Predicting Drug-Induced Liver Injury with Bayesian Machine Learning

Dominic P. Williams[†§*], Stanley E. Lazic[‡§•], Alison J. Foster[†], Elizaveta Semenova[‡] & Paul Morgan[†]

[†]. Safety Platforms, Clinical Pharmacology and Safety Sciences [‡]. Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge U.K.

[§] These authors contributed equally to the manuscript

Email: alison.foster2@astraeneca.com; stan.lazic@prioris.ai;

Dominic.P.Williams@astrazeneca.com; Elizaveta.Semenova@astrazeneca.com;

Paul.Morgan@astrazeneca.com

[*]Corresponding author: Dominic Williams, Safety Platforms, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Cambridge Science Park, Cambridge, Cambridgeshire, CB4 0WG, United Kingdom, email address: Dominic.P.Williams@astrazeneca.com

[•] Current address: Prioris.ai Inc., 459-207 Bank Street, Ottawa, K2P 2N2, Canada

Abstract

Drug induced liver injury (DILI) can require significant risk management in drug development, and on occasion can cause morbidity or mortality, leading to drug attrition. Optimising candidates preclinically can minimise hepatotoxicity risk but it is difficult to predict due to multiple aetiologies encompassing DILI, often with multifactorial and overlapping mechanisms. In addition to epidemiological risk factors, physicochemical properties, dose, disposition, lipophilicity, and hepatic metabolic function are also relevant for DILI risk. Better human-relevant, predictive models are required to improve hepatotoxicity risk assessment in drug discovery. Our hypothesis is that integrating mechanistically relevant hepatic safety assays with Bayesian machine learning will improve hepatic safety risk prediction. We present a quantitative and mechanistic risk assessment for candidate nomination using data from *in vitro* assays (hepatic spheroids, BSEP, mitochondrial toxicity and bioactivation), together with physicochemical (cLogP) and exposure ($\mathrm{Cmax}_{\mathrm{total}}$) variables from a chemically diverse compound set (33 no/low-, 40 medium- and 23 high-severity DILI compounds). The Bayesian model predicts the continuous underlying DILI severity and uses a data-driven prior distribution over the parameters to prevent overfitting. The model quantifies the probability that a compound falls into either no/low, medium, or high-severity categories, with a balanced accuracy of 63% on held-out samples, and a continuous prediction of DILI severity along with uncertainty in the prediction. For a binary yes/no DILI prediction, the model has a balanced accuracy of 86%, a sensitivity of 87%, a specificity of 85%, a positive predictive value of 92%, and a negative predictive value of 78%. Combining physiologically relevant assays, improved alignment with FDA recommendations, and optimal statistical integration of assay data, leads to improved DILI risk prediction.

Introduction

Predicting drug-induced liver injury is a challenge for drug developers and regulators. Preclinical animal studies are required for investigational new drug approvals by health regulatory agencies, however, retrospective analysis has revealed such tests fail in predicting risk for drug-induced liver injury (DILI) in about 45% of clinical trials.[1] These findings reinforce the need to improve the predictive power and to understand contributing factors of both human-based *in vitro* and *in silico* DILI models [2], consequently a variety of approaches have been reported to estimate human DILI risks both *in silico* [3-6] and *in vitro* [7-10].

It has been reported that oral medications of high lipophilicity (i.e., logP >3) administered at daily doses of >100 mg are associated with an increased risk of developing DILI [2] and that incorporating additional factors, especially those relevant to mechanisms of DILI [7], facilitates the development of quantitative predictive models. Moreover, high systemic exposure to drugs is associated with increased risk of DILI [11-12] and there is general acceptance, as well as considerable evidence, that low doses are desirable. Ideally, doses <100 mg/day should be the goal for oral drugs [13]. For example, total daily dose was the major differentiating factor between two groups of drugs in an evaluation of the top 200 oral drugs in 2009 (based on prescription and sales in the USA) and of 68 drugs recalled or associated with a black box warning due to idiosyncratic toxicity.[14] Likewise, the vast majority of oral drugs with reported idiosyncratic liver toxicity are administered at high clinical doses [15].

Bioactivation to reactive metabolites (RM) and subsequent covalent binding is a mechanism of DILI frequently discussed in the literature and has been shown to cause liver injury through direct toxic events [16] or through activation of the immune system [17]. However, there is no simple correlation between drug bioactivation *in vitro* and adverse drug reactions (ADRs) in the clinic. This association is obviously limited because not all drugs that undergo bioactivation by human drug-metabolizing enzymes are associated with ADRs in the clinic, and drug bioactivation is not always necessary to initiate hepatic drug toxicity.[18] Approximately 80% of drugs associated with toxicity contain a structural alert and of these ~65% formation of reactive metabolites has been suggested as a causative factor in the toxicity [14, 18].

Chemically reactive metabolites show considerable differences in electrophilicity, intracellular targets, pathway and degree of stress signalling, detoxication pathways and immunological recognition of the haptens (protein adducts) that they subsequently produce. Also little is understood about the relationship between these chemical factors and the mechanisms that underlie clinical hepatotoxicity. Numerous drugs have been reported to generate RM; however, their causative relationship for human DILI remains controversial and is at best inconclusive[18], partly due to the fact that protein adducts seen with drugs are not necessarily associated with liver injury.[14] A retrospective analysis suggested no correlation between incidence of liver toxicity observed *in vivo* in preclinical studies and the level of covalent binding.[18] Nonetheless, minimizing the potential of RM formation for drug candidates is strongly recommended[19-21], and a target threshold of <50 pmol of RM bound to 1 mg protein is considered to be safe.[22] Moreover, evidence suggests that although a drug's propensity to produce RM might be a contributing factor, it is not sufficient alone for predicting risk for developing DILI.[18] Other studies have indicated that RM formation combined with drug exposure (e.g., Cmax or daily dose) could improve prediction of risk for developing DILI.[23-27]

To date models used to predict DILI are typically either quantitative structure-activity relationship (QSAR) models designed to assess a specific hepatotoxicity mechanism (e.g. BSEP inhibition) or simple rules-based models. The previous AstraZeneca hepatotoxicity "bin and sum" hepatic risk assessment [28] qualitatively flagged the potential for hepatic risk but did not adequately enable projects to quantify DILI risk at the time of major investment decisions, particularly with respect to the relative contribution of *in vitro* assays to predicting liver toxicity and how to integrate and quantify the associated uncertainty of the predictions. Indeed this is a gap in previous hepatotoxicity predictive models in that they do not quantify probability and uncertainty in a simple way for practical use [29]. The use of Bayesian approaches in machine learning has, until recently [30], been limited to naïve Bayes, which makes simple and often unrealistic assumptions. In addition, most in silico models described in the literature focus on binary prediction (low vs high severity). It is more difficult to develop a model to predict multiple levels of DILI severity [31]. Here we have combined in silico, physicochemical and in vitro data, to seamlessly combine hepatic safety and statistical modelling to provide an integrated assessment using Bayesian machine learning, this is currently used within AstraZeneca when comparing a short-list of compounds for prioritisation. This model provides a quantitative assessment of DILI severity through the integration of five physiologically relevant, human cell-based *in vitro* assays along with additional physicochemical, exposure and metabolic parameters, each contributing to increased predictivity of the model. The model demonstrated excellent performance of the model across 96 compounds, 80 of which had been independently classified by the Food and Drug Administration (FDA) in the Liver Toxicity Knowledge Base (LTKB) database [32-33], 13 classified in Proctor *et al* 2017[10], 2 classified in Gustafsson *et al* 2013[9] and 1 annotated in house from literature reports. Since this novel approach generates a posterior distribution (summarising uncertain quantities in Bayesian models) of DILI severity, it provides an important ability to not only quantify the overall DILI severity, but equally to quantify the uncertainty in the individual prediction. Additionally, since this can vary substantially amongst compounds demonstrating varying assay characteristics, this is a key differentiator over any previous quantitative DILI predictive models. Firstly, as the model incorporates probability and uncertainty, it enables project teams to fully assess the likelihood of DILI severity in early clinical development including the need for dose adjustment in SAD/MAD. Secondly, the machine learning approach enables weighting of the hepatotoxicity assays for DILI severity which in turn fuels mechanistic studies to better understand the risk and minimise in drug design phase. Thirdly, the model has been developed to facilitate addition of new or improved hepatotoxicity assays as mechanistic understanding of DILI develops. Lastly, a user-friendly visualisation has been developed which includes a ‘front-end’ web interface enabling project teams to assess which compound properties are most contributing to DILI risk and which can subsequently be adjusted during drug design.

Materials & Methods

Chemicals and Reagents

All chemicals and cell culture plates were purchased from Sigma-Aldrich Company Ltd., Poole, Dorset, UK or VWR International, Leicester, UK unless stated otherwise. HepG2/C3A and THP1 cells were obtained from American Type Culture Collection (ATCC), Manassas, VA, USA. HepG2 cells were obtained from European Collection of Cell Cultures (ECACC), Salisbury, UK. Eagle's minimum essential media (EMEM), Dulbeccos modified essential media (DMEM) , HEPES, sodium pyruvate, galactose, foetal calf serum (FCS, Glutamax were obtained from ThermoFisher Scientific Loughborough, UK. CellTiter Glo® and 3D CellTiter Glo® were obtained from Promega UK Ltd., Southampton, UK. Test compounds (>95% purity) were obtained from Compound Management, AstraZeneca R&D, Macclesfield, UK.

Drug databases

LTKB provides a centralized repository of information for the study of DILI and predictive model development. The DILI classification data in LTKB has been used to develop the predictive model.[34] Eighty of the ninety six compounds evaluated in the predictive model had been independently classified by the FDA in the LKTB database.[32-33] For those compounds not present in the LTKB database, DILI annotation was taken from Proctor *et al* 2017 (thirteen compounds[10]), Gustafsson *et al* 2013 (two compounds[9]) and one was annotated in house using the criteria detailed in Proctor *et al* 2017[10] (see Supplementary Table 1). Human $Cmax_{total}$ values were collated from the literature, micromedex, Clarke's Analysis of Drugs and Poisons or FDA prescribing information (see Supplementary Table 1). cLogP values were obtained from the AstraZeneca D360 Scientific Informatics Platform, a scientific business intelligence application used across life science research for drug discovery and development, that supports analytics workflows requiring data access and analysis. Reports of bioactivation were collated from the literature and from the NIH Liver tox database (see Supplementary Table 1). For the consideration of bioactivation, we aligned our consideration with the published FDA approach [35], who performed a comprehensive literature search to retrieve publicly available reactive metabolite data. They considered experimental data based on the measurement of covalent binding either in in vitro or in vivo animal studies or thioether adducts and/or conjugates detected by mass spectrometric analysis [35]. Each of the 96 investigated drugs were manually checked in PubMed searches for evidence of reactive metabolite reports using the keywords "drugs AND (reactive metabolites OR covalent binding OR bioactivation OR glutathione conjugate OR active metabolites)," and a total of 42 had evidence of bioactivation. Similarly, the reactive metabolite-negatives were defined by lack of experimental evidence for reactive metabolites and no positive literature finding, and this left 54 reactive metabolite negatives [35].

Bayesian Model Definition

The Bayesian predictive model is outline in Figure 1 and the distribution of the DILI categories and assay values are shown in Supplementary Figure 1. Liver injury is, in reality, a multidimensional continuous variable, but we only observe discrete categories, which need to be predicted from the assay values and other data. The model estimates both the underlying continuous severity for each compound and the cut-points or thresholds that define the boundaries between the severity categories. The cut-points are the optimal values that separate the DILI categories and are estimated from the data.

The model is a proportional odds logistic regression model (also known as a cumulative logit or ordered logit model) and defined as

$$\mathrm{Severity_i} \sim \mathrm{OrderedLogistic}(\eta_i, c_1, c_2)$$

$$\mathrm{logit}(\eta_i) = X_{ij} \times \beta_j$$

$$c_1, c_2 \sim \mathrm{Normal}(0, 20)$$

$$\beta_j \sim \mathrm{Laplace}(\mu, \sigma)$$

$$\mu \sim \mathrm{Normal}(0, 2)$$

$$\sigma \sim \mathrm{HalfNormal}(0, 0.5).$$

Here, DILI severity is modelled as an ordered categorical variable with three levels (1 = no DILI, 2 = intermediate DILI, and 3 = severe DILI), and the cut-points between categories are determined by the parameters c_1, and c_2, which are estimated from the data as the optimal thresholds. The parameter η is the continuous prediction of DILI severity and is mapped to lie between zero and one using a logit transform. The subscript i indexes the compound, and since each η is subscripted, this indicates that each compound has its own DILI severity estimate. η is an unknown parameter that we estimate from the data, and its value depends on the matrix X, which contains the assay values and the other predictor variables (each row of X is a compound and each column is an assay). X also contains the interaction terms in the model and therefore has 29 columns (8 predictors plus 21 interactions), which are indexed by the subscript j. The interactions allow for the possibility that two weak signals might be highly predictive of toxicity—more than the sum of the two individual signals. $X_{ij} \times \beta_j$ is a short-hand notation for all terms to the right of the equal sign in Figure 1.

η is a deterministic function of the predictor variables contained in X and the β parameters (this is indicated with an equal sign instead of a $\sim$). The β parameters are also estimated from the data; they quantify the strength and direction of the relationship between the assay values and DILI severity. They can be thought of as weights, where a large value means that a change in a predictor variable, such as an IC_{50} value, is associated with a large change in DILI severity. We are estimating 31 parameters (29 β's plus the two cut-points) using only 96 compounds, and so are at risk of overfitting. A common solution to prevent overfitting is to shrink the β's towards zero, with the amount of shrinkage determined by a hyperparameter. This hyperparameter is not directly estimated from the data; typical non-Bayesian methods try many values, and the "best" value is empirically determined as the one that provides the best accuracy, and this value is used in the final model. Cross-validation (CV) is traditionally used to estimate the best value, which requires splitting the data into training and validation sets. However, classical CV has two drawbacks. First, it is difficult to obtain stable estimates with small data sets such as this one. Second, the best value of the hyperparameter is uncertain, but this is ignored when a single value is plugged-in to the predictive model. A Bayesian approach overcomes both of these problems because the hyperparameter is learned from the data—just like the other parameters, and so CV is unnecessary (Supplementary Figure 2). Furthermore, the uncertainty in the estimate is propagated through to the predictions.

To prevent overfitting, we model the β parameters as coming from a Laplace (double exponential) distribution. The Laplace distribution is symmetric and is defined by a mean (μ), and standard deviation (σ; see Supplementary Figure 3 for an example). σ is the hyperparameter that controls the amount of shrinkage— a small σ constrains the β's to be close to zero and makes the model less flexible. We used a Laplace distribution because it shrinks small β's close to zero but still allows for a few large β's, similar to classical LASSO or $L1$-regularisation. However, the choice of distribution had little effect on the predictions (Gaussian and Student-t distributions were also considered).

In the Bayesian framework, all unknowns that are estimated from the data require a prior distribution, which specifies information, if any, about the unknowns before seeing the data.

The priors that we need to specify are: the cut-points (c_1 and c_2), the mean of the distribution of β parameters (μ), and the standard deviation of the β's (σ). These priors given above are "uninformative" in that we are not including external information and the data will mainly determine the final values. The one exception is for μ, which we expect to be close to zero and so it has a narrow prior (often this parameter is set to zero for convenience, but we prefer to estimate all unknowns). σ is the key parameter because it controls the flexibility of the model and therefore the likelihood of over-fitting. Supplementary Figure 2 shows that the predictions are insensitive to the choice of prior for σ (as long as it is not too small so that the model underfits the data).

We developed the model in Stan (version 2.18[36]) using the RStan interface (version 2.18.2). We used the No-U-Turn sampler variant of Hamiltonian Monte Carlo with 3 chains of 10000 iterations, and discarded the first 5000 samples as warm up, leaving 15000 samples for final inferences. The usual numeric and graphical checks such as trace plots, posterior predictive checks, r-hat values, and so on indicated no problems with the fitted model. The R and Stan code are provided in the supplementary material, and to facilitate adoption and adaptation of the model, we have translated the code into Python (PyMC3) and Julia (Turing).

Assessment of HepG2/C3A Spheroid Cytotoxicity

The spheroid screen measures the effects of test compounds on the viability of a spheroid of HepG2/C3A cells, as visualised and quantitated by the addition of a 3D optimised ATP sensing Cell Titre Glo reagent (luminescence read, Promega). HepG2/C3A cells were seeded in EMEM containing 5% FCS v/v, 1% non-essential amino acids v/v and 1% Glutamax v/v at 95 µl/well and 1000 cells/well into 96-well ultra-low attachment (ULA) plates or 40 µl/well and 500 cells/well into 384-well ULA plates, and allowed to self-assemble into spheroids over 5 days at 37^{o}C with 5% CO^2. Compound treatment commenced 5 days after seeding, with spheroids exposed to 0-250 µM test compound (0.5% DMSO v/v final) for 4 days prior to assessment of cell viability using the CellTiter-Glo® 3D Cell Viability assay. Spheroids were incubated with 100 µl (96-well plates) or 30 µl (384-well plates) undiluted 3D CellTiter-Glo® assay reagent for 1 min prior to agitation for 5 mins. The assay was then performed as per the manufacturers protocol with luminescence determined on an Envison^{TM} Multiplate Reader (Perkin Elmer, Waltham, MA, USA).

Assessment of Mitochondrial Toxicity

Assessment of toxicity in HepG2 Cells in Galactose vs Glucose Medium (MitoTox) was performed as described previously.[28]

Briefly HepG2 cells (5000 cells/well, 96-well plate) were cultured in either DMEM glucose medium (DMEM with 4 mM L-glutamine, 4.5 mg/mL glucose, and 1 mM sodium pyruvate plus 10% FBS, 5 mM HEPES) or DMEM glucose free medium (DMEM with 4 mM L-glutamine plus 1 mM sodium pyruvate, 10% FBS, 5 mM HEPES, and 10 mM galactose) in the presence of test compound for 24 h. Cell viability was then determined by the CellTiter-Glo® Assay following the manufacturer's protocol, with luminescence determined on an Envison™ Multiplate Reader (Perkin Elmer, Waltham, MA, USA). The final concentrations of the test compounds ranged from 0 - 250 µM (0.5% v/v DMSO).

Assessment of BSEP Inhibition

Assessment of inhibition of the human BSEP (hBSEP) transporter was performed as described previously.[37]

Briefly the effect of the test compounds on ATP-dependent uptake of probe substrates into inside-out membrane vesicles prepared from baculovirus-infected Spodoptera frugiperda Sf21 insect cells expressing human BSEP were quantified using rapid filtration assays in 96-well plate format followed by scintillation counting or LC MS/MS analysis. The BSEP substrate was [^{3}H]taurocholate (PerkinElmer, Waltham, MA) or un-labelled taurocholate. The final concentrations of the test compounds ranged from 0 - 1000 µM (2% v/v DMSO). Incubations were also undertaken using 5 mM AMP in place of ATP at each test compound concentration to quantify substrate accumulation independent of active transport.

Assessment of THP-1 cytotoxicity

The THP-1 cytotoxicity screen is based upon determination of fluorescent signal generated by the reduction of non-fluorescent resazurin (7-Hydroxy-3H-phenoxazin-3-one 10-oxide) to the fluorescent resorufin (Alamar blue assay). Cellular reduction of resazurin is dependent on a pool of reductase or diaphorase enzymes derived from the mitochondria and cytosol. Therefore, Alamar blue can be used as an oxidation-reduction indicator in cell viability assays for mammalian cells.[38-39]

THP-1 cells were seeded in 96-well plates at 40,000/well and 95 µl/well in RPMI supplemented with 1% 2mM Glutamine v/v and 10% heat inactivated foetal bovine serum v/v and 5 µl test compound added immediately (0.5% DMSO v/v). The cells were then incubated at 37^{o}C with 5% CO^2 for 48 h. Resazurin solution (10 µl 450 µM) was added to all wells, the plates mixed and then incubated for a further 2 h at 37^{o}C with 5% CO^2, prior to the determination of fluorescence (excitation λ 560 nm, emission λ 590 nm) using an Envison^{TM} Multiplate Reader (Perkin Elmer, Waltham, MA, USA). The final concentrations of the test compounds ranged from 0 - 250 µM (0.5% v/v DMSO).

Estimation of cytotoxicity EC_{50} values

EC_{50} values, the half-maximal effective concentration, were estimated in GeneData (Genedata, Basel, CH).

Results

Figure 2 shows the prediction for one compound, along with the assay values. The x-axis is the continuous predicted DILI severity, and the two dashed vertical lines are the estimated cut-points. The shaded blue distribution shows the predicted DILI severity as well as the uncertainty in the prediction. The percentages near the x-axis represent the proportion of the blue distribution in each category. These values differ from, but are closely related to, the category predictions, which are given in the box labelled "Probability". To understand the difference, it is easier to consider an outcome with only two possibilities instead of three. Consider a coin-flipping experiment where we estimate the probability that a coin lands heads is 0.6, with a very narrow 95% CI of 0.58 to 0.62. If we take 0.5 as our cut-point, 100% of this predicted distribution is above 0.5 and 0% is below, but it doesn't follow that we are 100% certain that the coin will land heads. If we then flip this coin many times, we will get tails about 40% of the time (and 60% heads). Hence, we distinguish between our estimate for the propensity of the coin to land heads (known as the posterior distribution, which summarises our uncertainty in an unknown and unobservable parameter), and what will happen once the coin is flipped (known as the posterior predictive distribution, which summarises our uncertainty about a future observable outcome). Similarly, the blue distribution represents the underlying continuous but unobservable DILI severity, and the proportions in the box are our predictions for an observable outcome. For Acetaminophen therefore, we predict that the probability of being in categories 1, 2, and 3 are 0.06, 0.55, and 0.39, respectively. Category two is the most likely with a probability of 0.55, and that is the true category.

Summary statistics for the blue distribution are provided in the top right of Figure 2, and details of the assays and other data are provided below. If the model predicts a DILI risk, then the next question is “why”? The dots indicate which predictor variables are providing the risk signal, and the numbers on the right are the measured IC_{50} values and other data (Bioactivation (BA) is encoded as No = -1 and Yes = +1). See Table 1 for a description of each variable.

Spheroid C3A: IC50 (µM) cell health screen of cytotoxicity in HepG2 C3A spheroids.
BSEP: IC50 (µM) inhibition of the hepatic biliary transporter activity screen in vesicles.
THP1: IC50 (µM) a cell health screen using a monocytic cell line sensitive to chemotoxicity.
Glu: IC50 (µM) for cell health screen in HepG2 cell cultured in glucose media. Glu/Gal: Mitochondrial toxicity ratio from two in vitro assays measuring the IC50 (uM) at which compounds cause ATP depletion via direct impairment of mitochondrial function in cells cultured in galactose media vs glucose media.
cLogP: Physicochemical partition coefficient indicating lipophilicity.
Cmax: Total Cmax (µM) is the maximum (or peak) plasma concentration that a drug achieves after administration.
Bioactivation: A binary indicator (-1 or +1) of whether a compound can form a reactive metabolite. A value of "Unknown" splits the difference between "Yes" and "No" (encoded as a value of 0) and is intended to be a temporary placeholder until this information becomes available.

Table 1 Description of each of the assays / variables used for the Bayesian Predictive model.

We can see (qualitatively) that bioactivation, Cmax, and ClogP are the main drivers for the acetaminophen signal. This graph is the main output for the predictive model and is used by project teams to make decisions. Similar graphs for all 96 compounds are provided as Supplementary File 1 and supplementary code.

Figure 3 plots the results for all 96 compounds, where each point represents the median value of the posterior distribution. Overall accuracy when fitting the model on 95 compounds and using it to predict the remaining compound is 67%, compared with always predicting category 2, the most frequent, which gives an accuracy of 42%. Since the categories are unbalanced, the overall accuracy can be misleading, and we therefore calculated the balanced accuracy, which is 63%. The balanced accuracy is the average proportion correct across the three DILI categories, and it differs from the usual accuracy metric (the total proportion correct) when the number of samples in the categories is unequal. Figure 3 shows that there is better separation between categories 1 and 2 compared with categories 2 and 3. No category 1 compounds were misclassified as category 3, although Zomepirac came closest. Only one category 3 compound was misclassified as category 1, which was ximelagatran. The assays incorporated in the model are unlikely to be sufficiently complex to capture the idiosyncratic hepatotoxicity mechanism of ximelagatran toxicity [40].

We can also calculate the balanced accuracy for predicting “safe” versus “any DILI” (category 1 versus category 2 or 3), which equals 86% (95% CI = 78% to 93%), and a sensitivity and specificity of 87% and 85%, respectively. Furthermore, assuming that the prevalence of safe (category 1) compounds that we will observe in the future is the same as the proportion in this data set, the positive predictive value is 92% and the negative predictive value is 78%. However, we do not place much emphasis on these metrics because (1) the DILI severity categories do not exist in nature, they are just one way of mapping multidimensional clinical outcomes to numbers, (2) they do not take into account how bad a wrong prediction is (close to an arbitrary threshold versus strongly predicting the wrong category) or how good a correct prediction is (just on the right side of the threshold or confidently predicting the correct category), and (3) the classifications do not account for the severity of different types of incorrect decisions, such as progressing a toxic compound versus halting development of a safe compound. Instead, we prefer to examine the graphical displays (Figure 2) for decision making, but recognise that the metrics enable the performance of different models and assays to be compared.

Comparing predictions from structurally similar compounds

The performance of the Bayesian DILI model was further investigated to distinguish drug pairs as defined by their molecular structure (Tanimoto similarity >0.5) and similarity in mode of action but discordant toxicity (Figure 4a). Overall, the model is able to discriminate between different DILI severity categories for structurally similar compounds.

Discussion

Improved prediction and mitigation of risk is urgently required in safety sciences, particularly hepatic safety. Here we report on a novel approach to prediction of DILI risk incorporating quantitation of probability and uncertainty of DILI severity which is an improvement over previous published DILI models [2-3, 23, 28-29, 41-42] and in a form which is easily communicated to project teams. ,. We have improved on previous hepatic safety risk assessment models and visualisation through a number of new features which are detailed in Supplementary table 2. From a chemical and biological aspect this includes: 1) incorporation of a larger (96 compounds) and more chemically diverse compound set, 2) more relevant in vitro assays in terms of hepatic toxicity predictive capability and physiological relevance, 3) alignment of both the compound severity classification and the binary use of bioactivation as a predictive variable with the FDA criteria [35]. Statistical improvements include using a Bayesian model to optimally combine and integrate assay data which allows a flexible framework for the future addition or removal of assays based upon emerging science [43-46]. Finally, a key aspect to risk mitigation in a project setting is the simple communication of complex datasets and here we use a visualisation of the full probability distribution of DILI severity with comparisons to known DILI compounds allowing for contextualisation of uncertainty in predictions.. Indeed communication of uncertainty to project teams is particularly important in order to progress from a simplistic approach of a compound falling into either a ‘safe’ or ‘risk’ category [28].

The overall balanced accuracy for predicting the three DILI categories was 63%, compared with 33% for random guessing and 42% for always predicting the most frequent category. These results are an improvement over the only other 3-category DILI model that we are aware of by Hong et al., [31], but their results are not directly comparable to ours as both the compounds and the predictors are different. Distinguishing between category 2 and 3 was the most difficult, but when we collapse these categories and compare against category 1, the accuracy improves to 86%, with good sensitivity (87%) and specificity (85%). However, these metrics ignore the uncertainty in the predictions. Figure 3A illustrates the problem ignoring this uncertainty and the benefits of using a machine learning model that can make continuous predictions: most of the category 3 compounds that were misclassified as category 2 were very close to the threshold that separates the categories, and when looking at the full distributions, it is probable that similar go/no go decisions would be made for all compounds near the category 2/3 boundary.

We have successfully applied the Bayesian model to distinguish between drug pairs, which are defined by their similar chemical structure (Tanimoto similarity >0.5) and comparable pharmacological mode of action, although differing in their toxicity profile. Examples are shown in Figure 4, in each case the culprit drug demonstrated the highest probability of hepatic risk, with simple visualisation of the parameters contributing to the risk profile seen to the right of the associated curve. Buspirone demonstrated predominantly (0.54) low severity probability most likely due to a very low Cmax, yet part (0.43) of the curve is in the medium severity zone, suggesting a level of uncertainty most likely due to the ability of buspirone to undergo bioactivation. When we compare buspirone to nefazodone (known clinical DILI positive) we see a very strong shift to the right, suggesting a greater risk, most likely through a greater Cmax, ability to undergo bioactivation, potent BSEP IC_{50} and a high Glu/Gal ratio. The relative shape and positioning of the DILI probability curve (Figure 4) allows a simple visualisation of the confidence in the DILI prediction with a left-shifted curve predicting less DILI severity versus a right-shifted curve (e.g. ambrisentan vs bosentan, Figure 4) and a narrow curve denoting greater confidence in prediction of DILI category versus a broad curve (e.g. clozapine vs olanzapine, Figure 4).   Providing further confidence in the model, we obtained a similar prediction of severity between pioglitazone and troglitazone as Chen et al [35], with the majority of the pioglitazone profile falling into medium severity (0.6 in medium; 0.28 in high severity), a greater proportion of the profile for troglitazone falls into high severity (0.37), with the narrower shape of the curve indicating the greater confidence in the prediction for troglitazone.

It is important to highlight that many drug molecules have potential to form reactive metabolites, irrespective of the drug being classified as safe or known to cause DILI. With the increasing sensitivity of mass spectrometers, ultra-low formation of drug-glutathione or drug-protein conjugates are now able to be detected [47]. Bioactivation and Cmax were two of the highest predictors of DILI from variables used within the Bayesian model. Cmax has previously been shown to be predictive of DILI [10, 48]. We have found that point estimates of covalent binding to human hepatocytes are not predictive of clinical DILI, it is only once the body burden is calculated through incorporation of the daily dose the predictivity becomes significant. Aligning estimated exposure parameters for reactive metabolites between *in vivo* animal studies and *in vitro* hepatocyte experiments is extremely difficult, especially considering the breadth and variation in protein target expression between animals and humans. In cases where the estimated body burden was in a risk zone, additional safety endpoints in humans would be expected by the regulatory authorities. Given the resource requirements for radiochemical synthesis for ^{14}C, or ^{3}H labelling subject to lability of the tritium, this was very rarely carried out during the early discovery phase and was usually confined to preclinical development phase if at all. Similar to Chen et al., we use evidence of bioactivation as a binary predictive variable, based on FDA databases and literature. Assays used to determine bioactivation (GSH trapping, covalent binding etc) have inherent potential for both false positives and negatives. Lack of evidence/data may also contribute to inadequate use of bioactivation as a variable. Some drugs prone to bioactivation may lack good clinical evidence of hepatotoxicity. Other hepatotoxic compounds may not form GSH adducts, and not have been evaluated for covalent binding [35]. Additionally, the lack of standardised protocols used in these bioactivation assessments across the literature equates to uncertainty, yet we have found for the GSH trapping assay that higher substrate concentrations and longer assay incubations times leads to a greater chance of a positive result.

It has previously been shown within the IMI-1 MIP-DILI Consortium, in a multi-centre ring trial [46], that in 2D, HepG2 cells have a similar predictive capability to primary human hepatocytes [46]. We also found the 2D HepG2 assays had predictivity (for DILI) for the 96 compounds included here, we have incorporated the ATP IC50 for HepG2 grown in glucose as an endpoint, which is generated through the mitochondrial glu/gal ratio, which has also been demonstrated to have predictive capability for DILI [44-45]. The physiological relevance of the assays is increased through including a 3D HepG2 (clone C3A) spheroid assay that has an ATP endpoint assessed after a 4-day incubation. The C3A clone of HepG2 has been shown to demonstrate contact limited proliferation, and as 3D spheroids they have a workable timeframe of 32 days (from spheroids containing 500 or 1000 cells) before a necrotic core develops due to lack of oxygen availability [49]. They also demonstrate polarised MRP2 and Pgp expression, urea and albumin secretion capability and CYP450 expression *vs* the same cell line in 2D [49]. We recognise that the CYP450 activity is less than that of primary hepatocytes either in 2D or as spheroids, however, the 4 day time course of the assay allows some drug metabolism to occur. The arguments for inclusion of the other assays is that they demonstrated positive predictivity for hepatotoxicity in our model and have also demonstrated this within the literature for BSEP[50] and THP-1 [51], similarly for cLogP[3]. A forward-looking aspect of the model is such that as new assays supersede the current assays in predictive capability, these can be added into the model and the less predictive assays removed, the Bayesian model improves along with the assays. For example, within AstraZeneca we are currently developing a high throughput automated primary human hepatocyte spheroid assay. However, it is important to recognise that simple in vitro assays designed to risk assess thousands of compounds per year have to balance predictivity and throughput, which generally compromises complete physiological, pharmacological and toxicological equivalence to the hepatocyte *in vivo*.

This Bayesian model has been in use within AstraZeneca since August 2017, to inform selection of optimal candidate for projects approaching candidate drug investment decision (CDID) and during pre-IND safety studies. We have found it particularly useful in guiding the need for further bespoke studies to derisk candidates and to model potential DILI scenarios in either the absence of data or recognising some variables may change later in development (e.g. predicted vs observed clinical Cmax). For example, prior to performing reactive metabolite trapping assays, the model allows risk assessment in the absence and presence of bioactivation to visualise whether increased DILI risk wold occur with bioactivation. In the clinic, the melanin concentrating hormone receptor 1 antagonist, AZD1979 [52], demonstrated rapid alanine aminotransferase (ALT) elevations in several subjects after single-dose administration [53]. ALT elevations were not observed in any preclinical toxicology studies and could not be ascribed to any specific mechanism, resulting in project termination. We have applied the Bayesian predictive model to AZD1979 (supplementary figure 4) that demonstrated elevations of liver safety biomarkers in patients during a first time in human trial. The profile demonstrates AZD1979 has high hepatic safety risk. Use of this DILI model would have highlighted significant concerns about AZD1979 and likely these would have needed to be mitigated before progression.

In conclusion, the model developed can be used to predict the severity of DILI with associated uncertainty in the prediction, thereby assisting in selection of candidate drugs for development, guiding further derisking studies and supporting regulatory submission in drug safety reviews or support industry decision making.

Figure 1. The Bayesian Predictive model. Clinically characterised DILI positive and negative compounds are classified according to their DILI severity score. Assays (BSEP, Spheroid, THP1, interaction effects, etc) are used to estimate the unknown beta values, which quantify the strength and direction of the relationship between the assay values and DILI severity, which allows prediction of the underlying continuous severity.

Figure 2. Visualising the prediction for one compound. x-axis is the continuous predicted DILI severity, and the two dashed vertical lines are the estimated cut-points. The shaded blue distribution shows the predicted DILI severity as well as the uncertainty in the prediction, given by the width of the distribution. The black distribution is the average of all the category 3 compounds and is used as a visual reference. The percentages near the x-axis represent the proportion of the blue distribution in each category, rounded to the nearest whole number. The category predictions are given in the box labelled “Probability”, and true category is shown beside. The box labelled “Summary stats” is a breakdown of the distribution statistics. The assay data in the lower right has the assays/parameters listed on the left and the assay value (IC_{50}) on the right. The circle is on a sliding scale of low risk to high risk.

Figure 3. Model predictions for all compounds. Points are the median of the distributions for each compound and error bars are 95% CI.

Figure 4. Performance of the model on compounds with high molecular similarity and identical therapeutic indications, but different potential for hepatotoxicity. The order of severity of clinical hepatotoxicity is (a) buspirone<nefazodone; (b) ambrisentan<bosentan<sitaxentan; (c) olanzapine<clozapine; (d) pioglitazone<rosiglitazone<troglitazone. Distribution colour

indicates the true DILI category. Graphs on the left show the posterior distributions (estimated continuous DILI severity) and bargraphs on the right show the predictions (posterior predictive distributions).

Figure 1

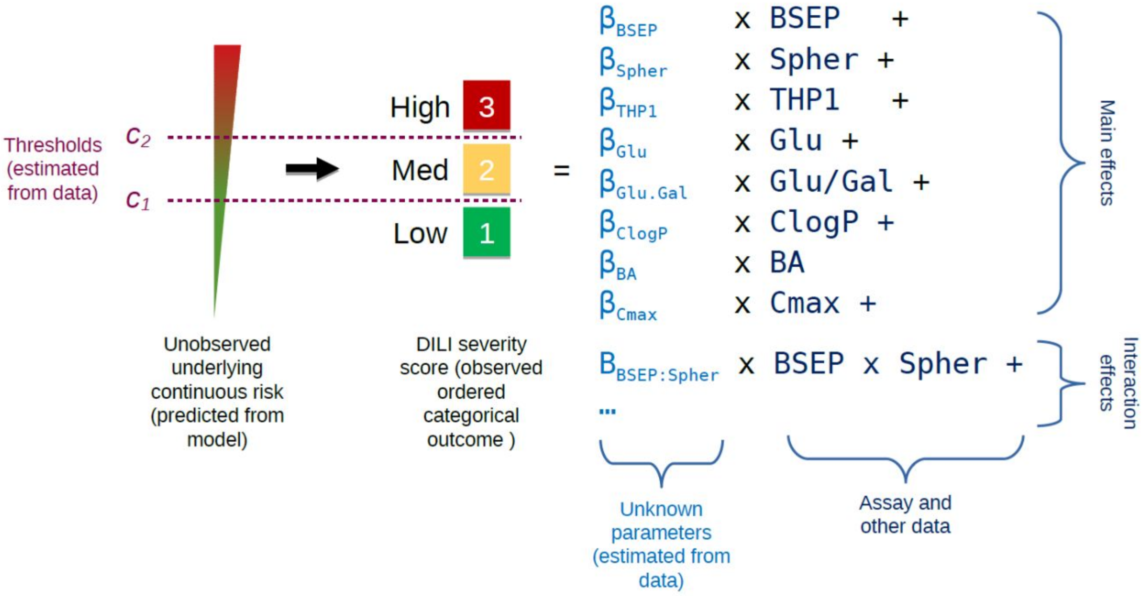

Thresholds (estimated from data)
c_2
c_1
High 3
Med 2
Low 1
=
β_{BSEP} x BSEP +
β_{Spher} x Spher +
β_{THP1} x THP1 +
β_{Glu} x Glu +
$\beta_{Glu.Gal}$ x Glu/Gal +
β_{ClogP} x ClogP +
β_{BA} x BA
β_{Cmax} x Cmax +
$B_{BSEP:Spher}$ x BSEP x Spher +
...
Main effects
Interaction effects
Unobserved underlying continuous risk (predicted from model)
DILI severity score (observed ordered categorical outcome)
Unknown parameters (estimated from data)
Assay and other data

Figure 2

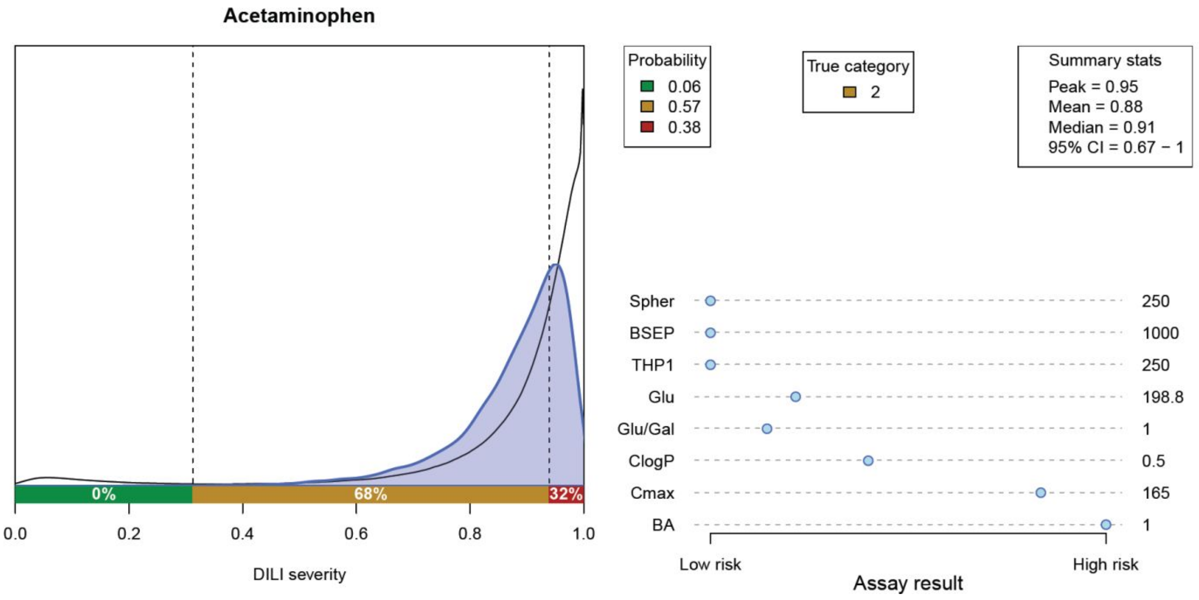

Acetaminophen
0%
68%
32%
0.0
0.2
0.4
0.6
0.8
1.0
DILI severity
Probability
🟩 0.06
🟨 0.57
🟥 0.38
True category
🟨 2
Summary stats
Peak = 0.95
Mean = 0.88
Median = 0.91
95% CI = 0.67 − 1
Spher 250
BSEP 1000
THP1 250
Glu 198.8
Glu/Gal 1
ClogP 0.5
Cmax 165
BA 1
Low risk
High risk
Assay result

Figure 3

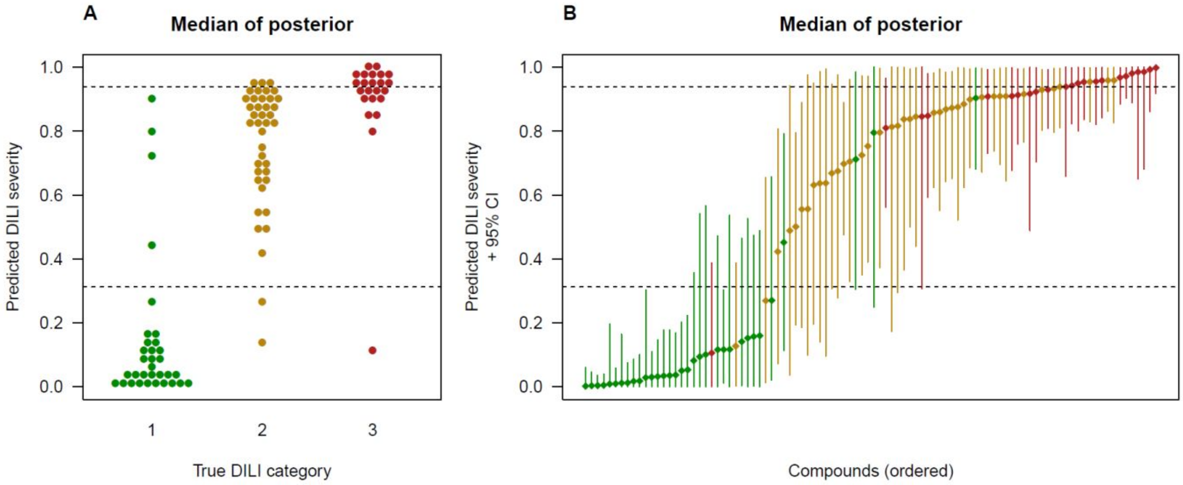
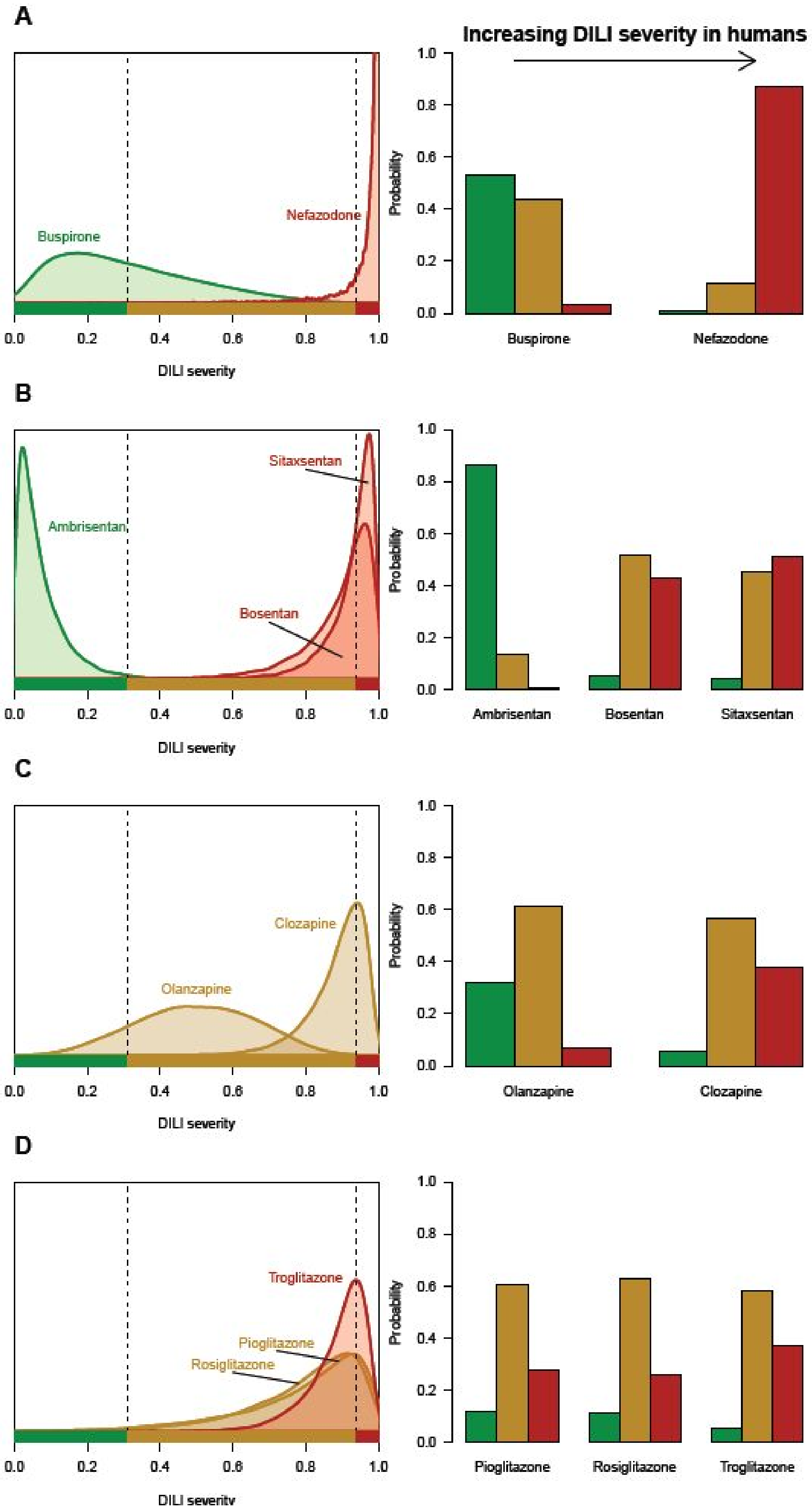

A
Median of posterior
Predicted DILI severity
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
True DILI category
B
Median of posterior
Predicted DILI severity + 95% CI
1.0
0.8
0.6
0.4
0.2
0.0
Compounds (ordered)

Figure 4


A
Increasing DILI severity in humans
Buspirone
Nefazodone
DILI severity
Probability
0.0
0.2
0.4
0.6
0.8
1.0
B
Ambrisentan
Sitaxsentan
Bosentan
C
Clozapine
Olanzapine
D
Troglitazone
Pioglitazone
Rosiglitazone

Conflict of Interest statement

No conflicts of interest

Financial Support statement

No financial support

Author Contributions

SL study concept, development of machine learning model, data analysis, manuscript preparation; DW study concept, manuscript preparation; AF research and collation of data, manuscript preparation; ES translation of R to Python and Julia code, and analysis; PM study concept.

Acknowledgements

We would like to thank Delyan Ivanov, Clare Sefton and Darren Jones for their excellent technical assistance.

References

1. Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; Dorato, M.; Van Deun, K.; Smith, P.; Berger, B.; Heller, A., Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol* **2000,** *32* (1), 56-67.

2. Chen, M.; Bisgin, H.; Tong, L.; Hong, H.; Fang, H.; Borlak, J.; Tong, W., Toward predictive models for drug-induced liver injury in humans: are we there yet? *Biomarkers in Medicine* **2014,** *8* (2), 201-213.

3. Chen, M.; Borlak, J.; Tong, W., High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology* **2013,** *58* (1), 388-96.

4. Chen, M.; Hong, H.; Fang, H.; Kelly, R.; Zhou, G.; Borlak, J.; Tong, W., Quantitative structure-activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. *Toxicol Sci* **2013,** *136* (1), 242-9.

5. Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L., Deep Learning for Drug-Induced Liver Injury. *J Chem Inf Model* **2015,** *55* (10), 2085-93.

6. Zhang, H.; Ding, L.; Zou, Y.; Hu, S. Q.; Huang, H. G.; Kong, W. B.; Zhang, J., Predicting drug-induced liver injury in human with Naive Bayes classifier approach. *J Comput Aided Mol Des* **2016,** *30* (10), 889-898.

7. Aleo, M. D.; Luo, Y.; Swiss, R.; Bonin, P. D.; Potter, D. M.; Will, Y., Human drug-induced liver injury severity is highly associated with dual inhibition of liver mitochondrial function and bile salt export pump. *Hepatology* **2014,** *60* (3), 1015-22.

8. Atienzar, F. A.; Novik, E. I.; Gerets, H. H.; Parekh, A.; Delatour, C.; Cardenas, A.; MacDonald, J.; Yarmush, M. L.; Dhalluin, S., Predictivity of dog co-culture model, primary human hepatocytes and HepG2 cells for the detection of hepatotoxic drugs in humans. *Toxicol Appl Pharmacol* **2014,** *275* (1), 44-61.

9. Gustafsson, F.; Foster, A. J.; Sarda, S.; Bridgland-Taylor, M. H.; Kenna, J. G., A correlation between the in vitro drug toxicity of drugs to cell lines that express human P450s and their propensity to cause liver injury in humans. *Toxicol Sci* **2014,** *137* (1), 189-211.

10. Proctor, W. R.; Foster, A. J.; Vogt, J.; Summers, C.; Middleton, B.; Pilling, M. A.; Shienson, D.; Kijanska, M.; Strobel, S.; Kelm, J. M.; Morgan, P.; Messner, S.; Williams, D., Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. *Arch Toxicol* **2017,** *91* (8), 2849-2863.

11. Uetrecht, J. P., New concepts in immunology relevant to idiosyncratic drug reactions: the "danger hypothesis" and innate immune system. *Chem Res Toxicol* **1999,** *12* (5), 387-95.

12. Walgren, J. L.; Mitchell, M. D.; Thompson, D. C., Role of Metabolism in Drug-Induced Idiosyncratic Hepatotoxicity. *Critical Reviews in Toxicology* **2005,** *35* (4), 325-361.

13. Bayliss, M. K.; Butler, J.; Feldman, P. L.; Green, D. V.; Leeson, P. D.; Palovich, M. R.; Taylor, A. J., Quality guidelines for oral drug candidates: dose, solubility and lipophilicity. *Drug Discov Today* **2016,** *21* (10), 1719-1727.

14. Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D., Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem Res Toxicol* **2011,** *24* (9), 1345-410.

15. Smith, D. A.; Obach, R. S., SEEING THROUGH THE MIST: ABUNDANCE VERSUS PERCENTAGE. COMMENTARY ON METABOLITES IN SAFETY TESTING. *Drug Metab Dispos* **2005,** *33*, 1409-1417.

16. Dahlin, D. C.; Miwa, G. T.; Lu, A. Y.; Nelson, S. D., N-acetyl-p-benzoquinone imine: a cytochrome P-450-mediated oxidation product of acetaminophen. *Proc. Natl. Acad. Sci. USA* **1984,** *81* (5), 1327-1331.

17. Lecoeur, S.; Bonierbale, E.; Challine, D.; Gautier, J.-C.; Valadon, P.; Dansette, P. M.; Catinot, R.; Ballet, F.; Mansuy, D.; Beaune, P. H., Specificity of in Vitro Covalent Binding of Tienilic Acid Metabolites to Human Liver Microsomes in Relationship to the Type of Hepatotoxicity: Comparison with Two Directly Hepatotoxic Drugs. *Chem Res Toxicol* **1994,** *7*, 434-442.

18. Park, B. K.; Boobis, A.; Clarke, S.; Goldring, C. E. P.; Jones, D.; Kenna, J. G.; Lambert, C.; Laverty, H. G.; Naisbitt, D. J.; Nelson, S.; Nicoll-Griffith, D. A.; Obach, R. S.; Routledge, P.; Smith, D. A.; Tweedie, D. J.; Vermeulen, N.; Williams, D. P.; Wilson, I. D.; Baillie, T. A., Managing the challenge of chemically reactive metabolites in drug development. *Nature Reviews Drug Discovery* **2011,** *10*, 292-306.

19. Srivastava, A.; Maggs, J. L.; Antoine, D. J.; Williams, D. P.; Smith, D. A.; Park, B. K., Role of Reactive Metabolites in Drug-Induced Hepatotoxicity. *In: Uetrecht J. (eds) Adverse Drug Reactions. Handbook of Experimental Pharmacology, vol 196. Springer, Berlin, Heidelberg* **2009**.

20. Park, B. K.; Laverty, H.; Srivastava, A.; Antoine, D. J.; Naisbitt, D. J.; williams, D. P., Drug bioactivation and protein adduct formation in the pathogenesis of drug-induced toxicity. *Chemico-Biological Interactions* **2011,** *192* (1-2), 30-36.

21. Kalgutkar, A. S.; Gardner, I.; Obach, R. S.; Shaffer, C. L.; Callegari, E.; Henne, K. R.; Mutlib, A. E.; Dalvie, D. K.; Lee, J. S.; Nakai, Y.; O'Donnel, J. P.; Boer, J.; Harriman, S. P., A Comprehensive Listing of Bioactivation Pathways of Organic Functional Groups. *Current Drug Metabolism* **2005,** *6*, 161-225.

22. Evans, D. C.; Watt, A. P.; Nicoll-Griffith, D. A.; Baillie, T. A., Drug−Protein Adducts: An Industry Perspective on Minimizing the Potential for Drug Bioactivation in Drug Discovery and Development. *Chem Res Toxicol* **2004,** *17* (1), 3-16.

23. Nakayama, S.; Atsumi, R.; Takakusa, H.; Kobayashi, Y.; Kurihara, A.; Nagai, Y.; Daisuke, N.; Okazaki, O., A Zone Classification System for Risk Assessment of Idiosyncratic Drug Toxicity Using Daily Dose and Covalent Binding. *Drug Metab Dispos* **2009,** *37* (9), 1970-1977.

24. Usui, T.; Mise, M.; Hashizume, T.; Yabuki, M.; Komuro, S., Evaluation of the Potential for Drug-Induced Liver Injury Based on in Vitro Covalent Binding to Human Liver Proteins. *Drug Metab Dispos* **2009,** *37* (12), 2383-2392.

25. Yu, K.; Geng, X.; Chen, M.; Zhang, J.; Wang, B.; Ilic, K.; Tong, W., High Daily Dose and Being a Substrate of Cytochrome P450 Enzymes Are Two Important Predictors of Drug-Induced Liver Injury. *Drug Metab Dispos* **2014,** *42* (4), 744-750.

26. Lammert, C.; Bjornsson, E.; Niklasson, A.; Chalasani, N., Oral medications with significant hepatic metabolism at higher risk for hepatic adverse events. *Hepatology* **2010,** *51* (2), 615-620.

27. Vuppalanchi, R.; Gotur, R.; Reddy, K. R.; Fontana, R. J.; Ghabril, M.; Kosinski, A. S.; Gu, J.; Serrano, J.; Chalasani, N., Relationship between characteristics of medications and drug-induced liver disease phenotype and outcome. *Clin Gastroenterol Hepatol* **2014,** *12* (9), 1550-5.

28. Thompson, R. A.; Isin, E. M.; Li, Y.; Weidolf, L.; Page, K.; Wilson, I.; Swallow, S.; Middleton, B.; Stahl, S.; Foster, A. J.; Dolgos, H.; Weaver, R.; Kenna, J. G., In vitro approach

to assess the potential for risk of idiosyncratic adverse reactions caused by candidate drugs. *Chem Res Toxicol* **2012,** *25* (8), 1616-32.

29. Chan, R.; Benet, L. Z., Evaluation of the Relevance of DILI Predictive Hypotheses in Early Drug Development: Review of In Vitro Methodologies vs BDDCS Classification. *Toxicol Res (Camb)* **2018,** *7* (3), 358-370.

30. Lazic, S. E.; Edmunds, N.; Pollard, C. E., Predicting Drug Safety and Communicating Risk: Benefits of a Bayesian Approach. *Toxicol Sci* **2018,** *162* (1), 89-98.

31. Hong, H.; Thakkar, S.; Chen, M.; Tong, W., Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. *Sci Rep* **2017,** *7* (1), 17311.

32. Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W., FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov Today* **2011,** *16* (15-16), 697-703.

33. Chen, M.; Suzuki, A.; Thakker, S.; Yu, K.; Hu, C.; Tong, W., DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today* **2016,** *21* (4), 648-653.

34. Thakker, S.; Chen, M.; Fang, H.; Liu, Z.; Roberts, R.; Tong, W., The Liver Toxicity Knowledge Base (LKTB) and Drug-Induced Liver Injury (DILI) Classification for Assessment of Human Liver Injury. *Expert Review of Gastroenterology and Hepatology* **2018,** *12* (1), 31-38.

35. Chen, M.; Borlak, J.; Tong, W., A Model to Predict Severity of Drug-Induced Liver Injury in Humans. *Hepatology* **2016,** *64* (3), 931-940.

36. Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A., Stan: A probabilistic programming language. Journal of Statistical Software. *Journal of Statistical Software* **2017,** *76* (1), 1-32.

37. Dawson, S.; Stahl, S.; Paul, N.; Barber, J.; Kenna, J. G., In vitro inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. *Drug Metab Dispos* **2012,** *40* (1), 130-8.

38. McMillan, M. K.; Li, L.; Parker, J. B.; Patel, L.; Zhong, Z.; Gunnett, J. W.; Powers, W. J.; Johnson, M. D., An improved resazurin-based cytotoxicity assay for hepatic cells. *Cell Biol Toxicol* **2002,** *18* (3), 157-173.

39. O'Brien, J.; Wilson, I.; Orton, T.; Pognan, F., Investigation of the Alamar Blue (resazurin) fluorescent dye for the assessment of mammalian cell cytotoxicity. *European Journal of Biochem* **2003,** *267*, 5421-5426.

40. Keisu, M.; Andersson, T. B., Drug-induced liver injury in humans: the case of ximelagatran. *Handb Exp Pharmacol* **2010,** (196), 407-18.

41. Sakatis, M. Z.; Reese, M. J.; Harrell, A. W.; Taylor, M. A.; Baines, I. A.; Chen, L.; Bloomer, J. C.; Yang, E. Y.; Ellens, H. M.; Ambroso, J. L.; Lovatt, C. A.; Ayrton, A. D.; Clarke, S. E., Preclinical strategy to reduce clinical hepatotoxicity using in vitro bioactivation data for >200 compounds. *Chem Res Toxicol* **2012,** *25* (10), 2067-82.

42. Schadt, S.; Simon, S.; Kustermann, S.; Boess, F.; McGinnis, C.; Brink, A.; Lieven, R.; Fowler, S.; Youdim, K.; Ullah, M.; Marschmann, M.; Zihlmann, C.; Siegrist, Y. M.; Cascais, A. C.; Di Lenarda, E.; Durr, E.; Schaub, N.; Ang, X.; Starke, V.; Singer, T.; Alvarez-Sanchez, R.; Roth, A. B.; Schuler, F.; Funk, C., Minimizing DILI risk in drug discovery - A screening tool for drug candidates. *Toxicol In Vitro* **2015,** *30* (1 Pt B), 429-37.

43. Bell, C. C.; Dankers, A. C. A.; Lauschke, V. M.; Sison-Young, R.; Jenkins, R.; Rowe, C.; Goldring, C. E.; Park, K.; Regan, S. L.; Walker, T.; Schofield, C.; Baze, A.; Foster, A. J.; Williams, D. P.; van de Ven, A. W. M.; Jacobs, F.; Houdt, J. V.; Lahteenmaki, T.; Snoeys, J.; Juhila, S.; Richert, L.; Ingelman-Sundberg, M., Comparison of Hepatic 2D Sandwich Cultures and 3D Spheroids for Long-term Toxicity Applications: A Multicenter Study. *Toxicol Sci* **2018,** *162* (2), 655-666.

44. Kamalian, L.; Chadwick, A. E.; Bayliss, M.; French, N. S.; Monshouwer, M.; Snoeys, J.; Park, B. K., The utility of HepG2 cells to identify direct mitochondrial dysfunction in the absence of cell death. *Toxicol In Vitro* **2015,** *29* (4), 732-40.

45. Kamalian, L.; Douglas, O.; Jolly, C. E.; Snoeys, J.; Simic, D.; Monshouwer, M.; Williams, D. P.; Park, B. K.; Chadwick, A. E., Acute Metabolic Switch Assay Using Glucose/Galactose Medium in HepaRG Cells to Detect Mitochondrial Toxicity. *Curr Protoc Toxicol* **2019,** *80* (1), e76.

46. Sison-Young, R. L.; Lauschke, V. M.; Johann, E.; Alexandre, E.; Antherieu, S.; Aerts, H.; Gerets, H. H. J.; Labbe, G.; Hoet, D.; Dorau, M.; Schofield, C. A.; Lovatt, C. A.; Holder, J. C.; Stahl, S. H.; Richert, L.; Kitteringham, N. R.; Jones, R. P.; Elmasry, M.; Weaver, R. J.; Hewitt, P. G.; Ingelman-Sundberg, M.; Goldring, C. E.; Park, B. K., A multicenter assessment of single-cell models aligned to standard measures of cell health for prediction of acute hepatotoxicity. *Arch Toxicol* **2017,** *91* (3), 1385-1400.

47. Ogese, M. O.; Jenkins, R. E.; Adair, K.; Tailor, A.; Meng, X.; Faulkner, L. L.; Enyindah, B. O.; Schofield, A.; Diaz-Nieto, R.; Ressel, L.; Eagle, G. L.; Kitteringham, N. R.; Goldring, C. E.; Park, B. K.; Naisbitt, D. J.; Betts, C., Exosomal transport of hepatocyte-derived drug-modified proteins to the immune system. *Hepatology* **2019**.

48. Shah, F.; Leung, L.; Barton, H. A.; Will, Y.; Rodrigues, A. D.; Greene, N.; Aleo, M. D., Setting Clinical Exposure Levels of Concern for Drug-Induced Liver Injury (DILI) Using Mechanistic in vitro Assays. *Toxicol Sci* **2015,** *147* (2), 500-14.

49. Gaskell, H.; Sharma, P.; Colley, H. E.; Murdoch, C.; Williams, D. P.; Webb, S. D., Characterization of a functional C3A liver spheroid model. *Toxicol Res (Camb)* **2016,** *5* (4), 1053-1065.

50. Kenna, J. G.; Taskar, K. S.; Battista, C.; Bourdet, D. L.; Brouwer, K. L. R.; Brouwer, K. R.; Dai, D.; Funk, C.; Hafey, M. J.; Lai, Y.; Maher, J.; Pak, Y. A.; Pedersen, J. M.; Polli, J. W.; Rodrigues, A. D.; Watkins, P. B.; Yang, K.; Yucha, R. W.; International Transporter, C., Can Bile Salt Export Pump Inhibition Testing in Drug Discovery and Development Reduce Liver Injury Risk? An International Transporter Consortium Perspective. *Clin Pharmacol Ther* **2018,** *104* (5), 916-932.

51. Edling, Y.; Sivertsson, L.; Andersson, T. B.; Porsmyr-Palmertz, M.; Ingelman-Sundberg, M., Pro-inflammatory response and adverse drug reactions: mechanisms of action of ximelagatran on chemokine and cytokine activation in a monocyte in vitro model. *Toxicol In Vitro* **2008,** *22* (6), 1588-94.

52. Ploj, K.; Benthem, L.; Kakol-Palm, D.; Gennemark, P.; Andersson, L.; Bjursell, M.; Borjesson, J.; Karrberg, L.; Mansson, M.; Antonsson, M.; Johansson, A.; Iverson, S.; Carlsson, B.; Turnbull, A.; Linden, D., Effects of a novel potent melanin-concentrating hormone receptor 1 antagonist, AZD1979, on body weight homeostasis in mice and dogs. *Br J Pharmacol* **2016,** *173* (18), 2739-51.

53. Morgan, P.; Brown, D. G.; Lennard, S.; Anderton, M. J.; Barrett, J. C.; Eriksson, U.; Fidock, M.; Hamren, B.; Johnson, A.; March, R. E.; Matcham, J.; Mettetal, J.; Nicholls, D. J.; Platz, S.; Rees, S.; Snowden, M. A.; Pangalos, M. N., Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat Rev Drug Discov* **2018,** *17* (3), 167-181.