SCIENTIFIC REPORTS

natureresearch

OPEN

Development of an *in silico* prediction system of human renal excretion and clearance from chemical structure information incorporating fraction unbound in plasma as a descriptor

Reiko Watanabe [1*], Rikiya Ohashi[1,2*], Tsuyoshi Esaki [1,3], Hitoshi Kawashima[1], Yayoi Natsume-Kitatani[1,4], Chioko Nagao[4] & Kenji Mizuguchi [1,4]

Prediction of pharmacokinetic profiles of new chemical entities is essential in drug development to minimize the risks of potential withdrawals. The excretion of unchanged compounds by the kidney constitutes a major route in drug elimination and plays an important role in pharmacokinetics. Herein, we created *in silico* prediction models of the fraction of drug excreted unchanged in the urine (f_e) and renal clearance (CL_r), with datasets of 411 and 401 compounds using freely available software; notably, all models require chemical structure information alone. The binary classification model for f_e demonstrated a balanced accuracy of 0.74. The two-step prediction system for CL_r was generated using a combination of the classification model to predict excretion-type compounds and regression models to predict the CL_r value for each excretion type. The accuracies of the regression models increased upon adding a descriptor, which was the observed and predicted fraction unbound in plasma ($f_{u,p}$); 78.6% of the samples in the higher range of renal clearance fell within 2-fold error with predicted $f_{u,p}$ value. Our prediction system for renal excretion is freely available to the public and can be used as a practical tool for prioritization and optimization of compound synthesis in the early stage of drug discovery.

The excretion process in the urine involves three main processes: glomerular filtration, tubular secretion, and reabsorption[1]. In glomerular filtration, only the unbound drugs in plasma are filtrated and enter the tubular lumen depending on the glomerular filtration rate (GFR) and the extent of the drug fraction unbound in plasma ($f_{u,p}$). Active tubular secretion is mediated by several transporters for numerous acidic, basic, and some large neutral compounds. A variety of transporters are expressed predominantly in the proximal tubule, executing sequential uptake and efflux that facilitates renal tubular secretion[2]. Reabsorption is mediated by passive diffusion and reuptake by transporters, with the former being especially important for exogenous compounds. Thus, renal excretion is a result of complicated multiple-transport systems, with previous studies reporting that compounds can be classified into reabsorption, intermediate, and secretion type depending on the ratio of renal clearance (CL_r) to glomerular filtration[3–5].

Two important pharmacological indicators in renal drug excretion include the fraction of drug excreted unchanged in urine (f_e) and renal clearance (CL_r). f_e is an important quantitative indicator showing the contribution of renal excretion for overall drug elimination and CL_r is defined as the proportionality term between urinary excretion rate of unchanged drug and plasma concentration[1]. Predicting the degree of f_e during the drug

[1]Laboratory of Bioinformatics, AI Center for Health and Biomedical Research, National Institute of Biomedical Innovation Health and Nutrition, Osaka, Japan. [2]Discovery Technology Laboratories, Mitsubishi Tanabe Pharma Corporation, Saitama, Japan. [3]The Center for Data Science Education and Research, Shiga University, Shiga, Japan. [4]Laboratory of In-silico Drug Design, Center of Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan. *email: reiko-watanabe@nibiohn.go.jp; rikiya-ohashi@nibiohn.go.jp

discovery stage is important to determine the basic principal for the subsequent development stage. Moreover, the use of renal excreted-type drugs should in general be avoided or administered at low dosages for patients with renal failure[6,7].

The pharmacokinetic profile of a drug is an amalgamation of various properties, such as dissolution, intestinal absorption, plasma protein binding, metabolism, biliary excretion, distribution, and renal excretion. Recently, computer-aided drug design using *in silico* models to predict the absorption, distribution, metabolism, excretion, and toxicity (ADMET) parameters[8–10] have attracted considerable attention in the field of drug development. This approach is effective to evaluate the physicochemical properties and *in vivo* pharmacokinetics during the early stages of drug discovery. In addition, the use of *in silico* prediction techniques minimizes the expenses and risks of subsequent withdrawals during clinical trials.

Properly validated *in silico* models for ADMET prediction can assist drug design by helping medicinal chemists prioritize suitable lead compounds in the optimization process of early drug discovery. Whereas industrial medicinal chemists may have access to comprehensive commercial suites to predict ADMET properties, this process is difficult for most academic researchers. Alternatively, models built using freely available computational tools can be easily shared with other researchers or can be integrated into other packages. Therefore, such models would constitute valuable assets for both academia and industry.

To the best of our knowledge, no models to predict f_e and CL_r based only on structure information have been developed using freely available software. For the prediction of f_e, Doddareddy *et al.*[11] generated a binary classification model of f_e from structural information calculated using Volsurf and Molconn-Z, with threshold values of f_e set to 0.2 in a dataset containing 130 compounds. This resulted in 65–80% of all test sets to be correctly predicted. Kusama *et al.*[12] established a binary classification model to predict the major clearance pathways and provided an online prediction system, CPathPred, which was subsequently improved by Toshimoto *et al.*[13] and Wakayama *et al.*[14]. In the latter prediction model[14], threshold values of f_e were set to 0.25 for the prediction of renal excretion, yielding an F-measure of 0.67 on the test set for renal excretion with the input of four fundamental parameters (charge, molecular weight [MW], *log*D, and $f_{u,p}$). To predict the CL_r, allometric scaling approaches and *in vitro–in vivo* extrapolation approaches have been extensively utilized. Nevertheless, although allometric scaling is a practical tool, it requires *in vivo* CL_r data in several animal species, which may be difficult to obtain by academic researchers[15,16]. The *in vitro–in vivo* extrapolation approaches have successfully determined and incorporated *in vitro* permeability data from Caco-2 or LLCPK1 cells into prediction models[17–19]; however, it remains necessary to experimentally determine the individual scaling factors. Furthermore, unique quantitative structure-pharmacokinetics relationships have been constructed to predict the CL_r of drugs or drug-like compounds in humans[20].

Although the accuracy of previously reported models has been improved[14,20], such models rely upon either the direct input of experimental values or commercial software for the calculation of descriptors or values of pKa and *log*D. It is difficult to find a free software that can calculate *log*D; moreover, even though ChemAxon (Marvin)[21] has the ability to calculate pKa on an individual basis, it is not possible to calculate this value for multiple compounds simultaneously using a command line. As it is essential to perform calculations batch-wise when new structures are brought into our prediction system, we could not find suitable free software to calculate *log*D and pK_a for the purpose of this open model.

Previously, we constructed prediction models of the human unbound fraction in plasma ($f_{u,p}$)[22], with the $f_{u,p}$ prediction models released via a freely available tool ($f_{u,p}$ Predictor, http://adme.nibiohn.go.jp/fup/). As approximately 10% of the blood volume is filtered at the glomerulus by the hydraulic pressure exerted by the arterial blood and, as a general rule, only the unbound drug in plasma is filtered, the value of $f_{u,p}$ significantly impacts the renal glomerular filtration[23]. Accordingly, Dave *et al.*[20] pointed out that the $f_{u,p}$ represents the most important determinant of CL_r prediction. Moreover, $f_{u,p}$ has been included as one of the four default descriptors in f_e prediction in several reports[12–14]. Thus, we considered that our $f_{u,p}$ prediction models[22] might be expanded to predict f_e and CL_r.

Here, we created f_e and CL_r datasets of 411 and 401 compounds, respectively, and generated two types of predictions: 1) binary classification models of f_e and 2) a two-step prediction system of CL_r through a combination of the classification and regression models, incorporating structure information without any experimental values but with predicted $f_{u,p}$ values, using freely available software. Moreover, the contribution of $f_{u,p}$ to the accuracy of regression models for CL_r prediction was considered. These *in silico* prediction models are freely available.

Methods

Data set preparation and descriptor calculation.   The dataset for f_e prediction was acquired from Benet *et al.*[24] and PharmaPendium[25]. The dataset for CL_r prediction was acquired from the ChEMBL database and the dataset reported by Varma *et al.*[3,26,27] and Ito *et al.*[5]. Both datasets were created after careful curation to select the values of f_e or CL_r in healthy adult humans for a single administration to obtain higher prediction accuracy[28]. The details of curation are provided in Supplementary Methods.

For the f_e, a dataset containing 411 compounds (343 from Benet *et al.*[24,27] and 68 from PharmaPendium) with f_e, $f_{u,p}$, and structure information was assembled (Dataset_f_e). The list of 343 compounds and their f_e values are summarized in Supplementary Table S1; detailed information for the 68 compounds acquired from PharmaPendium has not been presented owing to licensing restrictions.

For the CL_r, a dataset containing 401 compounds with experimental CL_r including $f_{u,p}$ values and structure information was assembled (Dataset_CL_r); the clearance ratio (CR)[5], which is also referred to as the renal extraction ratio[29], to categorize compounds into three excretion types was calculated using the following equation:

$$CR = CLr/(fu, p \times \text{GFR})$$

The GFR used in this study was 1.8 mL/min/kg (126 mL/min in a 70 Kg man). The compounds were categorized into three types based on their CR. The compounds that displayed CR < 0.67, 0.67 ≤ CR < 1.5, or 1.5 ≤ CR were classified into reabsorption (R) type (net reabsorbed compounds), intermediate (IM) type (apparently not reabsorbed or secreted compounds), and secretion (S) type (net secreted compounds), respectively[5]. Predicted $f_{u,p}$ was calculated using our previously developed $f_{u,p}$ predictor[22]. Ionization profiles in the data set were extracted from the ChEMBL database.

We employed the open source programs Mordred (ver. 1.0.0)[30] and PaDEL-Descriptor[31] to calculate the two-dimensional (2D) descriptors and fingerprints (Extended, KlekotaRoth, and AtomPairs2D), respectively. $LogD$pH7.4 and pK_a (apKa) values were calculated using ChemAxon calculator plugin software (Budapest, Hungary) because of the importance of $LogD$ and pK_a as pharmacokinetic parameters; these values were used only for visualizing the chemical space by principal component analysis (PCA).

Data analysis. Data analysis was performed in R (version 3.5.1[32]), and the results were visualised using the ggplot2[33] and ggfortify[34] packages. In total, 11 descriptors, i.e., MW, topological polar surface area, SLogP, LogD pH 7.4, apKa, bpKa, hydrogen bond acceptor (HBAcc), hydrogen bond donor (HBDon), number of aromatic atoms (nAromAtom), number of aromatic bonds (nAromBond), and the number of rotatable bonds (nRot), were used for PCA.

Processes of model construction. The caret[35] package in R was used to build the prediction models. An overview of the common process in model construction is shown in Supplementary Scheme S1. The data sets were split into training and test sets using random selection at a ratio of 8:2. In the training set, descriptors that showed near-zero-variance and absolute correlations >0.90 were identified and excluded by calculating the frequency ratio using the nearZeroVar function and by creating a correlation matrix using the findCorrelation function in the caret package. Thereafter, descriptors that significantly contributed to the prediction accuracy were selected using the Boruta[36] algorithm to automatically rank and omit descriptors based on the random forest (RF) classification algorithm with the training set. Boruta is a wrapper built around the RF classification algorithm implemented in the R package randomForest[37], which provides unbiased and stable selection of important and non-important attributes.

Prediction models were constructed using various machine learning techniques including linear and non-linear methods; i.e., RF, support vector machine (SVM with radial functions), artificial neural network (ANN), and partial least squares (PLS), to obtain the most accurate model for our data set. To adopt each technique, the train function was passed with method parameters set as rf, svm, nnet, and pls in the caret package. We used the automatic grid search of each tuning parameter with four (tuneLength = 4) values of each in the caret package to prioritize the optimal parameters for our predictions and models were created using a 10-fold cross validation. For 3-class classification, the RF algorithms can naturally handle multiclass classification, whereas all-versus-all and all-versus-rest approaches were used for multiclass SVM in the e1071 package[38] and multinomial log-linear models via neural networks in the nnet package[39], respectively. The generated models were evaluated with the test set. Kappa (True accuracy), balanced accuracy, sensitivity, and specificity obtained from the confusion matrix in classification models, and r-squared (r^2, coefficient of determination) and root mean squared error (RMSE) in regression models were used to evaluate their performance on the test set. The best models were chosen according to the value of Kappa or r^2 of the test set in the classification and regression model, respectively.

Model construction for f_e and CL_r prediction. As descriptors, more than 1600 2D descriptors calculated via Mordred and 5640 Extended, KlekotaRoth, and AtomPairs2D fingerprints generated using PaDEL-Descriptor were prepared, and descriptors for which the calculation failed were excluded (Supplementary Information 3). The 6974 and 6976 descriptors in f_e and CL_r prediction models were initially used for model construction and descriptors selected using the Boruta[36] algorithm were finally applied for the predictions. Dataset_f_e was split into 328 and 83 compounds for training and test sets, respectively, using random selection and the prediction model was constructed. Dataset_CL_r containing 401 compounds was split by random selection at a 1:9 ratio into 41 and 360 compounds to isolate the external test set. Thereafter, the other 360 compounds were split at 8:2 (278 and 72 compounds) for 3-class classification models; in parallel, the other 360 compounds were classified into three excretion types; 94 reabsorption (R), 86 intermediate (IM), and 180 secretion (S) type compounds according to their CR calculated using CL_r, $f_{u,p}$, and GFR values. Subsets were defined as Dataset_CL_r_R, Dataset_CL_r_IM, and Dataset_CL_r_S, respectively. An overview of CL_r model construction is shown in Supplementary Scheme S2.

Results

Distribution and chemical space analysis in Dataset_f_e and Dataset_CL_r.

Dataset_f_e and Dataset_CL_r, consisting of 411 and 401 compounds, respectively, were weighted towards the lower range of f_e and CL_r, with 220 compounds that were overlapped. Distribution of f_e in Dataset_f_e and CL_r with a logarithmic scale in Dataset_CL_r are shown in Fig. 1a,b, and that of CL_r in the original scale is shown in Supplementary Fig. S1; this characteristic was also observed regarding the data sets used in previous reports[11,20]. The chemical spaces of the two datasets were visualized by PCA along with classification, with the threshold set to 0.30 in Dataset_f_e (Fig. 1c) and with CR types such as R, IM, and S in Dataset_CL_r (Fig. 1d). A total of 11 descriptors, all of which are generally considered to be important parameters for synthetic expansion, were used for the analysis. Compounds with higher f_e were less lipophilic than those with lower f_e, reflecting the fact that water soluble drugs generally undergo renal excretion. In Dataset_CL_r, most of chemical space in R, IM and S type were overlapped, and it was difficult to separate the three classes using these 11 descriptors, indicating that R, IM, S compounds have similar physicochemical properties (Fig. 1d). It was considered reasonable that R type compounds showed a lower CL_r, S type compounds showed a higher CL_r and IM type compounds showed medium CL_r (Fig. 1e). The averages of CL_r were 0.20, 1.02, and 2.50 mL/min/kg in R, IM, and S types, respectively. The relationship between f_e and CL_r or

(a)

(b)

(c)

(d)

(e)
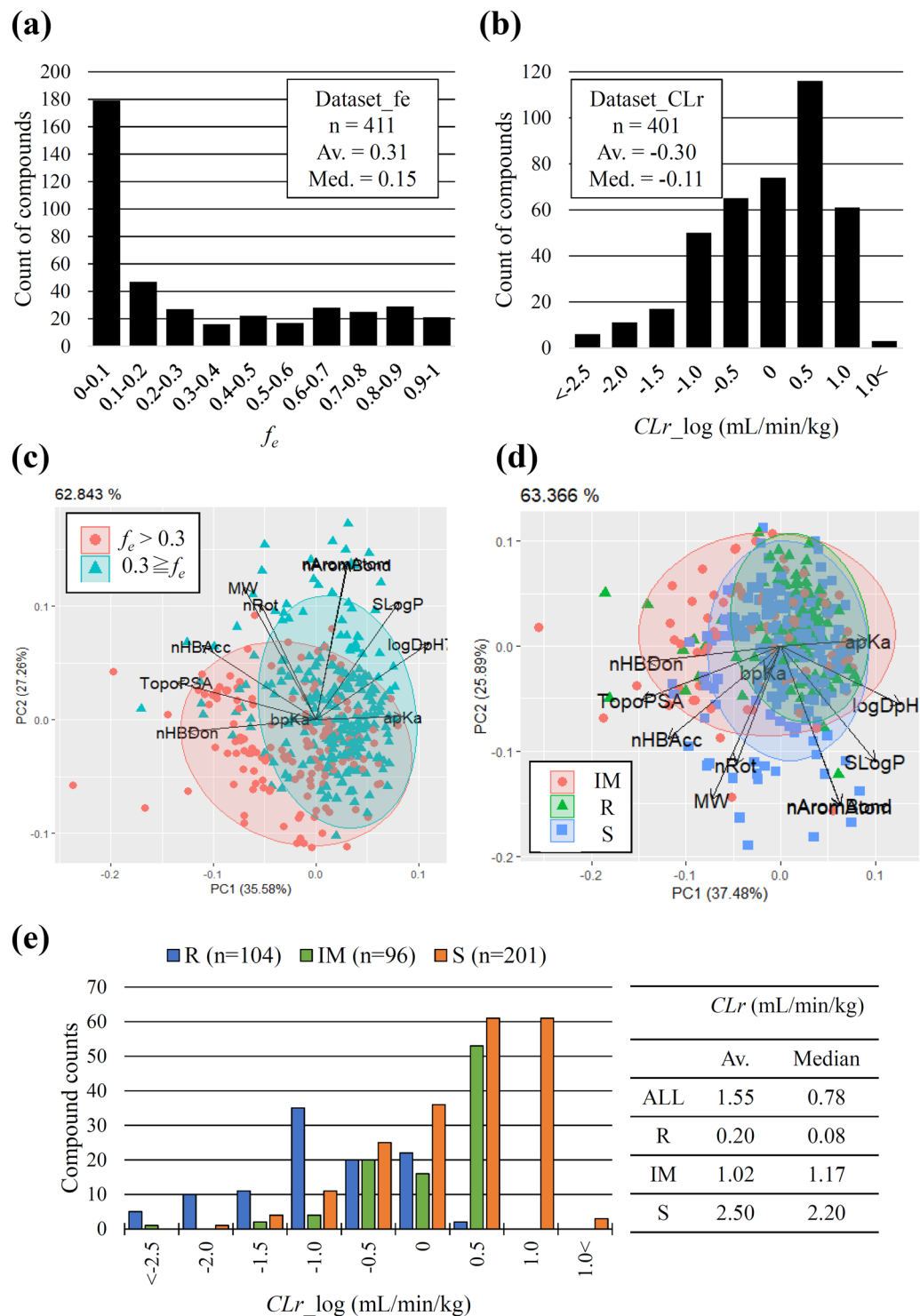
Figure 1. (**a**) Distribution of f_e in Dataset_f_e consisting of 411 compounds. Average and median are shown in the top-right. (**b**) Distribution of CL_r with logarithmic scale in Dataset_CL_r consisting of 401 compounds. Average and median are shown in the top left. (**c**) The chemical space of Dataset_f_e with classification by the threshold set to 0.30. The frames indicate 95% normal confidence ellipses in the assembled 411 compounds with $f_e \geq 0.3$ (red) and $f_e < 0.3$ (green). (**d**) The chemical space of Dataset_CL_r in 96 inte,mediate (IM, red circle), 104 reabsorption (R, green triangle), and 201 secretion (S, blue square) types. (**e**) Plot of compound counts depending on CR type. Average and median of CL_r in each CR type are shown on the right.

observed $f_{u,p}$ in logarithmic scale, depends on the ionization properties of the compounds, was also analysed. No trend existed in the distribution of CLr in each ionization property and the assembled data set spanned a chemical space similar to that of the approved drugs (Supplementary Fig. S2).

Descriptor	Selected descriptors	Training or Test	Parameter	RF[a] (Model_*fe*)	SVM[a]	ANN[a]	PLS[a]
without $f_{u,p}$	51	Training	Kappa	0.50	0.46	0.50	0.52
		Test	Kappa	0.49[b]	0.29	0.37	0.38
			Bal. Acc.	0.74	0.63	0.69	0.68
			Sensitivity	0.65	0.39	0.61	0.45
			Specificity	0.84	0.88	0.76	0.90

Table 1. Statistical results of the binary classification models for f_e prediction by each of the four models. [a]RF, Random forest; SVM, Support Vector Machine with radial functions; ANN, artificial neural network; PLS, partial least squares. [b]The highest kappa in the test set among four models.

Classification models to predict the extent of f_e.

Binary classification models were created with f_e threshold value set to 0.30 to define the low and high/medium classes, with 158 and 253 compounds classified into the high/medium and low class, respectively. These thresholds were chosen according to previous reports[2,14]. Fifty one descriptors were finally selected in the training set using the Boruta algorithm[36]. Prediction models were trained in a training set comprising 328 compounds, to which four machine learning methods (RF, SVM with radial, ANN, and PLS) were applied. Each model was validated on the common test set containing 83 compounds; the statistical results of the models are summarized in Table 1. Kappa was 0.46–0.52 and 0.29–0.49 in the training and test set, respectively. Balanced accuracy and specificity, which is the ratio to successfully distinguish the low f_e class, were 0.63–0.74 and 0.76–0.90 in the test set. RF showed the highest Kappa in the test set; RF parameters (ntree and ntry) were 500 and 14, and the model was defined as Model_f_e. In parallel, to evaluate the statistical influence of $f_{u,p}$ as a descriptor to f_e prediction accuracy, prediction models of f_e were constructed with or without $f_{u,p}$ values (observed and predicted). Paired t-test analysis revealed no significant difference between the Kappa of Model_f_e and those of other models with $f_{u,p}$ (Supplementary Table S2).

SLogP was the most important descriptor in all the models, whereas $f_{u,p}$ was listed as a second important descriptor in the models with $f_{u,p}$. The top ranked descriptors according to their variable importance for the best models are listed in Supplementary Table S3, and the main important descriptors were common to all the three models including other lipophilic descriptors such as SlogP. In addition, topological descriptors such as ATS (Moreau-Broto autocorrelation), MATS (Moran autocorrelation), GATS (Geary autocorrelation), chi related index (Molecular connectivity), and ETA (Extended topochemical atom) were also determined as important descriptors.

Relationship between CL_r and $f_{u,p}$.

The relationship between CL_r and $f_{u,p}$ was analysed in Dataset_CL_r. The correlation coefficient (r) between CL_r and observed $f_{u,p}$ in logarithmic scale was moderate ($r = 0.54$) (Fig. 2a); however, the correlation between CL_r and observed $f_{u,p}$ was increased ($r = 0.72, 0.98, 0.80$ in R, IM, and S type, respectively) in the subsets with the CR types (Fig. 3b), suggesting that $f_{u,p}$ values used as a descriptor are likely effective to create CL_r prediction models in the sub-clustered dataset by CR types. In comparison, the correlation did not change in a subset of Dataset_CL_r with ionization properties (Supplementary Fig. S3). In addition, $f_{u,p}$ in the IM type was significantly higher than that in the other types (Fig. 2c). This indicated that the mechanism of renal excretion in these compounds is mainly glomerular filtration, with the contribution of secretion by transporters or reabsorption by lipophilicity being low.

Furthermore, upon comparison of the observed and predicted $f_{u,p}$ values, as shown in Supplementary Fig. S4, a correlation could be seen between observed and predicted $f_{u,p}$ values ($r = 0.84$), with 72.8% and 84.0% of the predicted $f_{u,p}$ values falling within 2-fold and 3-fold error, respectively. This indicated that the $f_{u,p}$ predicted by $f_{u,p}$ predictor[22] correlated well with the observed $f_{u,p}$.

Prediction models for CL_r.

A comprehensive CL_r prediction model incorporating the whole Dataset_CL_r using several machine learning methods was constructed for a randomly selected training set. This was validated by the test set with or without $f_{u,p}$ values. Although the average of r^2 appeared to slightly increase (from 0.24 to 0.32) when $f_{u,p}$ was added as a descriptor, the highest r^2 of all the models was 0.4 in the test set (Supplementary Table S4). As previously reported by Dave et al.[20], a single model was not able to predict the renal clearance of all examined compounds.

As a next step, subsets of Dataset_CL_r by CR type were generated and defined as Dataset_CL_r_R, Dataset_CL_r_IM, and Dataset_CL_r_S as described in the experimental section. Regression models to predict the value of CL_r were generated using four machine learning methods (RF, SVM with radial functions, ANN, and PLS). Three types of descriptors were applied: 1) 6,976 descriptors, 2) 6,976 descriptors + predicted $f_{u,p}$, and 3) 6,976 descriptors with observed $f_{u,p}$ in each dataset. The statistical results of each model are summarized in Table 2, and r^2 of the best model and average of r^2 among several models with different randomized split of training and test set are shown. The p-values were calculated using the paired t-test with r^2 against models without $f_{u,p}$. All the models showed a significantly higher r^2 when $f_{u,p}$ values were applied as descriptors: r^2 in the test set increased from 0.38 to 0.66, 0.56 to 0.92, and 0.41 to 0.62 in the R, IM, and S type, respectively when the observed $f_{u,p}$ was included as a descriptor, indicating that inclusion of $f_{u,p}$ values as a descriptor increased the accuracy of the prediction model. In addition, r^2 in the test set also increased significantly with predicted $f_{u,p}$ values, and its r^2 values were slightly lower than those of the models with observed $f_{u,p}$. In the model with predicted $f_{u,p}$ values, the PLS in R types and RF in IM and S type showed the best prediction capability, defined as Model_CL_r_R, Model_CL_r_IM, and Model_CL_r_S, respectively. Fold error of the best models are also summarized; the percentage of samples within
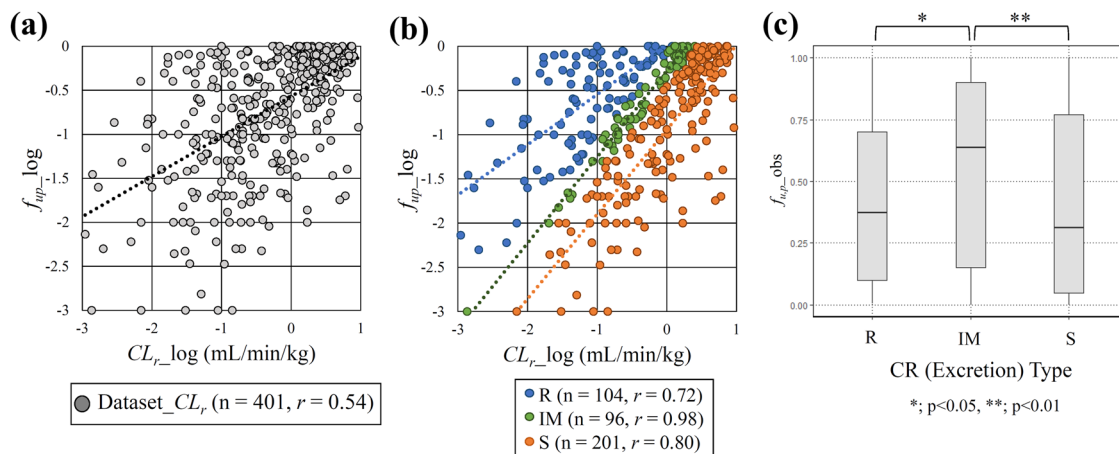
Figure 2. Relationship between CL_r in logarithmic scale and observed $f_{u,p}$. (**a**) Whole Dataset_CL_r (401 compounds), and (**b**) sub-categorized by CR type (104, 96, and 201 compounds in reabsorption [R], intermediate [IM] and secretion [S] type, respectively). (**c**) Boxplot of observed $f_{u,p}$ in each excretion type. n; compound counts, r; correlation coefficient.
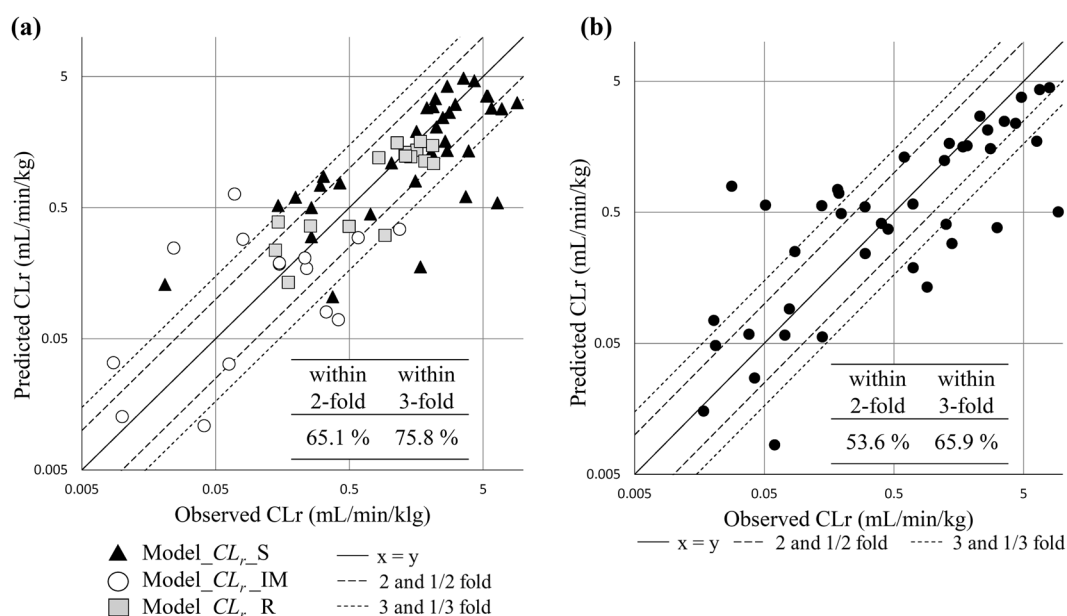
Figure 3. Plot of predicted and observed CL_r by three regression models with predicted $f_{u,p}$ value. (**a**) in the test set (66 compounds) and (**b**) external test set (41 compounds).

2-fold error increased from 37.5% to 56.3% in R type, 68.8% to 100% in IM type, and from 48.6% to 62.9% in S type compounds using the observed $f_{u,p}$ as a descriptor. The percentage of samples within the 2-fold error also increased with predicted $f_{u,p}$, as compared with that in the models without $f_{u,p}$ (to 43.8, 87.5, and 57.1% in R, IM, and S type, respectively). To ensure that this result was not derived from the inclusion of training compounds in the $f_{u,p}$ prediction model, whose $f_{u,p}$ can be predicted accurately in general, compounds included in the training set of the $f_{u,p}$ prediction model were excluded from the test set, with fold errors indicated in parentheses. Although the number of data sets in R type was small and this could accordingly not be compared accurately, a same trend was observed when using the entire data set in IM and S type. Predicted and observed CLr using Model_CL_r_R, Model_CL_r_IM, and Model_CL_r_S in the test set and the external test set containing 41 compounds were plotted in Fig. 3a,b, 75.8% and 65.9% of the compounds fell into within 3-fold error, respectively.

The top ranked descriptors according to their variable importance for the three defined best models and a description of those descriptors are summarized in Supplementary Tables S5 and S6. Predicted $f_{u,p}$ was the most important descriptor in all the models.

To actualize the CL_r prediction using structure information alone, three-class classification models to distinguish CR types (R, IM, and S) were constructed. The statistical results are summarized in Table 3. The RF models showed the highest Kappa (true accuracy) value of 0.32 in the test set, and balanced accuracy of 0.70,

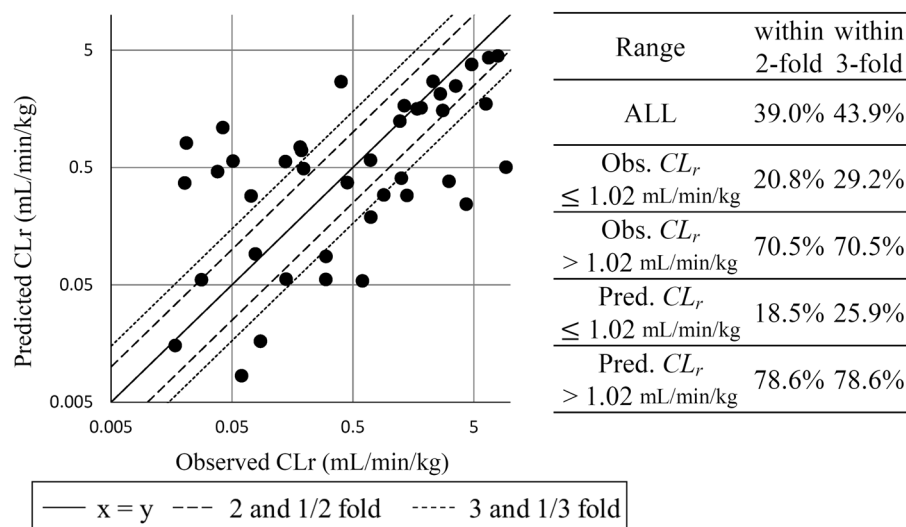| CR type	Descriptor set	Training or Test	The best model				Average	Method[a]
			r^2	RMSE	Within 2-fold error (%)	Within 3-fold error (%)	r^2	
Reabsorption Type (R)	Without $f_{u,p}$	Training	0.48	0.56	—	—	0.50	RF
		Test	0.38	0.61	37.5 (33.3)	43.8 (33.3)	0.23	
	With observed $f_{u,p}$	Training	0.71	0.44	—	—	0.62*	RF
		Test	0.66	0.46	56.3 (33.3)	62.5 (33.3)	0.53*	
	With predicted $f_{u,p}$ (Model_CL_r_R)	Training	0.57	0.51	—	—	0.52*	PLS
		Test	0.52	0.54	43.8 (16.7)	50.0 (33.3)	0.47*	
Intermediate Type (IM)	Without $f_{u,p}$	Training	0.65	0.38	—	—	0.65	SVM
		Test	0.56	0.28	68.8 (60.0)	93.8 (90.0)	0.43	
	With observed $f_{u,p}$	Training	0.95	0.17	—	—	0.94*	RF
		Test	0.92	0.12	100 (100)	100 (100)	0.88*	
	With predicted $f_{u,p}$ (Model_CL_r_IM)	Training	0.77	0.29	—	—	0.82*	RF
		Test	0.74	0.21	87.5 (83.3)	100 (100)	0.68*	
Secretion Type (S)	Without $f_{u,p}$	Training	0.43	0.51	—	—	0.46	RF
		Test	0.41	0.46	48.6 (35.0)	68.6 (60.0)	0.36	
	With observed $f_{u,p}$	Training	0.64	0.39	—	—	0.65*	RF
		Test	0.62	0.37	62.9 (55.0)	80.0 (75.0)	0.57*	
	With predicted $f_{u,p}$ (Model_CL_r_S)	Training	0.60	0.42	—	—	0.58*	RF
		Test	0.58	0.40	57.1 (50.0)	80.0 (65.0)	0.46*	

Table 2. Statistical results and fold error of the best regression models for CL_r prediction with or without $f_{u,p}$. [a]RF, Random forest; SVM, Support Vector Machine with radial functions; PLS, partial least squares; RMSE, root mean squared error. *p-value calculated using the paired t-test with Kappa against model without $f_{u,p}$ in each CR type ($p < 0.05$).

Model	Selected descriptors (n)	Training or Test set	Parameter	CR type	RF[a] (Model_CL_r_CR)	SVM[a]	ANN[a]	PLS[a]
Without $f_{u,p}$	15	Training	Kappa	—	0.34	0.34	0.32	0.29
		Test	Kappa	—	0.32[b]	0.19	0.18	0.22
			Sensitivity	R	0.56	0.56	0.56	0.50
				IM	0.29	0.12	0.41	0.18
				S	0.75	0.69	0.47	0.75
			Balanced Accuracy	R	0.7	0.69	0.68	0.68
				IM	0.58	0.50	0.59	0.54
				S	0.68	0.58	0.52	0.59

Table 3. Statistical results of the 3-class classification models for CL_r prediction. [a]RF, Random forest; SVM, Support Vector Machine; ANN, artificial neural network; PLS, partial least square. [b]The highest Kappa shown in the test set.

0.58, and 0.68 in R, IM, and S type, respectively, and were defined as Model_CL_r_CR. Although sensitivity in the R and IM type was not sufficiently high (0.56 and 0.29, respectively), 75% of S type compounds were successfully categorized into the correct type. The other raw parameters are shown in Supplementary Table S7. We also constructed three-class classification models with or without $f_{u,p}$; no significant difference in the accuracy were detected (Supplementary Table S8).

CL_r was predicted with the two-step prediction using CL_r regression models (Model_CL_r_R, Model_CL_r_IM, and Model_CL_r_S) following the prediction of CR type by a three-class classification model (Model_CL_r_CR). An external test set consisting of 41 compounds that were not included to generate any model was used for the validation. The observed and predicted CL_r values are plotted in Fig. 4; 39.0% and 43.9% of the predicted CL_r values fell into 2- and 3-fold error ranges, respectively. An external validation set was then split into the higher and lower range of observed or predicted CL_r with an average value of CL_r in IM type compounds ($CL_r = 1.02$ mL/min/kg). When the compounds were split according to observed value of CL_r, 70.5% of the compounds fell within 2-fold error in the higher range, and 20.8% and 29.2% of the observed CL_r values fell within 2- and 3-fold error in the lower range of CL_r. When the compounds were split by predicted value of CL_r, more compounds fell within 2- and 3-fold error in the higher range than in the lower range (78.6% in the higher range and 18.5% and 25.9% in the lower range). Using a combination of the classification model of CR type and the regression model of CL_r in R, IM, and S type, CL_r could be predicted from the structure information using only the freely available software, especially in the higher range of CL_r. We also tried two step CL_r prediction models with or without $f_{u,p}$ and fold error into 2- and 3- fold were not different (Supplementary Table S9).

Range	within 2-fold	within 3-fold
ALL	39.0%	43.9%
Obs. CL_r ≤ 1.02 mL/min/kg	20.8%	29.2%
Obs. CL_r > 1.02 mL/min/kg	70.5%	70.5%
Pred. CL_r ≤ 1.02 mL/min/kg	18.5%	25.9%
Pred. CL_r > 1.02 mL/min/kg	78.6%	78.6%

Figure 4. Plot of predicted and observed CL_r in the external validation set consisting of 41 compounds by the two-step prediction system with predicted $f_{u,p}$ value.

Discussion

We developed an *in silico* prediction system to classify compounds into their degree of unchanged excretion in the urine and to predict the value of CL_r using freely available tools without requiring any experimental data. Initially, a binary prediction model of f_e was successfully generated; the threshold was set to 0.30 according to Varma *et al.*[40], to define the compounds that are well- or poorly-eliminated in the urine. The inclusion of $f_{u,p}$ did not significantly affect the Kappa in the f_e prediction models; rather, Model_f_e without $f_{u,p}$ was sufficiently able to predict f_e, equivalent to the results of previous studies[11–14]. The majority of the important variables identified in the generated models to predict f_e were common, such that descriptors related to lipophilicity such as SLogP, topological descriptors related to electronic energy, and ionization potential indicators such as AATS, GATS, MATS, and chi comprised the key components of the models. Because lipophilicity is an important determinant for the choice between liver and renal excretion, it is natural that SLogP was the most important descriptor in all the models. In addition, hydrogen bonding interaction descriptors, including ionization potential, total energy, electronic energy, and sum of the total net charge were included in the previously constructed models[4]. Therefore, the inclusion of the descriptors related to lipophilicity, electronic energy, and ionization potential led to the models being able to successfully capture the key factors for f_e prediction. Drug metabolism is generally important as one of the determinants for f_e, because the compounds that are well metabolized show smaller values of f_e[27,40–42]. We believe that it is ideal to predict f_e in consideration of metabolic clearance as a task in future model construction because our f_e prediction model did not take metabolism into consideration; this matter should be addressed in future studies wherein metabolic information has been collected.

In general, renal impairment alters drug efficacy, often increasing their pharmacological and toxicological effects owing to high concentrations[7]. Moreover, hepatic clearance is known to be impaired in patients with end-stage renal disease because of the accumulation of uremic toxins, which is influenced by the expression of several CYPs[43–45]. Information on renal clearance is useful in the early stages of drug discovery, not only for understanding pharmacokinetic profiles but also for avoiding potential risk in the population with renal impairment, as well as in those with renal disease and advanced age[4]. Our binary model (Model_f_e) can be used to screen lead compounds in the early stage of drug discovery (Fig. 5 left). For example, Model_f_e is appropriate for selecting compounds showing low f_e that are not eliminated via the kidney, with an assumption that the drug could be administered to patients with renal impairment.

We concluded that a single *in silico* CL_r prediction model was unable to predict CL_r even if the $f_{u,p}$ value was applied as a descriptor, and no discernible linkages between CL_r and ionization property were observed in our study. Comprehensive prediction will be difficult because renal excretion is a result of multiple processes with different mechanisms such as glomerular filtration, secretion, and reabsorption, which are mediated by active transport and passive diffusion by lipophilicity. This interpretation is in accordance with those of Dave *et al.*[20], who also reported that splitting these compounds according to their ionization property did not improve prediction accuracy of CL_r. Dave *et al.*[20] finally constructed quantitative structure-pharmacokinetics relationships models that could be used to predict CL_r of compounds that (1) undergo net reabsorption, and (2) are substrates and/or inhibitors of human renal transporters. Although the models were accurate, the experimental information, such as class of the compounds in the Biopharmaceutics Drug Disposition Classification System (BDDCS)[24] and whether those compounds are substrates and/or an inhibitor of renal transporters, is required in advance to determine suitable prediction models. Thus, we aimed to generate a CL_r prediction model in which an external input is not required, using only chemical structure information for devising a practical tool in drug design processes prior to chemical synthesis.
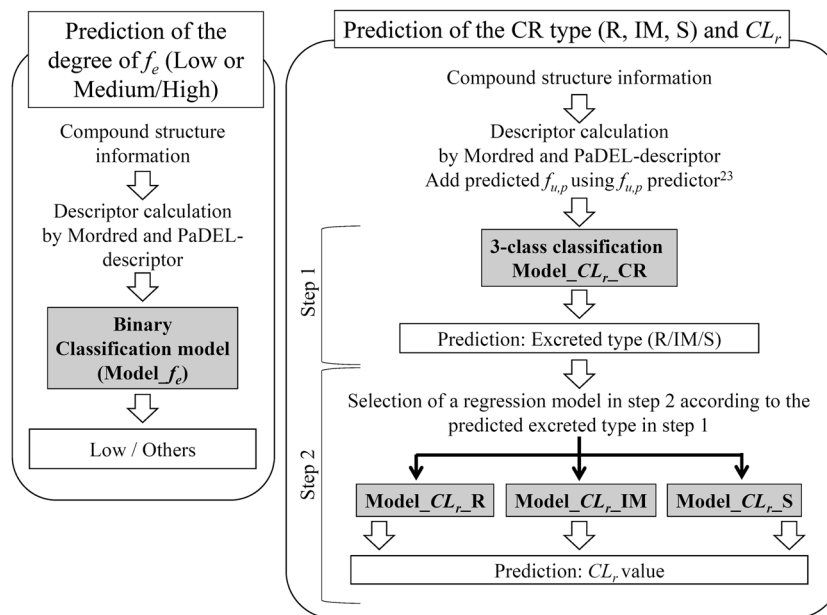
Figure 5. Application of the generated prediction models. Left: *In silico* prediction system for f_e in humans. Right: Two step *in silico* prediction system for CL_r in humans. R; Reabsorption, IM; Intermediate, S; Secretion.

Previously, $f_{u,p}$ was reported as the most important determinant of renal excretion[5,12,20]. However, the inclusion of $f_{u,p}$ as a descriptor did not significantly affect f_e and CR type prediction accuracy when the whole dataset was used in this study. In contrast, r^2 of the regression models with the subset of each CR type was significantly increased when observed and predicted $f_{u,p}$ values were included (Table 2). The results suggest that because of the multiple mechanisms of renal excretion, the impact of $f_{u,p}$ was observably low in the overall prediction, whereas when Dataset_CL_r was subclustered into three CR types, the influence of $f_{u,p}$ became more visible among the compounds with similar mechanisms.

The appearance of a drug in the urine is the net result of glomerular filtration, secretion, and reabsorption, for which CL_r is defined as follows:

$$CLr = (1 - FR)\ (fu,\ p \times GFR + CLs)$$

where FR and CL_s are the fractions reabsorbed from the lumen and the secretion clearance, respectively. When the compounds belong to R, IM, and S types, CL_r is expressed by the following respective equations:

$$\text{Reabsorption type (R): } CLr = (1 - FR)\ (fu,\ p \times GFR)$$

$$\text{Intermediate type (IM): } CLr = fu,\ p \times GFR$$

$$\text{Secretion type (S): } CLr = fu,\ p \times GFR + CLs$$

All the R, IM, and S type are proportionally affected by $f_{u,p}$, and $f_{u,p}$ directly affects the value of CL_r especially in the IM type. On the other hand, FR and CL_s can also affect the values of CL_r in addition to $f_{u,p}$ in the R and S type; information on renal transporters or metabolism related to FR and CL_s is important for CL_r prediction in these types. In addition, when the averages of r^2 in Table 2 were compared, r^2 was increased to the greatest degree in the IM type model (from 0.43 to 0.88 in the test set).

As shown in Figs. 3 and 4, the two-step prediction model of CL_r was generated using a combination of several models. As a first step, the CR type could be predicted using a three-class classification model (Model_CL_r_CR). As a second step, one of the three regression models (Model_CL_r_R, Model_CL_r_IM, or Model_CL_r_S) was chosen according to the prediction results of Model_CL_r_CR; then the final values of CL_r were predicted. It should be mentioned that 12 out of 13 compounds that were miss-classified in the first three-class classification did not fall within 3-fold error in the final CL_r prediction, indicating that improved accuracy in step 1 is necessary. Although it was difficult to identify a commonality among miss-classified compounds, cationic charges were frequently included in these miss-classified compounds (Fig. S5). Addition of similar compounds to the dataset or inclusion of pK_a or logD information as descriptors which are related to charges will be effective to get higher accuracy. In the present study, we could not include pK_a or logD as a descriptor because of the difficulties to find the freely available pK_a or logD calculators suitable for our prediction system. Therefore, it is necessary to take into consideration that the accuracy of CL_r prediction is low, particularly when the value of predicted CL_r is <1.02 mL/min/kg. However, in contrast, 78.6% of the compounds in the higher range of predicted CL_r were within 2-fold error, indicating that the results of compounds predicted to be >1.02 mL/min/kg are sufficiently reliable. This can be

used for the designing of compounds and subsequent optimization of lead compounds in the early stages of drug discovery (Fig. 4 right).

Our dataset is one of the largest among those previously reported[3,14]. However, several hundreds of compounds were not sufficient to account for all potential diversity. We hope to further expand the number of compounds although it has been difficult to retrieve quality data from the public databases in the present circumstance. It is, therefore, desirable to develop an integrated database with curated data of high quality and sufficient compounds to cover a larger chemical space.

We have developed a prediction system of renal excretion focused on f_e and CL_r based on structure information alone using freely available software, which is available to the public. The prediction of CL_r values from structure information was made possible using a two-step prediction, with three regression models to predict the value of CL_r depending on CR type, following three-class classification into three CR types. Moreover, the accuracies of the regression models were increased by adding observed and predicted $f_{u,p}$ values, with contribution of $f_{u,p}$ being the highest in the regression models of IM type. In the external validation set, 78.6% of the samples fell within 2-fold error in the higher range of CL_r. These prediction systems of renal excretion are expected to be practical tools, helping medicinal chemists to prioritize the actual synthesis of compounds during the drug design process before synthesis. A new web resource (http://adme.nibiohn.go.jp/renal_ex) has been established to access the online system for the prediction of overall renal excretion, as described in this study.

References

1. Rowland, M., Tozer, T. N. & Rowland, M. *Clinical pharmacokinetics and pharmacodynamics: concepts and applications*. 4th edn, (Lippincott William & Wilkins, 2011).
2. Morrissey, K. M., Stocker, S. L., Wittwer, M. B., Xu, L. & Giacomini, K. M. Renal transporters in drug development. *Annu Rev Pharmacol Toxicol* **53**, 503–529, https://doi.org/10.1146/annurev-pharmtox-011112-140317 (2013).
3. Varma, M. V. *et al.* Physicochemical determinants of human renal clearance. *J Med Chem* **52**, 4844–4852, https://doi.org/10.1021/jm900403j (2009).
4. Feng, B., LaPerle, J. L., Chang, G. & Varma, M. V. Renal clearance in drug discovery and development: molecular descriptors, drug transporters and disease state. *Expert Opin Drug Metab Toxicol* **6**, 939–952, https://doi.org/10.1517/17425255.2010.482930 (2010).
5. Ito, S. *et al*. Relationship between the urinary excretion mechanisms of drugs and their physicochemical properties. *J Pharm Sci* **102**, 3294–3301, https://doi.org/10.1002/jps.23599 (2013).
6. Delco, F., Tchambaz, L., Schlienger, R., Drewe, J. & Krahenbuhl, S. Dose adjustment in patients with liver disease. *Drug Saf* **28**, 529–545, https://doi.org/10.2165/00002018-200528060-00005 (2005).
7. Doogue, M. P. & Polasek, T. M. Drug dosing in renal disease. *Clin Biochem Rev* **32**, 69–73 (2011).
8. Wang, Y. *et al*. In silico ADME/T modelling for rational drug design. *Q Rev Biophys* **48**, 488–515, https://doi.org/10.1017/S0033583515000190 (2015).
9. Morales, J. F., Montoto, S. S., Fagiolino, P. & Ruiz, M. E. Current State and Future Perspectives in QSAR Models to Predict Blood-Brain Barrier Penetration in Central Nervous System Drug R&D. *Mini Rev Med Chem* **17**, 247–257 (2017).
10. Bergstrom, C. A. S. & Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int J Pharm* **540**, 185–193, https://doi.org/10.1016/j.ijpharm.2018.01.044 (2018).
11. Doddareddy, M., Cho, Y., Koh, H., Kim, D. & Pae, A. In silico renal clearance model using classical Volsurf approach. *J Chem Inf Model* **46**, 1312–1320 (2006).
12. Kusama, M. *et al*. In silico classification of major clearance pathways of drugs with their physiochemical parameters. *Drug Metab Dispos* **38**, 1362–1370, https://doi.org/10.1124/dmd.110.032789 (2010).
13. Toshimoto, K. *et al*. In silico prediction of major drug clearance pathways by support vector machines with feature-selected descriptors. *Drug Metab Dispos* **42**, 1811–1819, https://doi.org/10.1124/dmd.114.057893 (2014).
14. Wakayama, N. *et al*. In Silico Prediction of Major Clearance Pathways of Drugs among 9 Routes with Two-Step Support Vector Machines. *Pharm Res* **35**, 197, https://doi.org/10.1007/s11095-018-2479-1 (2018).
15. Kunze, A., Huwyler, J., Poller, B., Gutmann, H. & Camenisch, G. In vitro-in vivo extrapolation method to predict human renal clearance of drugs. *J Pharm Sci* **103**, 994–1001, https://doi.org/10.1002/jps.23851 (2014).
16. Scotcher, D., Jones, C., Rostami-Hodjegan, A. & Galetin, A. Novel minimal physiologically-based model for the prediction of passive tubular reabsorption and renal excretion clearance. *Eur J Pharm Sci* **94**, 59–71, https://doi.org/10.1016/j.ejps.2016.03.018 (2016).
17. Liu, D. *et al*. A unified strategy in selection of the best allometric scaling methods to predict human clearance based on drug disposition pathway. *Xenobiotica* **46**, 1105–1111, https://doi.org/10.1080/00498254.2016.1205761 (2016).
18. Paine, S. W., Menochet, K., Denton, R., McGinnity, D. F. & Riley, R. J. Prediction of human renal clearance from preclinical species for a diverse set of drugs that exhibit both active secretion and net reabsorption. *Drug Metab Dispos* **39**, 1008–1013, https://doi.org/10.1124/dmd.110.037267 (2011).
19. Huang, W. & Isoherranen, N. Development of a Dynamic Physiologically Based Mechanistic Kidney Model to Predict Renal Clearance. *CPT Pharmacometrics Syst Pharmacol* **7**, 593–602, https://doi.org/10.1002/psp4.12321 (2018).
20. Dave, R. A. & Morris, M. E. Quantitative structure-pharmacokinetic relationships for the prediction of renal clearance in humans. *Drug Metab Dispos* **43**, 73–81, https://doi.org/10.1124/dmd.114.059857 (2015).
21. ChemAxon. *Marvin: A full featured chemical editor for making science accessible on all platforms*, https://chemaxon.com/products/marvin
22. Watanabe, R. *et al.* Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol Pharm* **15**, 5302–5311, https://doi.org/10.1021/acs.molpharmaceut.8b00785 (2018).
23. Bohnert, T. & Gan, L. S. Plasma protein binding: from discovery to development. *J Pharm Sci* **102**, 2953–2994, https://doi.org/10.1002/jps.23614 (2013).
24. Benet, L. Z., Broccatelli, F. & Oprea, T. I. BDDCS applied to over 900 drugs. *AAPS J* **13**, 519–547, https://doi.org/10.1208/s12248-011-9290-9 (2011).
25. Elsevier. *PharmaPendium: Fully searchable drug approval documents and extracted data to inform critical drug development decisions*, https://www.elsevier.com/
26. Varma, M. V. *et al*. Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. *J Med Chem* **53**, 1098–1108, https://doi.org/10.1021/jm901371v (2010).

27. Hosey, M. C., Chan, R. & Benet, Z. L. BDDCS Predictions, Self-Correcting Aspects of BDDCS Assignments, BDDCS Assignment Corrections, and Classification for more than 175 Additional Drugs. *AAPS J.* **18**, 251–260, https://doi.org/10.1208/s12248-015-9845-2 (2016).
28. Esaki, T. *et al*. Data curation can improve the prediction accuracy of metabolic intrinsic clearance. *Mol. Inf.* **37**, 1800086 (2018).
29. Tucker, G. T. Measurement of the renal clearance of drugs. *Br J Clin Pharmacol* **12**, 761–770, https://doi.org/10.1111/j.1365-2125.1981.tb01304.x (1981).
30. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J Cheminform* **10**, 4, https://doi.org/10.1186/s13321-018-0258-y (2018).
31. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**, 1466–1474, https://doi.org/10.1002/jcc.21707 (2011).
32. R Core Team. *R: A language and environment for statistical computing.*, https://www.R-project.org/ (2016).
33. Wickham, H. ggplot2: Elegant Graphics for Data Analysis in *Use R!* (Springer,, Switzerland, 2016).
34. Yuan Tang, M. H. and Wenxuan L. ggfortify: Unified Interface to Visualize Statistical Results of Popular R Packages. *R J 8.2*, 478–489 (2016).
35. Kuhn, M. Building predictive models in R using the caret package. *J Stat Softw.* **28**, 1–26 (2008).
36. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J Stat Softw* **36**, 1–13 (2010).
37. Liaw., A. & Wiener., M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
38. Meyer, D. *et al*. LIBSVM: a library for support vector machines, https://cran.r-project.org/web/packages/e1071/index.html (2001).
39. Ripley, B. & Venables, W. nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models, http://www.stats.ox.ac.uk/pub/MASS4/ (2016).
40. Varma, M. V., Steyn, S. J., Allerton, C. & El-Kattan, A. F. Predicting Clearance Mechanism in Drug Discovery: Extended Clearance Classification System (ECCS). *Pharm Res* **32**, 3785–3802, https://doi.org/10.1007/s11095-015-1749-4 (2015).
41. El-Kattan, A. F. *et al*. Projecting ADME Behavior and Drug-Drug Interactions in Early Discovery and Development: Application of the Extended Clearance Classification System. *Pharm Res* **33**, 3021–3030, https://doi.org/10.1007/s11095-016-2024-z (2016).
42. Varma, M. V., Pang, K. S., Isoherranen, N. & Zhao, P. Dealing with the complex drug-drug interactions: towards mechanistic models. *Biopharm Drug Dispos* **36**, 71–92, https://doi.org/10.1002/bdd.1934 (2015).
43. Tsujimoto, M. *et al*. Effects of decreased vitamin D and accumulated uremic toxin on human CYP3A4 activity in patients with end-stage renal disease. *Toxins (Basel)* **5**, 1475–1485, https://doi.org/10.3390/toxins5081475 (2013).
44. Yeung, C. K., Shen, D. D., Thummel, K. E. & Himmelfarb, J. Effects of chronic kidney disease and uremia on hepatic drug metabolism and transport. *Kidney Int* **85**, 522–528, https://doi.org/10.1038/ki.2013.399 (2014).
45. Ladda, M. A. & Goralski, K. B. The Effects of CKD on Cytochrome P450-Mediated Drug Metabolism. *Adv Chronic Kidney Dis* **23**, 67–75, https://doi.org/10.1053/j.ackd.2015.10.002 (2016).

Acknowledgements

Author contributions

Reiko Watanabe designed the study, wrote the manuscript, collected and curated data, and analysed data. Rikiya Ohashi contributed to design the study, interpretation of data, and assisted in the preparation of the manuscript. All other authors have contributed to data interpretation and review the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55325-1.

Correspondence and requests for materials should be addressed to R.W. or R.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.