Quantitative Structure – Pharmacokinetics Relationships Analysis of Basic Drugs: Volume of Distribution

Zvetanka Zhivkova , Tsvetelina Mandova, Irini Doytchinova

Faculty of Pharmacy, Medical University of Sofia, Sofia 1000, Bulgaria

Received, June 1, 2015; Revised, July 16, 2015; Accepted, September 23, 2015; Published, October 10, 2015

ABSTRACT - Purpose. The early prediction of pharmacokinetic behavior is of paramount importance for saving time and resources and for increasing the success of new drug candidates. The steady-state volume of distribution (VD_{ss}) is one of the key pharmacokinetic parameters required for the design of a suitable dosage regimen. The aim of the study is to propose a quantitative structure – pharmacokinetics relationships (QSPkR) for VD_{ss} of basic drugs. **Methods:** The data set consists of 216 basic drugs, divided to a modeling (n = 180) and external validation set (n = 36). 179 structural and physicochemical descriptors are calculated using validated commercial software. Genetic algorithm, stepwise regression and multiple linear regression are applied for variable selection and model development. The models are validated by internal and external test sets. **Results:** A number of significant QSPkRs are developed. The most frequently emerged descriptors are used to derive the final consensus model for VD_{ss} with good explanatory (r^2 0.663) and predictive ability ($q^2_{LOO\text{-}CV}$ 0.606 and r^2_{pred} 0.593). The model reveals clear structural features determining VD_{ss} of basic drugs which are summarized in a short list of criteria for rapid discrimination between drugs with a large and small VD_{ss}. **Conclusions**: Descriptors like lipophilicity, fraction ionized as a base at pH 7.4, number of cycles and fused aromatic rings, presence of Cl and F atoms contribute positively to VD_{ss}, while polarity and presence of strong electrophiles have a negative effect.

INTRODUCTION

The progress of computer-aided drug design techniques has led to an extensively increasing number of structures with drug-like properties and activities. Unfortunately, only few of them pass successfully though all stages of drug development and become drugs. In the past, one of the main reasons for drug failure was the unfavorable pharmacokinetic behavior (absorption, distribution, metabolism or excretion – ADME) (1). The understanding for the importance of pharmacokinetics inspired an intense research focused on the early prediction of the ADME properties of drug candidates before the expensive preclinical and clinical studies. As a result, the drug failure due to pharmacokinetics and bioavailability problems has fallen markedly from 40% in 1991 to 10% in 2000 (2).

One of the most reliable and widely used approaches for ADME prediction is the computational (*in silico*) modeling. It enables construction of quantitative structure – pharmacokinetics relationships (QSPkRs) based on molecular descriptors. The QSPkR models allow prediction of ADME properties even of virtual compounds, accelerate the identification of new drug candidates and reduce the cost of drug development process.

The volume of distribution VD is important pharmacokinetic parameter relating the amount of the drug in the body A to its plasma concentration, C:

$$V = A/C \qquad (1)$$

It has been defined as a hypothetical volume of body fluid that would be required to dissolve the total amount of drug at the same concentration as

Correspondence Author: Dr. Zvetanka Zhivkova, Faculty of Pharmacy, Medical University of Sofia, 2 Dunav St., 1000 Sofia, Bulgaria. Tel.: +359 2 9236514, Email: zzhivkova@pharmfac.acad.bg

that found in plasma (3). Three types of volume of distribution are classically reported in the literature: VD of the central compartment (VD_c), VD during the terminal phase (VD_β or VD_{area}) and VD at steady state (VD_{ss}). They differ in the times of sampling: just after *iv* administration, during the terminal phase of drug disposition, or after reaching of steady state, respectively (4). VD_{ss} is considered as the most reliable indicator for drug distribution in the body (5). It determines the half-life of the drug and serves as a key parameter for setting up a suitable dosage regimen (4, 6).

VD_{ss} is the most frequently predicted ADME parameter and a good number of QSPkR models have been published in the last two decades. They differ in the size and in the content of the datasets, the descriptors used, and the methods for model derivation and validation. A few studies concern congeneric series of drugs (7 – 10). In general, the QSPkRs proposed on congeneric series have a higher predictive power as a similar distribution behavior is expected. However, these models are local models, valid only within the studied series, while construction of a global model requires a large dataset encompassing diverse chemical spaces. The earlier models on diverse datasets are based on inconsistent data collected from literature, including different types of distribution volume (VD_c, VD_β or VD_{ss}), following different routes of administration (11 – 17). In 2008 Obach et al. (5) published the largest and best curated database so far containing the major pharmacokinetic parameters of 670 drugs, including VD_{ss} after *iv* administration. This database was used for the development of several successful models for VD_{ss} prediction (18 – 20).

A wide diversity of descriptors is used in the models for VD prediction, like lipophilicity of drugs, ionization state parameters, constitutional, topological, electrotopological, chemical, geometrical, quantum chemical descriptors (8 – 20), fraction bound to plasma proteins (7, 14, 15, 17), VD_{ss} in rat and dog (12). The models are derived by different statistical and machine learning methods as artificial neural networks (ANN) (7, 9, 13), multiple linear regression (MLR) (8, 10, 12, 14, 15, 18, 19), partial least squares (PLS) (10 – 12, 16, 18, 20), Bayesian neural networks (BNN) (16), classification and regression trees (CART) (16), mixed determinant analysis – random forest (MDA – RF) (17), recursive partitioning classification (RPC) (20).

Despite of the huge number of descriptors used in the QSPkR models for VD prediction, most of them contain mainly parameters, characterizing drugs lipophilicity (logP, logD and water solubility at different pH values, etc.) and the ionization state of the molecules (pK_a of the base, fraction ionized or non-ionized as base or as acid, etc.). A good agreement exists on the fact that more lipophilic drugs have larger VD_{ss}s (7, 11, 14, 15, 17, 18). The fraction ionized as a base at pH 7.4 also increases VD_{ss}, while the fraction ionized as an acid has a negative impact (14, 15, 17). All considered descriptors discriminate between acids and bases however there is no information about the structural features affecting VD_{ss}. According to the models, acids are expected to have small VD_{ss} and bases – large ones, which is consistent with the observations. VD_{ss} reflects the drug ability to cross membranes and to bind in tissues. The bases have high affinity to membrane phospholipids due to interactions between the drug cationic centers and the phospholipid acidic groups. The basic drugs bind to plasma alpha-1-acid glycoprotein and albumin with moderate to strong affinity depending on lipophilicity and also are accumulated by ion-trapping into lysosomes. Therefore, bases indeed have extensive VD_{ss}s (21). Acids have high affinity for albumin and the high albumin concentration in plasma results in a high plasma protein binding. The ionization at physiological pH 7.4 prevents their distribution in tissues and in general, acids have small VD_{ss}s.

Obviously, acids and bases follow different distribution patterns and it is reasonably to construct separate QSPkR models for acids and bases in order to identify the main structural features governing the value of VD_{ss}. There are only few reports on separate QSPkR modeling of VD_{ss} of bases and acids (14, 15, 19). In the studies of Ghafourian et al. (14, 15) the separate models show lower predictive ability as compared to the model on the whole dataset – mainly due to the limited number of drugs involved in the study. Recently, we developed robust, predictive and easy interpretable models for VD_{ss} of 132 acidic drugs from Obach's database (5), which revealed the main structural features affecting the distribution of acidic drugs in the body (19). The present study is focused on the relationship between the structure of basic drugs and their VD_{ss}s.

METHODS

Datasets

The whole dataset used in the present study comprised of 216 basic drugs belonging to different chemical and therapeutic classes. It was collected from Obach's database presenting data for the main pharmacokinetic parameters of 670 drugs following *iv* administration in human (5). In our previous study (19) we classified the molecules as acids, bases, neutral and zwitterions on the basis of their extent of ionization at the physiological pH 7.4. The fractions of a drug ionized as an acid (f_A) and as a base (f_B) were calculated according to the equations:

$$f_A = \frac{1}{1+10^{(pK_a - 7.4)}} \quad (2)$$

$$f_B = \frac{1}{1+10^{(7.4 - pK_a)}} \quad (3)$$

The mol-files of the drugs were derived and verified from several public databases (DrugBank (22), Chemical Book (23), Japan Chemical Substance Dictionary (24) and ChEBI (25). The pK_a values were calculated by ACD/LogD version 9.08 software (Advanced Chemistry Development Inc., Ontario, Canada). In case of multiple acidic/basic centers, the pK_a of the strongest one was considered. A drug was defined as a base, if $f_B > 0.02$ and $f_A = 0$.

The whole dataset was divided randomly into modeling and external validation set. To this end the molecules were arranged in an ascending order according to their VD_{ss} and one of every six drugs was allocated to a different subset. Thus, six subsets, each comprising 36 drugs, were generated. One of the subsets (randomly) was excluded as an external validation set and later was used for assessment of the predictive ability of the final model. The remaining five subsets composed the modeling set. In turn, each subset in the modeling set was used once as a test set for the model, developed on the training set, consisting of the remaining 4 subsets (leave-group-out validation). In summary, five training sets, five test sets and one external validation set were used in the study (Table 1). The experimental VD_{ss} values were logarithmically transformed in order to get close to a normal distribution.

Table 1. Training, test and external validation sets used in the study.

Training set	Subsets included	Test set
A	2 + 4 + 5 + 6	1
B	1 + 4 + 5 + 6	2
C	1 + 2 + 5 + 6	4
D	1 + 2 + 4 + 6	5
E	1 + 2 + 4 + 5	6
External validation set	3	-

Molecular Descriptors and Variable Selection

The chemical structures were described by 179 molecular descriptors calculated by ACD/LogD version 9.08 (Advanced Chemical Development, Inc.) and MDL QSAR version 2.2 (MDL Information Systems Inc, San Leandro, CA). The descriptors included electrotopological, molecular connectivity and kappa shape indices (26, 27), descriptive properties (number of specific atoms and groups, rings, circles, hydrogen-bond donors and acceptors, etc.), molecular 2D (molecular weight, logP, $logD_{7.4}$, PSA, etc.) and 3D properties (dipole moment, volume, surface, etc.).

A three-step variable selection procedure was applied to identify the most significant predictors. Initially, for every training set, descriptors with non-zero values for less than three molecules were eliminated. Next, descriptors were selected by genetic algorithm (GA) (28). Finally, the selected descriptors (usually less than 15) entered a forward stepwise linear regression with F-to-enter 4.00 and F-to-remove 3.99. Both genetic and stepwise regression algorithms were used as implemented in the MDL QSAR package.

Generation of QSPkR Models

The QSPkR models were developed by multiple linear regression (MLR) technique. Using different combinations of descriptors, a number of QSPkR models were constructed for each training set. Drugs which $logVD_{ss}$ values were predicted with residuals not obeying the normal distribution law were considered as outliers. They were removed from the training sets and the models were rebuilt. The models were primarily assessed by explained variance r^2 given by the equation:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(\log VD_{ss_{obs,i}} - \log VD_{ss_{calc,i}}\right)^2}{\sum_{i=1}^{n}\left(\log VD_{ss_{obs,i}} - \log VD_{ss_{obs,mean}}\right)^2} \quad (4)$$

where $VD_{ss_{obs,i}}$ and $VD_{ss_{calc,i}}$ are the observed and the calculated by the model values of VD_{ss} for the i^{th} drug and $\log VD_{ss_{obs,mean}}$ – the mean observed $\log VD_{ss}$ value for the set. Only models with $r^2 > 0.6$ were subjected to validation. The most significant descriptors involved in the best models were further used for development of a consensus QSPkR model for $\log VD_{ss}$ prediction.

Validation of the Models

The generated QSPkR models were validated by randomization test, leave-one-out cross-validation (LOO-CV) and leave-group-out validation. The model performance was assessed by cross-validated coefficient $q^2_{LOO\text{-}CV}$, prediction coefficient r^2_{pred} for the test set, mean fold error of prediction MFEP and accuracy:

$$q^2_{LOO-CV} = 1 - \frac{\sum_{i=1}^{n}\left(\log VD_{ss_{obs,i}} - \log VD_{ss_{pred,i}}\right)^2}{\sum_{i=1}^{n}\left(\log VD_{ss_{obs,i}} - \log VD_{ss_{obs,mean}}\right)^2} \quad (5)$$

$$r^2_{pred} = 1 - \frac{\sum_{i=1}^{n}\left(\log VD^{ss}_{obs,i} - \log VD^{ss}_{pred,i}\right)^2_{test}}{\sum_{i=1}^{n}\left(\log VD^{ss}_{obs,i} - \log VD^{ss}_{obs,mean}\right)^2_{test}} \quad (6)$$

Where $VD_{ss_{obs,i}}$ and $VD_{ss_{pred,i}}$ are the observed and the predicted by the model values of VD_{ss} for the i^{th} drug in the training set or in the test set, and $\log VD_{ss_{obs,mean}}$ – the mean observed $\log VD_{ss}$ value.

$$MFEP = \frac{10^{\left|\log VD_{ss_{obs,i}} - \log VD_{ss_{pred,i}}\right|}}{n} \quad (7)$$

Accuracy of prediction was assessed as a percent of drugs with VD_{ss} predicted with less than two- or three-fold error.

RESULTS

Analysis of the Datasets

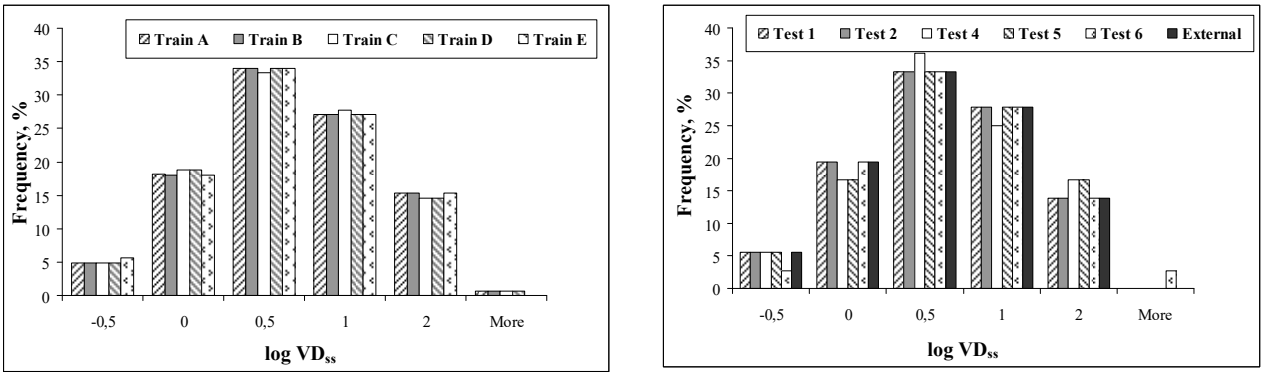
The dataset used in the present study consisted of 216 basic drugs with diverse chemical structure and therapeutic usage. The structures covered a broad chemical space: the molecular weight was in the range 129 – 1431 g/mol (mean 360, median 324), logP varied between -5.8 and 8.9 (mean 2.53, median 2.83), and $logD_{7.4}$ – between -8.7 and 6.9 (mean 1.0, median 1.28). The fraction ionized as a base at the physiological pH 7.4 f_B ranged between 0.015 and 1.00 with 60% of the drugs almost completely ionized (with $f_B > 0.95$). The VD_{ss}s covered a wide interval from 0.073 to 140 L/kg (mean 6.08, median 2.5) and $\log VD_{ss}$s showed a normal distribution (mean 0.41, median 0.40). The unbound fractions f_u were available for 182 drugs and suggested moderate to high plasma protein binding with $f_u \leq 0.1$ (plasma protein binding exceeding 90%) for 32% of the drugs, and $f_u \geq 0.9$ (negligible plasma protein binding) for 8% of the drugs.

In order to develop robust and predictive QSPkR models, the whole dataset was divided into two subsets – an external validation set (36 drugs) and a modeling set (180 drugs). In turn, the modeling set was divided into five training and five test sets as described in **Methods**. The $\log VD_{ss}$s had normal distribution for all subsets (Figure 1).

QSPkR Models for $LogVD_{ss}$

Numerous significant models were generated on the five training sets using different initial combinations of descriptors. The models were validated as described in **Methods**. The best models for every training set in terms of explained variance r^2, cross-validation coefficient on the training set $q^2_{LOO\text{-}CV}$, prediction coefficient for the test set r^2_{pred}, mean fold error of prediction MFEP, and accuracy are given in Table 2.

Although the training sets differed of each other by 20% of the included drugs, the generated QSPkR models were very similar in terms of selected variables, statistics and outliers. The explained variance of the best models r^2 varied between 0.635 and 0.700 (mean 0.666). The values of $q^2_{LOO\text{-}CV}$ and r^2_{pred} ranging from 0.578 to 0.664 (mean 0.604) and from 0.444 to 0.602 (mean 0.538), respectively, were indicative for the good predictive ability of the models. The values of r^2_{rand}



a. b.

Figure 1. Histogram of $logVD_{ss}$ values: a. for the training sets; b. for the test and external validation sets

Table 2. QSPkR models developed on the five training sets and validated by the corresponding test sets

Training set	Model	r^2	q^2_{LOO-CV}	r^2_{pred}	MFEP	Accuracy %
A	$\log VD^{ss} = 0.117(\pm0.014)\log P + 0.627(\pm0.102) f_B + 0.056(\pm0.012)SsCl + 0.081(\pm0.036)SaaaC + 0.021(\pm0.004)SsF + 0.175(\pm0.031)G\,\min + 0.027(\pm0.007)ncirc - 0.556$	0.652	0.578	0.505	2.41	64
outliers from the training set: azythromycin, chloroquine, mibefradil, topixantrone, triamterene						
B	$\log VD^{ss} = 0.113(\pm0.013)\log P - 0.034(\pm0.011)Dipole + 0.523(\pm0.097)f_B + 0.033(\pm0.006)ncirc + 0.058(\pm0.012)SsCl + 0.241(\pm0.045)SsF\_acnt + 0.150(\pm0.029)G\,\min - 0.070(\pm0.012)SdsN - 0.292$	0.667	0.594	0.564	2.26	50
outliers from the training set: azythromycin, cloroquine, netilmicin, procyclidine, pyrimethamine, topixantrone, triamterene						
C	$\log VD^{ss} = 0.129(\pm0.015)\log P + 0.208(\pm0.034)G\,\min + 0.098(\pm0.022)xp9 + 0.648(\pm0.111)f_B + 0.248(\pm0.052)SsF\_acnt + 0.041(\pm0.013)SsCl + 0.077(\pm0.035)SaaaC - 0.589$	0.635	0.588	0.574	2.12	53
outliers from the training set: azythromycin, chloroquine, fentanyl, maprotiline, mibefradil, netilmicin, tolterodine; outliers from the test set: topixantrone, triamterene						
D	$\log VD^{ss} = 0.157(\pm0.013)\log P - 0.002(\pm0.0003)Volume + 19.06(\pm3.53)xch10 + 0.053(\pm0.011)SsCl + 0.494(\pm0.094)f_B - 0.071(\pm0.025)SssssC + 0.085(\pm0.029)SaaN\_acnt + 0.126$	0.700	0.664	0.444	2.25	58
outliers from the training set: azimilide, azythromycin, chloroquine, disopyramide, oxybutynin, procyclidine, tamsulosin, topixantrone, triamterene; outliers from the test set: pyrimethamine, repinotan, vinblastine						
E	$\log VD^{ss} = 0.098(\pm0.016)\log P + 34.78(\pm14.26)xvch10 + 0.677(\pm0.109)f_B + 0.387(\pm0.084)SsCl\_acnt - 0.052(\pm0.013)Dipole + 0.079(\pm0.035)SaaaC\_acnt + 0.139(\pm0.045)SddssS - 0.040(\pm0.011)SsssCH\_acnt + 0.145(\pm0.035)xvp8 - 0.317$	0.656	0.594	0.602	1.93	71
outliers from the training set: melperone, oxybutynin, procyclidine, topixantrone, triamterene, vinblastine; outliers from the test set: azythromycin, chloroquine						

Table 3. The most frequently emerging descriptors in the QSPkR models for $logVD_{ss}$

Descriptor	Encoded structural information	Effect on VD^{ss}	Frequency,%
logP, $logD_{7.4}$	Lipophilicity parameter	Positive	100
f_B	Fraction of the drug ionized as a base at *pH* 7.4	Positive	100
SsCl SsCl_acnt	Sum of all Cl E-state values, or count of all Cl atoms in the molecule	Positive	100
SaaaC SaaaC_acnt	Sum of all aaaC (aromatic C in fused rings) E-state values, or count of all aaaC groups in the molecule	Positive	59
SsF SsF_acnt	Sum of all F E-state values, or count of all F atoms in the molecule	Positive	53
xch10 xvch10	Simple or valence 10-order chain connectivity index (presence of 10-member ring system)	Positive	53
Dipole	Dipole moment of the molecule	Negative	53
SddssS SddssS_acnt	Sum of all ddssS (sulphonyl) E-state values, or count of all ddssS groups in the molecule	Negative	47
G_{min}	Minimum E-state value in the molecule	Positive	47
G_{max}	Maximum E-state value in the molecule	Negative	
ncirc	Number of cycles	Positive	29
SssssC SssssC_acnt	Sum of all ssssC (quaternary C) E-state values, or count of all ssssC atoms in the molecule	Positive	23
Surface, Volume		Negative	18
SdsN	Sum of all =N- E-state values	Negative	18
SaaN_acnt	Count of all aaN (aromatic N) groups	Positive	18
SsssCH_acnt	Count of all sssCH (tertiary C) groups in the molecule	Negative	12
xp9	Simple 9th order path connectivity index	Positive	12
SdssC	Sum of all dssC E-state values in the molecule	Positive	6
xvp8	Valence 8th order path connectivity index	Positive	6

(between 0.051 and 0.072, mean 0.057) suggested that no chance correlations were developed. No intercorrelation between the descriptors in the models was observed ($r < 0.65$). Several drugs were identified as outliers by almost all of the models (azythromycin, chloroquine, pyrimethamine, topixantrone, triamterene. The most frequently emerged descriptors in the best three models for each training set are listed in Table 3. For sets D and E four best models were taken as they had very close statistics.

The 27 most frequently emerging descriptors were used for development of the final QSPkR model on the whole modeling dataset comprising 180 basic drugs. Eight drugs were identified as outliers, and their removal resulted in the following consensus model:

Consensus model

$$\log VD_{\text{я}} = 0.124(\pm 0.013)\log P - 0.040(\pm 0.011)\text{Dipole} + 0.026(\pm 0.006)\text{ncirc} +$$
$$+ 0.565(\pm 0.090)f_B + 0.333(\pm 0.065)\text{SsCl\_acnt} + 0.254(\pm 0.046)\text{SsF\_acnt} +$$
$$+ 0.155(\pm 0.029)G\,\text{min} + 0.079(\pm 0.031)\text{SaaaC\_acnt} - 0.071(\pm 0.032)\text{SdssC} - 0.366$$

$n = 172$; $r^2 = 0.663$; $q^2_{LOO-CV} = 0.606$; $F = 35.43$; $r^2_{rand} = 0.048$

Outliers*:* azythromycin, chloroquine, mibefradil, netilmicin, procyclidine, pyrimethamine, topixantrone, triamterene

Evaluation of the Predictive Ability of the QSPkR Models

The predictive ability of the best proposed QSPkR models for $logVD_{ss}$ of basic drugs was assessed using the external validation set. The predictive statistics of the models is summarized in Table 4. The plot of predicted by the consensus model versus the experimental values of $logVD_{ss}$ for the external validation set is presented in Figure 2.

The values of r^2_{pred} ranged from 0.529 to 0.593 (mean 0.555) and MFEP was between 2.24 and 2.38 (mean 2.31). The predicted values for VD_{ss} were within the two fold error for 53% of the drugs (on average). As expected, the consensus model showed the best performance.

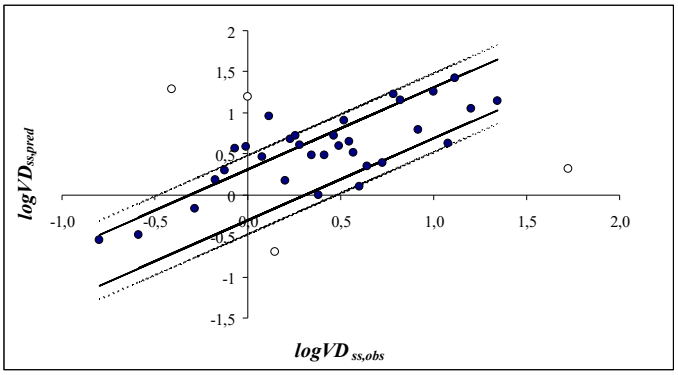


Figure 2. Predicted by the consensus model *vs*. experimental $logVD_{ss}$ values for the external validation set. The four outliers are shown as blank circles. The lines and the dotted lines represent the twofold and three fold error limits.

Table 4 Predictive statistics of the best QSPkR models for $logVD_{ss}$ of basic drugs evaluated by the external validation set (n = 36)

Model	r^2_{pred}	MFEP	Accuracy, %		Outliers in the external validation set
			FEP < 2	FEP < 3	
Model 1	0.536	2.38	53	78	cetrorelix, pentamidine, sildenafil, ziprasidone
Model 2	0.579	2.24	44	88	cetrorelix, pentamidine, sildenafil, ziprasidone
Model 3	0.540	2.30	59	75	cetrorelix, pentamidine, sildenafil, ziprasidone
Model 4	0.551	2.35	57	74	pentamidine
Model 5	0.529	2.37	52	76	cetrorelix, pentamidine, ziprasidone
Consensus model	0.593	2.25	50	88	cetrorelix, pentamidine, sildenafil, ziprasidone

Criteria for VD_{ss} Prediction of Basic Drugs

The descriptors involved in the consensus model were used to propose a number of criteria for prediction of VD_{ss} of basic drugs. To this end the drugs were classified into three groups: with small (< 0.7L/kg), moderate (between 0.7 and 2L/kg), and large VD_{ss} (> 2L/kg). A cutoff value for each descriptor could be defined in order to distinguish between drugs with small and large VD_{ss}. The cutoffs for large VD_{ss} are listed in Table 5.

For most descriptors, the number of molecules meeting these criteria in each group increased as VD_{ss} increased (Figure 3). However, this was not true for f_B and SdssC. Drugs with small and large VD_{ss} had high values for f_B. According to the consensus model, the descriptor SdssC had a negative contribution in VD_{ss}, i.e. it was expected that drugs with negative values for SdssC would have large VD_{ss}. However, the distribution of the molecules with SdssC < 0 followed the opposite trend.

The remaining seven criteria were applied to the studied dataset. The small VD_{ss} group comprised 27 drugs with low lipophilicity (56% with logP < 0), small number of cycles, negative

Table 5. Criteria for a large VD_{ss} of basic drugs

Descriptor	Cutoff for VD_{ss} > 2 L/kg
logP	> 3
f_B	> 0.95
Dipole	< 4
ncirc	> 4
G_{min}	> 0
SdssC	< 0
aaaC_acnt	> 0
Cl_acnt	> 0
F_acnt	> 0

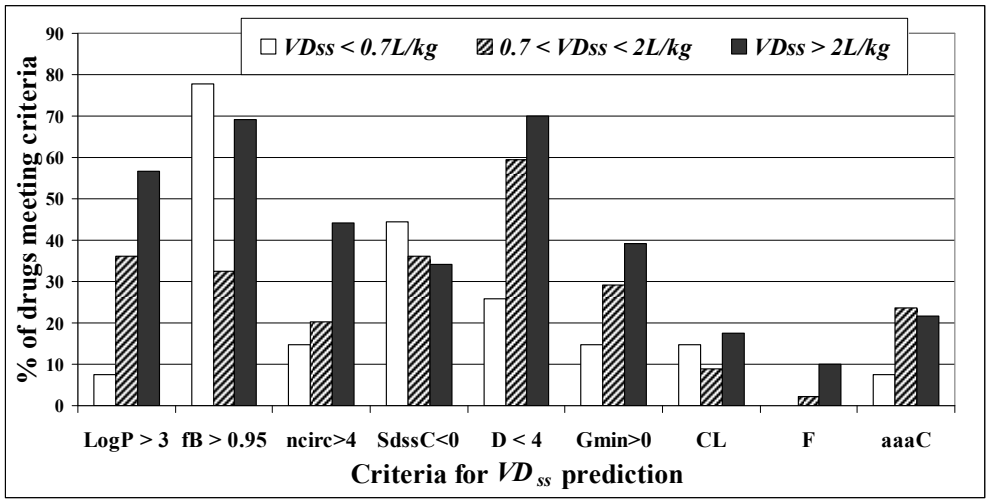


Figure 3. Percentage of drugs with small (blank), average (shaded) and large (black) VD_{ss} meeting the proposed criteria for VD_{ss} prediction

G_{min} for 93% of the drugs, high polarity (high Dipole), negligible content of Cl, F or aaaC atoms. The large VD_{ss} group consisted of 120 drugs. Most of them showed high lipophilicity (with average logP = 3.31), 44% of the structures contained more than 4 cycles, 40% of drugs had positive value of G_{min}, 68% were fairly polar with Dipole < 4, and almost the half contained one or more Cl, F or aaaC atoms. The distribution of the drugs with small, moderate and large VD^{ss} (in %) according to the number of met criteria is shown in Figure 4. None drug met all seven criteria.

It is evident that the criteria for large VD_{ss} defined in the present study are good enough to distinguish between drugs with small and large VD_{ss}. Sixty three percent of the drugs with small VD_{ss} meet neither criterion for large VD_{ss}. At the other extreme, 52% of the drugs with large VD_{ss} fulfill at least three criteria. Therefore, basic drugs meeting at least three of the following criteria: logP > 3, ncirc > 4, G_{min} > 0, Dipole < 4, presence of Cl, F or/and aaaC atom are expected to have VD_{ss} > 2 L/kg. Oppositely, molecules which meet neither criterion should have VD_{ss} < 0.7 L/kg.

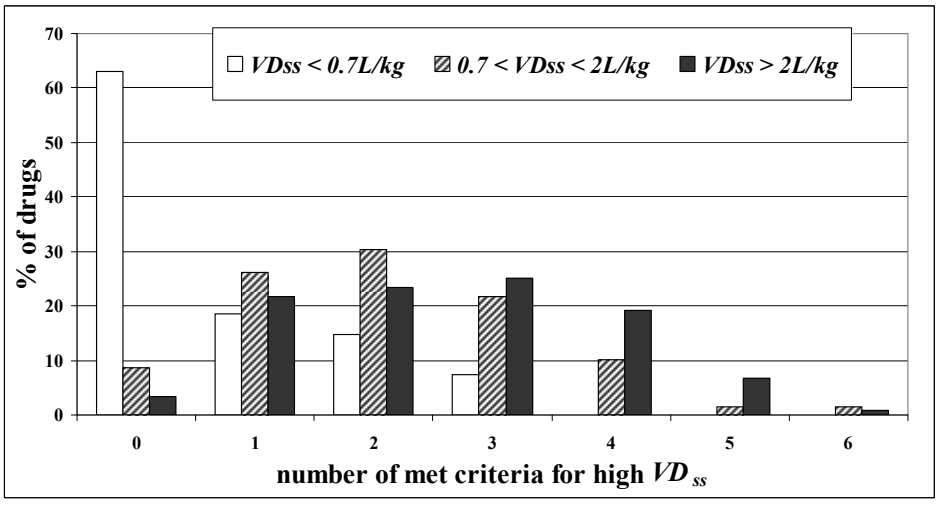


Figure 4. Distribution of the drugs (in %) according to the number of met criteria for high VD_{ss}

DISCUSSION

In general, the QSPkR models differ from the classical QSAR models in many aspects (29 – 32). First, the QSPkR models operate with *in vivo* data, collected from different labs, often using a wide variety of assay conditions. Next, the sets of compounds used in ADME prediction consist of structurally diverse molecules, having diverse pharmacokinetic and pharmacodynamic behavior. Finally, there is a trade-off between simple, ease to interpret models with lower predictivity and more complex "black box" models with better predictive ability. Despite these challenges, there has been no respite in the development of newer and better *in silico* models for ADME prediction.

The present study is focused on the development of QSPkR models for VD_{ss} of basic drugs. A dataset consisted of 216 molecules covering a wide chemical and biological space. The chemical structures were described with 179 descriptors. The VD_{ss}s were transformed to $logVD^{ss}$ in order to approach a normal distribution. A three-step variable selection was applied and a number of QSPkR models were proposed using MLR. In order to obtain robust models with high predictive ability, a rigorous validation procedure was applied, as recommended by Tropsha et al. (33). To this end the dataset was separated into six subsets of 36 drugs each. One of the subset was defined as an external validation set, the remaining five – as a modeling set. In turn, the modeling set was divided into training and test sets in a ratio 4:1 in five different combinations. The QSPkR models were developed on the training sets and validated by the test sets. The most frequently emerged descriptors entered a step-wise selection and a consensus QSPkR model was developed. All models were evaluated by the external validation set and showed very good predictive ability (r^2_{pred} in the range 0.529 – 0.579; MFEP between 2.24 and 2.38). As expected, the consensus model performed best: r^2 0.663, $q^2_{LOO\text{-}CV}$ 0.606, r^2_{pred} 0.593, MFEP 2.25. The values of both $q^2_{LOO\text{-}CV}$ and r^2_{pred} exceeded the value of 0.5 accepted as a threshold for predictive models in QSAR (34).

The final consensus model contains descriptors with clear physical sense. It reveals the most important structural features determining the value of VD_{ss} for basic drugs. The descriptor logP, encoding the lipophilicity of the molecule, appears to be the most significant determinant of VD_{ss}. It is responsible for about 50% of the explained variance. There are 20 drugs with negative logP in the dataset. Seventeen of them have small VD_{ss} (< 0.7L/kg). Oppositely, 10 of the 14 most lipophilic drugs (with logP > 5) have large VD_{ss} (> 2L/kg). The positive effect of logP on VD_{ss} has long been recognized (35). This is not surprising as a good lipophilicity is required for many processes involved in drug distribution: membrane permeability, binding to tissue components, accumulation in mast cells, etc.

The descriptor f_B indicates the fraction ionized as a base at pH 7.4. Drugs with high f_B values have large VD^{ss}. The presence of a strong basic center enables the ion-pair interactions with the charged acidic head groups of membrane phospholipids, the binding to phosphatidylserine in the cell membranes in several tissues and the ion trapping in lysosomes (21). The descriptor Dipole represents the dipole moment of the molecule – a measure of polarity. According to the consensus model, it has a negative contribution in VD_{ss}. This means that polar drugs should have small VD_{ss}s, as it is observed.

The molecular descriptor ncirc is equal to the total number of cycles in the molecular graph. One cycle can be counted several times if it is fused with another cycles. For example, for biphenyl ncirc = 2, while for naphthalene ncirc = 3 (two cycles with 6 edges and one common cycle with 10 edges). At equal composition, molecules with higher value of ncirc have lower volume and surface. Obviously, ncirc reflects the compactness of the molecule – higher compactness is favorable for both membrane permeation and tissue binding. The positive contribution to VD_{ss}, according to the consensus model, means that the more compact drugs should have larger VD_{ss}.

G_{min} coincides with the lowest E-state value in the molecule. The E-state value provides information about the electron accessibility to the atom. Terminal electronegative atoms are easily accessible and have higher E-state values, while atoms connected with electronegative ones (strong electrophiles) have lower values (26). The positive correlation between G_{min} and VD_{ss} means that drugs containing electrophile groups (CF_3, SO_2, CO, etc.) have small VD_{ss}.

The descriptor SdssC encodes the presence and electronic state of a carbon atom type –C=

(represented in the dataset as >C=O, >C=C< or >C=N–). Depending on the substituent, it can take positive as well as negative values. Molecules with many electronegative atoms and groups have negative values, while the prevalence of aromatic and aliphatic moieties results in positive ones. According to the consensus model, drugs with negative SdssC have large VD_{ss}.

The presence of Cl and F atoms contributes to the lipophilicity of the molecule, and also they might serve as hydrogen-bond acceptors. The atom type aaaC represents an aromatic C-atom connected with three aromatic C-atoms, i.e. belonging to two fused aromatic rings. The presence of such atoms increases the VD_{ss}.

The clear physical sense of the descriptors involved in the consensus model allowed us to define a list of criteria for discrimination between drugs with small and large VD_{ss}s. Values for logP higher than 3, more than 4 circles in the molecule, positive G_{min} values, dipole moments up to 4 D and the presence of Cl, F or/and fused aromatic rings are prerequisites for large VD_{ss}. Drugs which meet neither criterion are expected to have small VD_{ss} (< 0.7L/kg), while those meeting three or more criteria have VD_{ss} > 2L/kg. Applying these criteria to drugs with small (27 drugs) and large VD_{ss} (120 drugs) from the tested dataset, only six of them were mispredicted: two were overestimated and four were underestimated.

Eight drugs were identified as outliers of the consensus model from the modeling set and another four – from the external evaluation set (Table 6). Their incompatibility with the model could be due to several reasons: unique structural features, unusual distribution patterns, errors in molecule presentation or in the VD_{ss} values.

Table 6 Outliers from the consensus model.

Outlier	logP	$VD_{ss,obs}$, L/kg	$VD_{ss,pred}$, L/kg	Number of met criteria
netilmicin	-1.9	0.073	0.42 ↑	0
cetrorelix	2.62	0.39	19.2 ↑	3
pyrimethamine	2.75	0.43	2.67 ↑	2
procyclidine	3.93	0.74	4.26 ↑	2
ziprasidone	4	1	15.5 ↑	6
sildenafil	2.28	1.4	0.20 ↓	1
mibefradil	6.29	3.1	16.5 ↑	4
triamterene	0.18	13	0.64 ↓	3
azithromycin	3.33	33	2.14 ↓	2
pentamidine	2.47	53	4.52 ↓	1
topixantrone	1.31	57	4.51 ↓	3
chloroquine	3.69	140	9.5 ↓	4

The search in the literature showed that netilmicin and pyrimethamine are not real outliers. The VD_{ss} value of netilmicin in the Obach's database (5) is 0.073 L/kg. However, the value in the original reference is 0.2 – 0.3 L/kg (36) which is close to our prediction of 0.42 L/kg. An even higher value of 0.68 L/kg was announced by Wenk et al. (37). Similarly, the VD_{ss} value of pyrimethamine in the Obach's database is 0.43 L/kg, while others report for VD_{ss} ranging between 2.12 and 3.06 L/kg (38). Our predicted value of 2.67 L/kg falls in this range. Cetrorelix is highly overpredicted by the model, due to the presence of three positive criteria (a large number of cycles, presence of Cl and aaaC atoms). This drug has an extremely high molecular weight of 1431 g/mol which embarrasses the membrane permeability and localizes the drug in plasma where it is 86% bound to plasma proteins (5). The experimental VD_{ss} value of procyclidine is 0.74 L/kg (39) and seems unlikely small considering drug's high lipophilicity (logP 3.93). Ziprasidone meets all 6 criteria for large VD_{ss}, but its observed value is only 1 L/kg. It is extensively metabolized in liver (40) and is almost completely

bound to plasma proteins (f_u 0.0012) (5), which locate it mainly in the central compartment. The presence of a sulfonyl group in the molecule of sildenafil results in a high negative value of G_{min}, which in combination with the small ionized fraction (f_B = 0.04) predicts 0.2 L/kg VD_{ss} instead of the observed 1.4 L/kg. Mibefradil is correctly predicted as a drug with large VD_{ss}, and the large difference between the observed and predicted values could be explained by the extremely high lipophilicity (logP 6.29). The last five outliers have extremely large VD_{ss} (> 10 L/kg) implying considerable tissue accumulation and unique distribution patterns not captured by the models. Triamterene meets three criteria for large VD_{ss} but its low lipophilicity (logP 0.18) results in a small predicted value. An extensive binding to tissues in the central compartment was observed in rat leading to very slow elimination (41). This is consistent with the extensive hepatic metabolism and biliary excretion of the drug (42). Azythromycin meets only two criteria for large VD_{ss} and is underpredicted by the model. The presence of two basic centers in the molecule was considered as the main factor for its extremely large VD_{ss} (21). Additionally, there is a plenty of hydrogen-bond donors and acceptors in the molecule involved in tissue binding. Extensive uptake and slow release from tissues was suggested as the main reasons for the long half-life of the drug (43). High concentrations were observed in prostate, tonsils and other tissues (44). Pentamidine meets only one criterion; however it is moderately lipophilic and also has two equivalent basic centers. A high accumulation of the in rat liver lysosomes was reported for this drug (45). As a diamine substance, pentamidine is a substrate of the organic cation transporters facilitating the high distribution in kidneys, liver and bile (46). Topixantrone meets three criteria but is also underpredicted – probably due to the unfavorable low logP. The presence of two basic centers presupposes specific interactions in the cell membranes or tissues. A prominent affinity of the drug for DNA has been also suggested (47). Chloroquine is the drug with the largest VD_{ss} in the dataset and it is predicted as a large VD_{ss} drug. The great difference between the observed and predicted values makes it outlier. Ion trapping was suggested as the main factor for chloroquine accumulation in tissues (48). Very high concentrations were observed in rat kidneys, liver, spleen and lung with a tissue to blood ratio close to 300 (49). A remarkable affinity for melanin in skin and eye (mediated through a charge transfer process) and slow release from the pigmented tissues was suggested (50).

CONCLUSIONS

The present study presents a set of statistically significant, predictive and interpretable QSPkR models for VD_{ss} of basic drugs. The best of them, the consensus model, allows the prediction of 50% of the drugs in an external test set with less than 2-fold error, and 88% – with less than 3-fold error. The descriptors involved in the model reveal clear structural features determining the distribution of basic drugs. The lipophilicity, the ionization at pH 7.4, the presence of fused rings, Cl and F atoms contribute positively to VD_{ss}, while the polarity of the molecule and the presence of strong electrophiles have a negative effect. A list of criteria is proposed for discrimination between drugs with small and large VD_{ss}.

REFERENCES

1. Van de Waterbeemd H, Gifford E. ADMET in silico modeling: towards prediction paradise? Nature Reviews/Drug Discovery, 2003; 2:192-204.
2. Kola L, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Reviews/Drug Discovery, 2004; 3:711-715.
3. Dominguez R. Studies of renal excretion of creatinine. II. Volume of distribution. Proc Soc Exp Biol Med, 1934**;** 31:1146-1150.
4. Toutain PL, Bousquet – Melou A. Volumes of distribution. J Vet Pharmacol Therap*,* 2004; 27:441-453.
5. Obach RS, Lombardo F, Waters NJ. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos, 2008; 36(7):1385-1405. `
6. Rowland M, Tozer TN, Multiple-dose regimens, in Rowland M; Tozer TN (eds), Clinical pharmacokinetics and pharmacodynamics: concepts and applications, 4th ed, Baltimore Maryland, Lippincott Williams & Wilkins, pp 293-329, 2010.
7. Goburru J, Shelver W. Quantitative structure – pharmacokinetic relationships of beta blockers derived using neural networks. J Pharm Sci*,* 1995; 84:862-865.
8. Mager D, Jusko W. Quantitative structure – pharmacokinetic/pharmacodynamic relationships of

corticosteroids in man. J Pharm Sci, 2002; 91:3441-3451.
9. Turner JV, Maddalena DJ, Cutler DJ, Agatonovich – Kustrin S. Multiple pharmacokinetic prediction for a series of cephalosporins. J Pharm Sci 2003; 92:552-559.
10. Karalis,V, Tsantili – Kakoulidou A, Macheras P. Quantitative structure – pharmacokinetic relationships for dispositions parameters of cephalosporins. Eur J Pharm Sci, 2003; 20:115-123.
11. Chee NN, Xiao Y, Putnam W, Lum B, Tropsha A. Quantitative structure – pharmacokinetic parameters relationships analysis of antimicrobial agents in humans using simulated annealing K-nearest neighbor and PLS analysis methods. J Pharm Sci, 2004; 93(10):2535-2544.
12. Wajima T, Fukumura K, Yano Y, Oguma T. Prediction of human pharmacokinetics from animal data and molecular structural parameters using multivariate regression analysis: volume of distribution at steady state. J Pharm Pharmacol, 2003; 55:939-949.
13. Turner JV, Maddalena DJ, Cutler DJ. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. Int J Pharmac, 2004; 270:209-219.
14. Ghafourian T, Barzegar – Jalali M, Hakimiha N, Cronin MTD. Quantitative structure – pharmacokinetic relationship modeling: apparent model of distribution. J Pharm Pharmacol, 2004; 56:339-350.
15. Ghafourian T, Barzegar – Jalali M, Dastmalchi S, Khavari – Khorasani T, Hakimiha N, Nokhodchi A. QSPR models for the prediction of apparent volume of distribution. Int J Pharmac, 2006; 319: 82-97.
16. Gleeson MP, Waters NJ, Paine SW, Davis A. M. In silico human and rat V_{ss} quantitative structure – activity relationships models. J Med Chem, 2006; 49 (6):1953-1963.
17. Lombardo F, Obach RS, DiCapua FM, Bakken GA, Lu J, Potter DM, Gao F, Miller MD, Zhang Y. Hibrid mixture discriminant analysis 9 random forest computational model for the prediction of volume of distribution of drugs in human. J Med Chem, 2006; 49(7):2262-2267.
18. Berellini G, Springer C, Waters NJ, Lombardo F. In silico prediction of volume of distribution in human using linear and nonlinear models on a 669 compound data set. J Med Chem, 2009; 52:4488-4495.
19. Zhivkova Z, Doytchinova I. Prediction of steady – state volume of distribution of acidic drugs by quantitative structure – pharmacokinetic relationships. J Pharm Sci, 2012; 101(3):1253-1266.
20. del Amo EM, Ghemtio L, Xhaard H, Yliperttula M, Urtti A, Kidron H. Applying linear and non-linear methods for parallel prediction of volume of distribution and fraction of inbound drug. PLoS ONE, 2013, 8(10), e74758; doi:10.1371/journal.pone 0074758
21. Smith DA, Allerton C, Kalgutkar AS, Waterbeemd H, Walker DK. Distribution, in Smith DA; Allerton C; Kalgutkar AS; Waterbeemd H; Walker DK (eds), Pharmacokinetics and metabolism in drug design, 3rd ed, Weinheim, Wiley – VCH Verlag GmbH&Co. KgaA, pp. 61-79, 2012.
22. http://www.drugbank.ca/ Accessed 30.04.2015
23. http://www.chemicalbook.com/Accessed 30.04.2015
24. http://nikkajiweb.jst.go.jp/ Accessed 30.04.2015
25. http://www.ebi.ac.uk Accessed 30.04.2015
26. Hall LH, Kier LB, Electrotopological state. Structure modelling for QSAR and database analysis, in Devillers J; Balaban A (eds), Topological indices and related descriptors in QSAR and QSPR, London, Gordon and Breach, pp 491-562, 1999.
27. Hall LH, Kier LB, Molecular connectivity chi indices for database analysis and structure – property modelling, in Devillers J; Balaban A (eds), Topological indices and related descriptors in QSAR and QSPR, London, Gordon and Breach, pp 307-360, 1999.
28. Leardi R. Genetic algorithms in chemistry. J Chromatogr A, 2007; 1158:226-233.
29. Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweyko A, Li Y. In silico ADME/Tox: why models fail. J Comp Aided Mol Des, 2003; 17:83-92.
30. Yamashita F, Hashida M. In silico approaches for predicting ADME properties of drugs. Drug Metab Pharmacokin, 2004; 19:327-338.
31. Chohan KK, Paine SW, Waters NJ. Advancements in predictive in silico models for ADME. Curr Chem Biol, 2008; 2:215-228.
32. Gleeson PM, Hersey A, Hannongbua S. In silico ADME models: a general assessment of their utility in drug discovery applications. Curr Topics Med Chem, 2011; 11:358-381.
33. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inf, 2011; 29:476-488.
34. Roy K; Kar S; Das, RN, A primer on QSAR/QSPR modelling Fundamental concepts. Springer Cham, Heidelberg, New York, Dordrecht, London, 2015.
35. Fagerholm U. Prediction of human pharmacokinetics – evaluation of methods for prediction of volume of distribution. J Pharm Pharmacol, 2007; 59:1181-1190.
36. Welling PG, Baumueller A, Lau CC, Madsen PO. Netilmicin pharmacokinetics after single intravenous doses to elderly male patients. Antimicr Agents Chemother, 1977; 12:328-334.

37. Wenk M, Spring P, Vozeh S, Follath F. Multicompartment pharmacokinetics of netilmicin. Eur J Clin Pharmacol, 1979; 16(5):331-334.
38. Corvaisier S, Charpiat B, Mounier B, Wallon M, Leboucher G, Kurdi MA, Chailet J-F, Payron F. Population pharmacokinetics of pyrimethamine and sulfadoxine in children treated for congenital toxoplasmosis. Antimicrob Agents Chemother, 2004; 48(10):3794-3800.
39. Whiteman PD, Fowle ASE, Hamilton MJ, Peck AW, Bye A, Dean K, Webster A. Pharmacokinetics and pharmacodynamics of procyclidine in man. Eur J Clin Pharmacol, 1985; 28:73-78.
40. Beedham C, Miceli JJ, Obach RS. Ziprasidone metabolism, aldehyde oxidase, and clinical implications. J Clin Psychopharmacol, 2003; 23(3):229-232.
41. Kau ST, Rama Sastry BV. Distribution and pharmacokinetics of triamterene in rats. J Pharm Sci, 1977; 66(1):53-56.
42. Mutschler E, Gilfrich HJ, Knauf H, Moerke W, Voelker KD. Pharmacokinetics of triamterene. Clin Exp Hypertens A, 1983; 5(2):249-269.
43. Luke DR, Foulds G, Cohen SF, Levy B. Safety, toleration, and pharmacokinetics of intravenous azithromycin. Antimicr Agents Chemother, 1996; 40:2577-2581.
44. Foulds G, Shepard RM, Johnson RB. The pharmacokinetics of azithromycin in human serum and tissues. J Antimicr Chemother, 1990; 25(Suppl. A):73-82.
45. Glauman H. Pentamidine accumulates in rat liver lysosomes and inhibits phospholipids degradation. Pharm Toxicol, 1994; 74(1):17-22.
46. Ming X, Ju W, Wu H, Tidwell RR, Hall JE, Thakker DR. Transport of dicationic drugs pentamidine and furamidine by hOCTs. Drug Metab Dispos, 2008; 37(2):424-430.
47. Sissi C, Moro S, Richter S, Gatto B, Menta E, Spinelli S, Krapcho AP, Zunino F, Palumbo M. DNA-interactive aza-antrapyrazoles: biophysical and biochemical studies relevant to the mechanism of action. Mol Pharmacol, 2001; 59(1):96-103.
48. Adelusi SA. Tissue and blood concentration of chloroquine following chronic administration in the rat. J Pharm Pharmacol, 1982; 11:733-735.
49. Daniel WA, Bickel MH, Honegger UE. The contribution of lysosomal trapping in the uptake of desipramine and chloroquine by different tissues. Pharmacol Toxicol, 1995; 77(6):402-406.
50. Sams WM, Epstein JH. The affinity of melanin for chloroquine. J Invest Dermatol, 1965; 45:482-488.