for CKD predictions. We measured our machine learning models' predictive power by a set of well-established metrics to improve the comparability across models and studies. In contrast to previous studies, we have included a multitude of static and dynamic factors in our prediction models (data-driven machine learning modeling), which resulted mostly in improved performances for CKD prediction compared to manually built regression models based on a few predictor variables (table 2) [13, 23]. Our proof-of-concept study provides a “reality-check” of the feasability of machine learning prediction studies nested within large epidemiological cohorts.

To the best of our knowledge, this is the first study, in which different machine learning models have been developed and internally validated in people living with HIV for individualised CKD prediction. Previous studies have developed standard regression-based models and scores (e.g. by use of Poisson regression) for long-term CKD prediction, which had a good discrimination in external validation [5, 13, 23, 24]. For instance, as part of the ‘Data Collection on Adverse Events of Anti-HIV Drugs’ study, a full and short risk score were developed to predict CKD over 5 years (but not for shorter prediction horizons) ─ with the short risk score demonstrating a relatively good predictive performance in external validation (ROC-AUC, 0.85) [13, 24]: These widely used full and short risk scores were developed in individuals living with HIV who were not previously exposed to a potentially nephrotoxic antiretroviral agent and included nine and six predictor variables, respectively. In contrast to these two CKD risk scores, we used a set of machine learning algorithms and short-term prediction horizons ─ accounting for individuals with any antiretroviral treatment status and incorporating a variety of static and time-changing variables. These various short-term prediction horizons may be useful to differentiate acute and chronic kidney disease and to evaluate the dynamics and plausibility of machine learning predictions in single individuals over time. For individual CKD predictions, we achieved moderate to excellent discrimination with the given machine learning models. Therefore, our models can be investigated as part of a subsequent implementation study to assess the clinical utility and validity of the present machine learning models ─ also for complex cases (table 3).

Of interest, as illustrated in the variable importance plot of the gradient boosting model (figure 3), we observed a number of predictors, which are well-established risk factors for CKD (e.g. treatment with tenofovir disoproxil fumarate containing regimens [25]) as well as proxy variables and markers, which may not have a direct effect on CKD development (e.g. alkaline phosphatase). This observation highlights that predictive machine learning models may help to build novel causal hypotheses, which can be validated in subsequent causal studies. However, machine learning predictions and corresponding variable importance plots should not be used *per se* for causal inference, as it requires expert guidance and causal concepts.

While developing machine learning models for CKD prediction, we faced two main challenges. Firstly, the preparation and structuring of the datasets for machine learning training was time-consuming, as real-world HIV cohort data include a multitude of static and dynamic data, which are often measured irregularly. Nonetheless, we believe that our data representation can be valuable for future machine learning investigations relying on (HIV) cohort databases. Secondly, the machine learning model training and selection was computing-intensive and required a high-performance computing cluster.

Our study has some limitations. Firstly, our machine learning predictions models for CKD may not be generalisable to other healthcare settings and populations; specifically, the coding practices and parameters may differ between HIV cohorts, which may complicate the application of the same machine learning prediction models across HIV cohorts. Therefore, we did not intend to externally validate our machine learning prediction models as part of this proof-of-concept study. Secondly, as we used short prediction horizons, target leakage (i.e. models include information that is not yet available at the time of prediction) can result in biased and often too optimistic predictive performances. To safeguard against target leakage, we included only variables that were known at the prediction day [26]. However, we cannot exclude the possibility that a few parameters in our machine learning models (e.g. laboratory values) would be reported to the treating physician and/or clinical decision support tool some minutes or hours after a potential CKD prediction. Thirdly, follow-up studies should consider including proteinuria in the CKD outcome definition to capture CKD at earlier stages. With the present models, we are unable to predict proteinuria. Fourthly, a higher eGFR threshold >60 ml/min/1.73m^2 could have been chosen for patient selection to prevent immediate switches from the at risk status to the CKD status; however, this would have excluded a substantial proportion of individuals in the SHCS, which are at highest risk of eGFR deterioration. Lastly, our machine learning model training did not include genetic data (or other –omics data), which might have further improved the machine learning CKD predictions but which are often unavailable for a majority of individuals [27].

CONCLUSION

In people living with HIV, we observed state-of-the-art performances in forecasting individual CKD onsets with different machine learning algorithms: The underlying machine learning methods may help to advance personalised predictions of co-morbidities in various populations.

ACKNOWLEDGMENTS

Members of the Swiss HIV Cohort Study (SHCS): Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (president of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (chairman of the mother & child substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (chairman of the scientific board), Rudin C, Scherrer AU (head of data center), Schmid P, Speck R, Stöckle M (chairman of the clinical and laboratory committee), Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

Financial Support: This study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation [grant number 177499], by SHCS project #814 and by the SHCS Research Foundation. The data are gathered by the five Swiss university hospitals, two cantonal hospitals, 15 affiliated hospitals, and 36 private physicians (listed in www.shcs.ch/180-health-care-providers).

Contributions: JAR, JB, MB, CM, AR, HFG, RDK, and CAF developed the study protocol and drafted the manuscript. All authors critically reviewed the study protocol. GR and JB analysed the data with input from JAR. All authors critically reviewed the manuscript. All authors contributed to the design of the study and approved the final version of the manuscript.

Potential conflicts of interest: All authors declare no competing interests.

Reprints: Prof. Dr Manuel Battegay, MD, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland. Tel: +41 61 328 60 72, Fax: +41 61 265 31 98. E-mail address: manuel.battegay@usb.ch

REFERENCES

1. Gueler A, Moser A, Calmy A, et al. Life expectancy in HIV-positive persons in Switzerland: matched comparison with general population. AIDS **2017**;31(3):427─36.
2. Marcus JL, Chao CR, Leyden WA, et al. Narrowing the gap in life expectancy between HIV-infected and HIV-uninfected individuals with access to care. J Acquir Immune Defic Syndr **2016**;73(1):39─46.
3. Weber R, Ruppik M, Rickenbach M, et al. Decreasing mortality and changing patterns of causes of death in the Swiss HIV Cohort Study. HIV Med **2013**;14(4):195─207.
4. Wandeler G, Johnson LF, Egger M. Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population. Curr Opin HIV AIDS **2016**;11(5):492─500.
5. Mocroft A, Lundgren JD, Ross M, et al. Cumulative and current exposure to potentially nephrotoxic antiretrovirals and development of chronic kidney disease in HIV-positive individuals with a normal baseline estimated glomerular filtration rate: a prospective international cohort study. Lancet HIV **2016**;3(1):e23─32.
6. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Med **2019**;25(1):44─56.
7. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng **2018**;2(10):719─31.
8. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med **2019**;380(14):1347─58.
9. Swiss HIV Cohort Study, Schoeni-Affolter F, Ledergerber B, et al. Cohort profile: The Swiss HIV Cohort study. Int J Epidemiol **2010**;39(5):1179─89.

10. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol **2008**;61(4):344─9.

11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med **2015**;162(10):735─6.

12. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res **2016**;18(12):e323.

13. Mocroft A, Lundgren JD, Ross M, et al. Development and validation of a risk score for chronic kidney disease in HIV infection using prospective cohort data from the D:A:D study. PLoS Med **2015**;12(3):e1001809.

14. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med **2009**;150(9):604─12.

15. Cristelli MP, Cofan F, Rico N, et al. Estimation of renal function by CKD-EPI versus MDRD in a cohort of HIV-infected patients: a cross-sectional analysis. BMC Nephrology **2017**;18(1): 58.

16. Bonjoch A, Bayes B, Riba J, et al. Validation of estimated renal function measurements compared with the isotopic glomerular filtration rate in an HIV-infected cohort. Antiviral Research **2010**;88(3):347─54.

17. Gagneux-Brunon A, Delanaye P, Maillard N, et al. Performance of creatinine and cystatin C-based glomerular filtration rate estimating equations in a European HIV-positive cohort. AIDS **2013**;27(10):1573─81.

18. Zou H, Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society **2005**;67(2):301─20.

19. Breiman L. Random Forests. Machine Learning **2001**;45(1):5─32.

20. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics **2000**;29:1189─232.

21. Rosenblatt F. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Washington DC: Spartan Books, **1961**.

22. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation **1997**;9(8):1735─80.

23. Scherzer R, Gandhi M, Estrella MM, et al. A chronic kidney disease risk score to determine tenofovir safety in a prospective cohort of HIV-positive male veterans. AIDS **2014**;28(9):1289─95.

24. Woolnough EL, Hoy JF, Cheng AC, et al. Predictors of chronic kidney disease and utility of risk prediction scores in HIV-positive individuals. AIDS **2018**;32(13):1829─35.

25. Aloy B, Tazi I, Bagnis CI, et al. Is tenofovir alafenamide safer than tenofovir disoproxil fumarate for the kidneys? AIDS Rev **2016**;18(4):184─92.

26. Roth JA, Battegay M, Juchler F, Vogt JE, Widmer AF. Introduction to machine learning in digital healthcare epidemiology. Infect Control Hosp Epidemiol **2018**;39(12):1457─62.

27. Dietrich LG, Barcelo C, Thorball CW, et al. Contribution of genetic background and clinical D:A:D risk score to chronic kidney disease in Swiss HIV-positive persons with normal baseline estimated glomerular filtration rate. Clin Infect Dis **2019** [Epub ahead of print].

FIGURE LEGENDS

***Figure 1*: Study population**

Abbreviations: SHCS, Swiss HIV Cohort Study.

[a] Calculated using the 'Chronic Kidney Disease Epidemiology Collaboration' equation.

[b] We defined the baseline as the first creatinine measurement after January 1, 2002.

***Figure 2*: Overall glomerular filtration rates in people living with HIV (N = 12,761 individuals)**

Abbreviations: CKD, chronic kidney disease; GFR, glomerular filtration rate [ml/min/1.73 m^2].

Note: This figure refers to the glomerular filtration rate at the last visit of the visit sequences in the considered observation period that is used to make predictions for 90 days, 180 days, 270 days, and 365 days ahead subsequently. The middle line and box indicate the median and interquartile range, respectively. Whiskers cover the 1.5 interquartile range.

Figure 3**: Variable importance plot of the gradient boosting model; 180 days prediction horizon**

Abbreviations: GFR, glomerular filtration rate; SHAP, shapley additive explanation; std, standard deviation.

Note: This hypothesis-generating plot is for illustration purposes only. Suffix ‘2’ signifies that information from the latest visit was used, whereas suffix ‘1’ signifies that information from the preceding (penultimate) visit was used, both specified with respect to the visit sequence in the considered observation period. The different statistics (the median, standard deviation for numeric and max for the nominal variables) were computed for all the remaining visits in the target observed hospital visit sequence. The SHAP values describes for each variable and individual the change in the expected model prediction when conditioning on that variable.

***Table 1*: Main characteristics of the study population**

Variable/category		**All**		**Individuals without CKD[a]**		**Individuals with CKD[a]**	
		(N = 12,761)		**(N = 11,569)**		**(N = 1,192)**	
		N / median	IQR / %	N / median	IQR / %	N / median	IQR / %
Age in years	Baseline	39	33 to 46	48	33 to 45	38	40 to 57
	End of follow-up	49	41 to 56	56	41 to 55	49	50 to 65
Sex	Male	9,156	72	8,319	72	837	70
	Female	3,605	28	3,250	28	355	30
Ethnicity	White	9,964	78	8,851	77	1,113	93
	Black	1,825	14	1,783	15	42	4
	Hispanic	444	3	433	4	11	1
	Asian	482	4	458	4	24	2
	Other/unknown	46	0.4	44	0.4	2	0.2
Intravenous drug use prior to HIV diagnosis	Yes	2,287	18	2,047	18	240	20
	No	10,408	82	9,465	82	943	79

	Unknown	66	0.005	57	0.005	9	0.008
Ever smoked	Yes	7,906	62	7,158	62	748	63
	No	4,815	38	4,372	38	443	37
	Unknown	40	0.3	39	0.3	1	0.1
Hypertension	Yes	729	5.7	575	5.7	154	12.9
	No	11,963	94	10,928	94	1,035	86.8
	Unknown	69	0.5	66	0.5	3	0.3
eGFR[b]	Baseline	103	90 to 114	105	92 to 115	84	73 to 96
(ml/min/1.73m^2)	End of study	90	75 to 104	93	80 to 106	55	50 to 58
CD4 count (cells/µl)	Baseline	407	252 to 597	410	255 to 600	366	228 to 561
	End of study	615	426 to 830	621	437 to 839	536	362 to 759
Viral load	Baseline	883	0 to 35,173	1,040	0 to 36,000	174	0 to 23,459
(copies/ml)	End of study	0	0 to 0	0	0 to 0	0	0 to 0
Hepatitis B	Positive	510	4	464	4	46	4
	Negative	8,208	64	7,563	65	645	54
	Unknown	4,043	32	3,542	30	501	42

Hepatitis C	Positive	1,407	11	1,272	11	135	11
	Negative	10,022	79	9,142	79	880	74
	Unknown	1,332	10	1,155	10	177	15
	Baseline	2,259	18	2,100	18	159	13
Ever exposed to TDF	End of study	9,800	77	8,814	76	986	83
	Baseline	481	4	441	4	40	3
Ever exposed to ATV/r	End of study	3,629	28	3,135	27	494	41
	Baseline	1,783	14	1,577	14	206	17
Ever exposed to LPV/r	End of study	4,043	32	3,604	31	439	37

Abbreviations: ATV/r, atazanavir/ritonavir; CD4, cluster of differentiation 4; CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; HIV, human immunodeficiency virus, IQR, interquartile range; LPV/r, lopinavir/ritonavir; TDF, tenofovir disoproxil fumarate.

Note: All values are presented at baseline if not stated otherwise. We defined the baseline as the first creatinine measurement after January 1, 2002. Some potential risk factors are not presented, as these variables were not recorded during the entire study period.

[a] Within the observation period.

[b] Calculated using the ‘Chronic Kidney Disease Epidemiology Collaboration’ equation.

Table 2: **Performance of models to predict chronic kidney disease across different prediction horizons (N = 1,276 individuals; test set)**

Algorithm	**Visits used**	**Imputation method**	**F1-score**	**Precision**	**Recall**	**ROC-AUC**	**PR-AUC**
Prediction 90 days in advance							
Data-driven machine learning models (full models)							
— Multilayer perceptron	Last 2 visits[a]	Zero imputation	0.782	0.703	0.879	0.979	0.829
		Median forward	0.847	0.858	0.836	0.990	0.890
— Gradient boosting	Last 2 visits[a]	Zero imputation	0.874	0.852	0.897	0.994	0.933
		Median forward	0.890	0.875	0.905	0.996	0.956
— Random forest	Last 2 visits[a]	Zero imputation	0.583	0.942	0.422	0.995	0.943
		Median forward	0.836	0.918	0.767	0.994	0.931
— Elastic net	Last 2 visits[a]	Zero imputation	0.774	0.649	0.957	0.984	0.861
		Median forward	0.846	0.800	0.897	0.992	0.904
— Bidirectional recurrent neural network	Full sequence; all previous visits	Zero imputation	0.818	0.786	0.853	0.984	0.874
		Median forward	0.856	0.819	0.897	0.989	0.916

— Bidirectional attention recurrent neural network	Full sequence; all previous visits	Zero imputation	0.803	0.797	0.810	0.981	0.867
		Median forward	0.852	0.812	0.897	0.986	0.901
Manually built logistic regression model (short model)	Last 2 visits[a]	None	0.807	0.689	0.974	0.990	0.881
Prediction 180 days in advance							
Data-driven machine learning models (full models)							
— Multilayer perceptron	Last 2 visits[a]	Zero imputation	0.719	0.716	0.722	0.960	0.777
		Median forward	0.718	0.798	0.652	0.963	0.803
— Gradient boosting	Last 2 visits[a]	Zero imputation	0.656	0.859	0.530	0.969	0.833
		Median forward	0.789	0.815	0.765	0.970	0.860
— Random forest	Last 2 visits[a]	Zero imputation	0.115	>0.999	0.061	0.955	0.803
		Median forward	0.677	0.844	0.565	0.968	0.814
— Elastic net	Last 2 visits[a]	Zero imputation	0.698	0.629	0.783	0.952	0.768
		Median forward	0.767	0.777	0.757	0.959	0.787
— Bidirectional recurrent neural network	Full sequence; all previous	Zero imputation	0.722	0.732	0.713	0.965	0.759

	visits	Median forward	0.718	0.706	0.730	0.956	0.730
— Bidirectional attention recurrent neural network	Full sequence; all previous visits	Zero imputation	0.694	0.720	0.670	0.963	0.755
		Median forward	0.721	0.712	0.730	0.945	0.792
Manually built logistic regression model (short model)	Last 2 visits[a]	None	0.559	0.405	0.904	0.934	0.646
Prediction 270 days in advance							
Data-driven machine learning models (full models)							
— Multilayer perceptron	Last 2 visits[a]	Zero imputation	0.678	0.634	0.728	0.948	0.666
		Median forward	0.660	0.753	0.588	0.952	0.735
— Gradient boosting	Last 2 visits[a]	Zero imputation	0.290	0.833	0.175	0.944	0.702
		Median forward	0.689	0.745	0.640	0.957	0.728
— Random forest	Last 2 visits[a]	Zero imputation	0.068	>0.999	0.035	0.928	0.661
		Median forward	0.578	0.788	0.456	0.955	0.739
— Elastic net	Last 2 visits[a]	Zero imputation	0.647	0.566	0.754	0.942	0.702
		Median forward	0.650	0.756	0.570	0.943	0.716

— Bidirectional recurrent neural network	Full sequence; all previous visits	Zero imputation	0.605	0.581	0.632	0.938	0.649
		Median forward	0.661	0.632	0.693	0.940	0.737
— Bidirectional attention recurrent neural network	Full sequence; all previous visits	Zero imputation	0.664	0.630	0.702	0.931	0.678
		Median forward	0.664	0.699	0.632	0.934	0.693
Manually built logistic regression model (short model)	Last 2 visits[a]	None	0.453	0.310	0.842	0.893	0.504
Prediction 365 days in advance							
Data-driven machine learning models (full models)							
— Multilayer perceptron	Last 2 visits[a]	Zero imputation	0.641	0.691	0.598	0.950	0.699
		Median forward	0.628	0.776	0.527	0.950	0.722
— Gradient boosting	Last 2 visits[a]	Zero imputation	0.220	0.933	0.125	0.945	0.700
		Median forward	0.619	0.663	0.580	0.941	0.710
— Random forest	Last 2 visits[a]	Zero imputation	0.018	>0.999	0.009	0.941	0.705
		Median forward	0.527	0.800	0.393	0.952	0.725
— Elastic net	Last 2 visits[a]	Zero imputation	0.588	0.626	0.554	0.938	0.673

		Median forward	0.512	0.808	0.375	0.935	0.681
— Bidirectional recurrent neural network	Full sequence; all previous visits	Zero imputation	0.606	0.656	0.562	0.945	0.631
		Median forward	0.678	0.661	0.696	0.935	0.694
— Bidirectional attention recurrent neural network	Full sequence; all previous visits	Zero imputation	0.600	0.643	0.562	0.928	0.632
		Median forward	0.633	0.554	0.738	0.926	0.692
Manually built logistic regression model (short model)	Last 2 visits[a]	None	0.423	0.286	0.812	0.883	0.468

Abbreviations: ROC-AUC, area under the receiver operating characteristic curve; PR-AUC; area under the precision-recall curve.

[a] And summary statistics from earlier visits during the target observation period, as detailed in the methods section.

***Table 3*: How would you decide? Predicted and observed chronic kidney disease outcomes among three complex cases across prediction horizons (gradient boosting model estimates for illustration purposes)**

Individual	**Predicted outcome (CKD probability)**				**Observed outcome**				**Brief interpretation and key predictor for single individuals**
	Prediction horizon				Prediction horizon				
	90 days	180 days	270 days	365 days	90 days	180 days	270 days	365 days	
1	No CKD (0.34)	CKD (0.99)	CKD (0.51)	No CKD (0.01)	CKD	CKD	CKD	CKD	Platelet counts and various hematological parameters were strong predictors for CKD in this individual; however, this did not prevent false negative predictions at 90 and 365 days. There were dozens of moderate predictors of unclear clinical relevance: These factors have cancelled out at 365 days, as some were preventive and others suggested an incremental CKD risk. This

									example highlights that a clinician should review every machine learning prediction.
2	No CKD (0.18)	No CKD (0.00)	No CKD (0.00)	No CKD (0.00)	No CKD	No CKD	No CKD	No CKD	Absent cardiovascular risk factors (e.g. smoking) were strong predictors against CKD development. However, there were dozens of moderate predictors (potential preventive factors and risk factors) of unclear clinical relevance. The low CKD probability score across prediction horizons, together with a careful review of medical records, may be an indication for clinicians that CKD development is unlikely.
3	No CKD	CKD	No CKD	No CKD	No CKD	No CKD	No CKD	No CKD	Cardiovascular risk factors (e.g. high

	(0.28)	(0.71)	(0.00)	(0.02)		systolic blood pressure) and alcohol binge drinking increased the predicted CKD probability substantially ─ resulting in a false positive prediction at 180 days; however, high preceding eGFR values were strong predictors against CKD across prediction horizons.

Abbreviations: CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate.

Figure 1

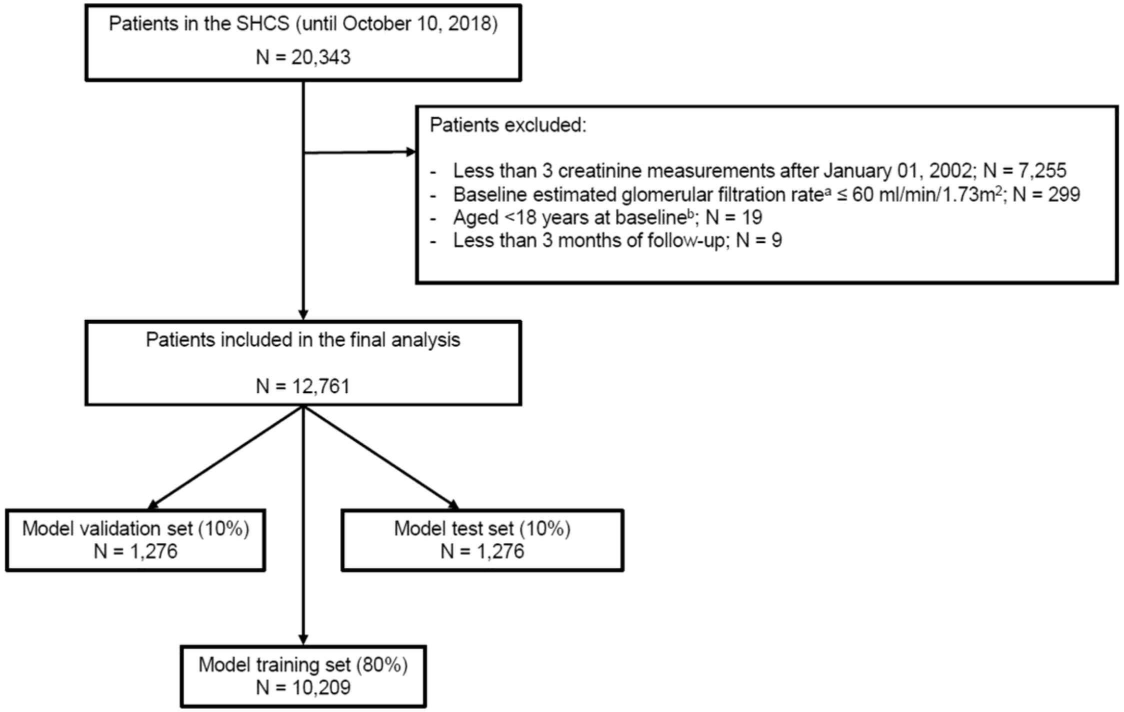

Patients in the SHCS (until October 10, 2018)
N = 20,343
Patients excluded:
- Less than 3 creatinine measurements after January 01, 2002; N = 7,255
- Baseline estimated glomerular filtration rate[a] ≤ 60 ml/min/1.73m^2; N = 299
- Aged <18 years at baseline[b]; N = 19
- Less than 3 months of follow-up; N = 9
Patients included in the final analysis
N = 12,761
Model validation set (10%)
N = 1,276
Model test set (10%)
N = 1,276
Model training set (80%)
N = 10,209

Figure 2

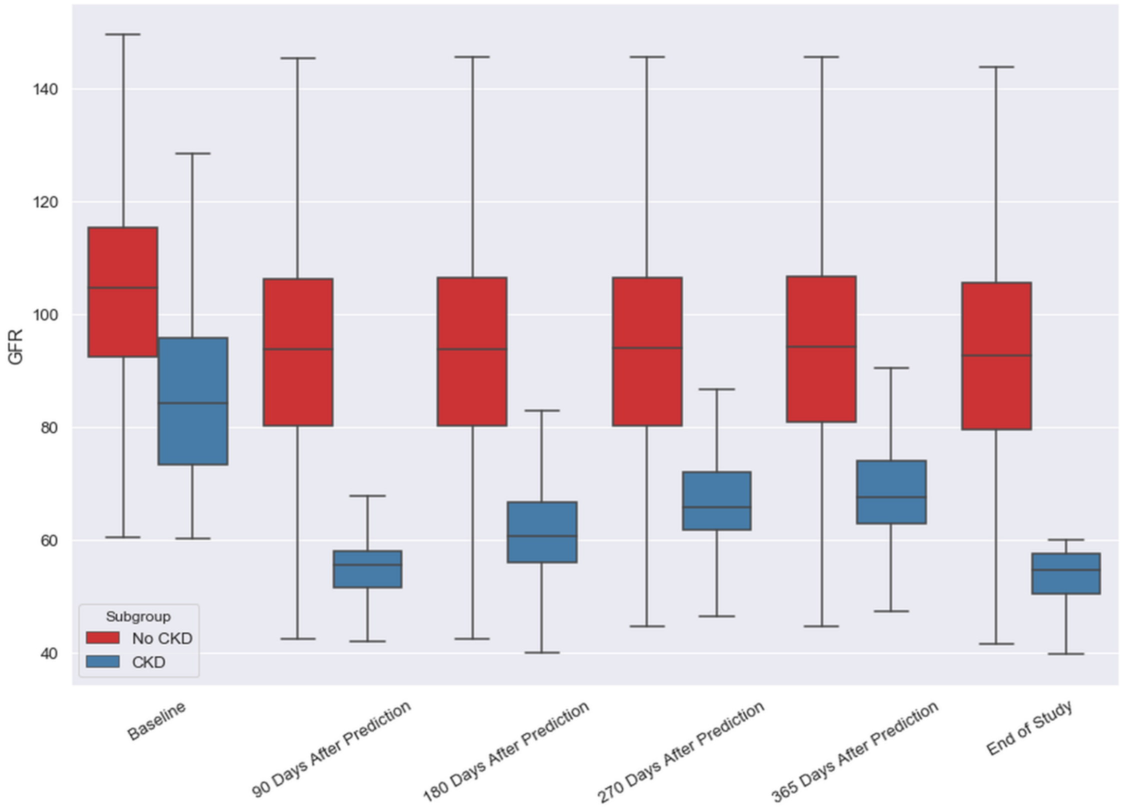
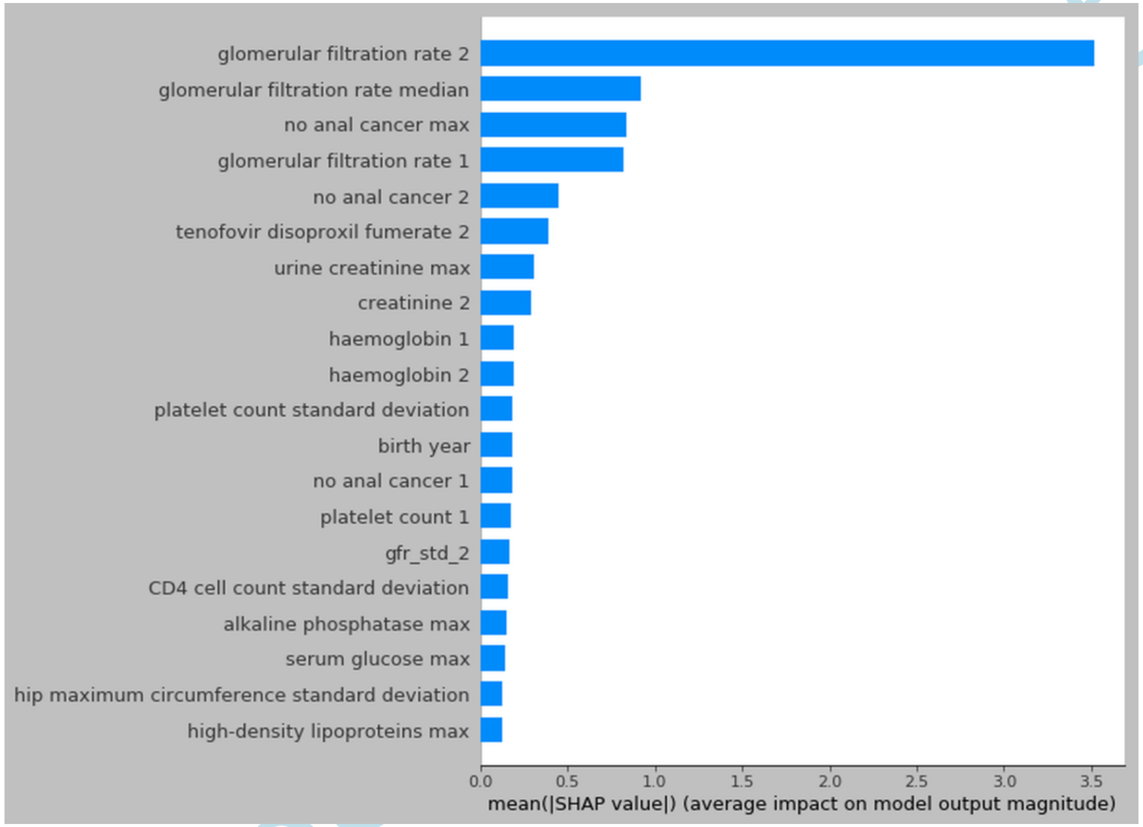

GFR
140
120
100
80
60
40
Subgroup
No CKD
CKD
Baseline
90 Days After Prediction
180 Days After Prediction
270 Days After Prediction
365 Days After Prediction
End of Study

Figure 3


glomerular filtration rate 2
glomerular filtration rate median
no anal cancer max
glomerular filtration rate 1
no anal cancer 2
tenofovir disoproxil fumerate 2
urine creatinine max
creatinine 2
haemoglobin 1
haemoglobin 2
platelet count standard deviation
birth year
no anal cancer 1
platelet count 1
gfr_std_2
CD4 cell count standard deviation
alkaline phosphatase max
serum glucose max
hip maximum circumference standard deviation
high-density lipoproteins max
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
mean(|SHAP value|) (average impact on model output magnitude)