Research Article

A Tiered, Bayesian Approach to Estimating Population Variability for Regulatory Decision-Making

Weihsueh A. Chiu[1], *Fred A. Wright*[2] *and Ivan Rusyn*[1]

[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA; [2]Bioinformatics Research Center and Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh, NC, USA

Summary

Characterizing human variability in susceptibility to chemical toxicity is a critical issue in regulatory decision-making, but is usually addressed by a default 10-fold safety/uncertainty factor. Feasibility of population-based *in vitro* experimental approaches to more accurately estimate human variability was demonstrated recently using a large (~1000) panel of lymphoblastoid cell lines. However, routine use of such a large population-based model poses cost and logistical challenges. We hypothesize that a Bayesian approach embedded in a tiered workflow provides efficient estimation of variability and enables a tailored and sensible approach to selection of appropriate sample size. We used the previously collected lymphoblastoid cell line *in vitro* toxicity data to develop a data-derived prior distribution for the uncertainty in the degree of population variability. The resulting prior for the toxicodynamic variability factor (the ratio between the median and 1% most sensitive individuals) has a median (90% CI) of 2.5 (1.4-9.6). We then performed computational experiments using a hierarchical Bayesian population model with lognormal population variability with samples sizes of *n* = 5 to 100 to determine the change in precision and accuracy with increasing sample size. We propose a tiered Bayesian strategy for fit-for-purpose population variability estimates: (1) a default using the data-derived prior distribution; (2) a pilot experiment using samples sizes of ~20 individuals that reduces prior uncertainty by > 50% with > 80% balanced accuracy for classification; and (3) a high confidence experiment using sample sizes of ~50-100. This approach efficiently uses *in vitro* data on population variability to inform decision-making.

Keywords: *in vitro*, variability, uncertainty, Bayesian

1 Introduction

The growing list of chemical substances in commerce and the complexity of exposures in the environment present enormous challenges for ensuring safety while promoting innovation. Because of the limitations of the current human and animal data-centric paradigm of chemical hazard and risk assessment in terms of cost, time, and throughput, the next generation of human health assessments has no choice but to use the information on chemical structure and from molecular and cell-based assays (NAS, 2007). Combined with the ever-increasing power of modern biomedical research tools to probe biological effects of chemicals at finer and finer resolutions, 21st century toxicology is taking shape (Tice et al., 2013; Kavlock et al., 2012).

In addition to addressing only a fraction of the chemicals in commerce, current hazard testing approaches usually do not take into account the genetic diversity within populations, overlooking uncertainties about how genetic variability might interact with environmental exposures to affect risk (Rusyn et al., 2010). As a result, while characterization of human variability in susceptibility to chemical toxicity is a critical issue in toxicology, public health, and risk assessment, it is usually addressed by a generic 10-fold safety/uncertainty factor despite encouragement to generate and use chemical-specific data (WHO/IPCS, 2005). The recent use of population-based animal *in vivo* (Rusyn et al., 2010; Chiu et al., 2014) and human *in vitro* (Abdo et al., 2015a,b; Eduati et al., 2015; Lock et al., 2012) experimental models that incorporate genetic diversity provides an opportunity to more precisely estimate human variability and increase confidence in decision-making. The technical feasibility and the scientific and practical value of large-scale *in vitro* population-based experimental approaches to more accurately estimate human

Received August 25, 2016;
Accepted December 9, 2016;
Epub December 13, 2016;
doi:10.14573/altex.1608251

variability, thereby avoiding the use of animals, has been firmly established in experiments with hundreds of single chemicals (Abdo et al., 2015b), as well as with several mixtures (Abdo et al., 2015a). Such an experimental approach fills a critical gap in large-scale *in vitro* toxicity testing programs, providing quantitative estimates of human toxicodynamic variability and generating testable hypotheses about the molecular mechanisms that may contribute to inter-individual variation in responses to particular agents. However, it is not feasible or practical to employ *in vitro* screening for population variability using thousands of cell lines to test thousands of chemicals and an infinite number of mixtures and real-life environmental samples.

Two approaches are possible to address the challenges in cost and effort of embedding population variability into large-scale *in vitro* testing programs. One solution is to develop computational models based on the already collected data, either to predict susceptibility to chemicals based on the constitutional genetic make-up of an individual or to forecast which chemicals may be most prone to eliciting widely divergent responses in a human population. Indeed, the large-scale population based *in vitro* toxicity data of Abdo et al. (2015b) enabled development of an *in silico* approach to predicting individual- and population-level toxicity associated with unknown compounds (Eduati et al., 2015). This exercise showed that *in silico* models that produced predictions which were statistically significantly better than random could be developed, but the correlations were modest for individual cytotoxicity response and only somewhat better for population-level responses, consistent with predictive performances for complex genetic traits. A second solution is to devise a tiered experimental strategy, flagging compounds with greater-than-default variability that may benefit from additional testing to more fully characterize the extent of a population-wide response.

Here we hypothesized that a Bayesian approach embedded in a tiered workflow will enable one to efficiently estimate population variability, and to sequentially determine the number of individuals needed to provide sufficiently accurate variability estimates. The acceptable degree of uncertainty in population variability differs depending on the risk assessment decision-making context as well as other sources of uncertainty. Our approach combines a data-derived default and Bayesian estimation of uncertainty to provide sufficient flexibility to develop fit-for-purpose estimates of human toxicodynamic variability as part of broader, more generic decision-making frameworks (e.g., Keisler and Linkov, 2014). This approach avoids the use of animals to fill a critical need for decision-making, and also provides a template to minimize sample sizes that can be applied to reduction in the use of animals, both of which are in keeping with the 3R concept of Russel and Burch (1959).

2 Materials and methods

2.1 Population cytotoxicity data and measures of toxicodynamic variability

The chemicals, cell lines, and cytotoxicity assays were previously described in Abdo et al. (2015b). Briefly, concentration-response data consisted of intracellular ATP concentrations evaluated 40 h after treatment with 170 unique chemicals at concentrations from 0.33 nM to 92 µM in lymphoblastoid cell lines from 1,086 individuals. Data were collected in 6 batches, and included some within-batch and some between batch replicates, with a total of 1-5 replicates for each chemical/cell-line combination. In all there were 351,914 individual concentration-response profiles, each consisting of 8 concentrations of a chemical in a specific cell line. The data were renormalized so that 0 corresponded to control levels (number of cells in each well) and -100 corresponded to a maximal response (complete loss of viable cells).

For each chemical and individual, the EC$_{10}$ (concentration associated with a 10% decline in viability) was used as an indicator of a toxicodynamic response. The variation across individuals in the EC$_{10}$ was then used as an indicator of population variability in the toxicodynamic response. Specifically, the *toxicodynamic variability factor at 1%* (TDVF$_{01}$) is defined as the ratio of the EC$_{10}$ for the median individual (EC$_{10,50}$) to the EC$_{10}$ for the most sensitive 1st percentile individual (EC$_{10,01}$): TDVF$_{01}$ = $_{\text{EC10,50}}$ / EC$_{10,01}$. Additionally, we define the *toxicodynamic variability magnitude* (TDVM) as the base 10 logarithm of the TDVF: TDVM = log$_{10}$(TDVF), or TDVF = 10^{TDVM}. The default fixed uncertainty factor for toxicodynamic variability is $10^{½}$ (WHO/IPCS, 2005), or half an order of magnitude, corresponding to TDVF = 3.16 and TDVM = ½.

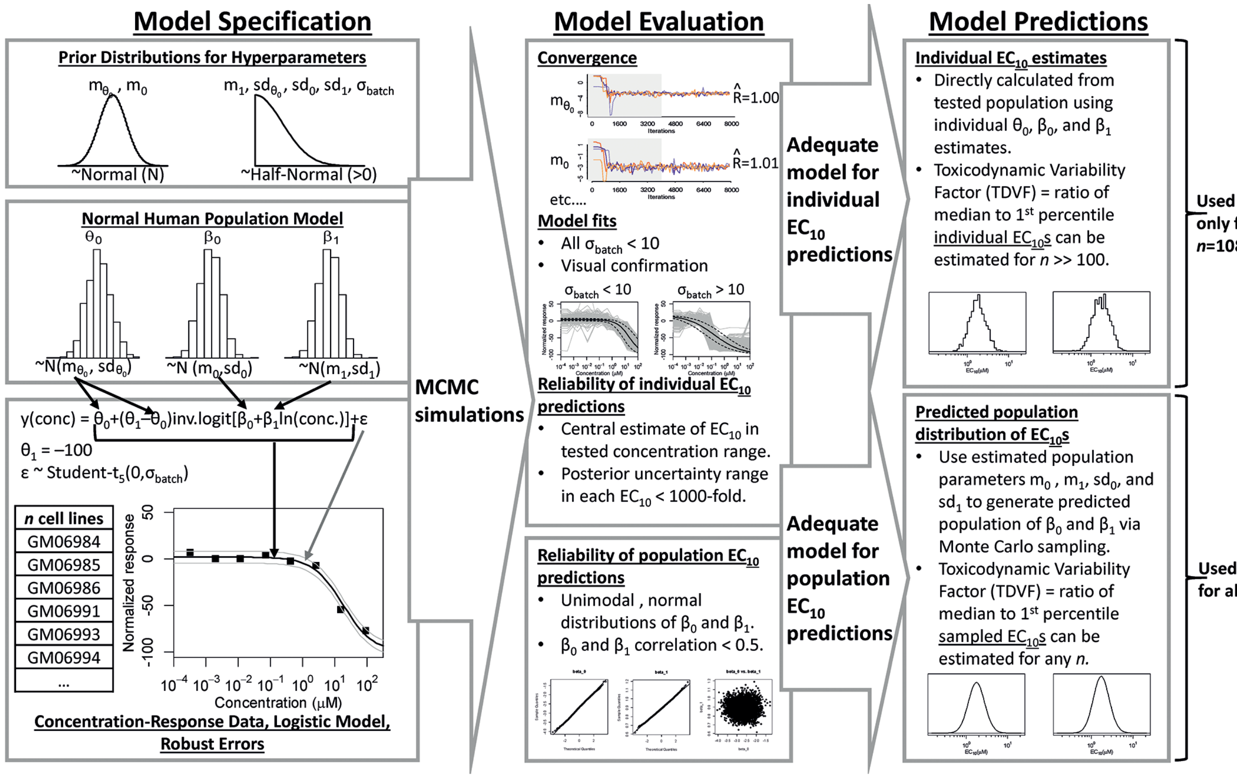
2.2 Estimating population variability for each chemical using a Bayesian approach

Abdo et al. (2015b) used maximum likelihood to fit a logistic model to each concentration-response dataset, averaging EC$_{10}$ estimates across replicates to estimate the EC$_{10}$ of each individual, with TDVF$_{01}$ estimated by the ratio between the median individual's EC$_{10}$ and the 1% most sensitive individual's EC$_{10}$, using a simple correction for measurement-related sampling variation based on the variation among replicates. This method of using the sample quantiles to estimate TDVF$_{01}$ is subject to increasing sample variation for sample sizes much less than ~1000, and is not feasible for smaller sample sizes < 100 (the 1% most sensitive EC$_{10}$ would not be part of the sample). However, if using a parametric distributional fit, then in principle any quantile can be estimated, along with (importantly) the uncertainty in this estimation.

Hierarchical Bayesian methods provide a natural approach to this type of challenge. The TDVF can be viewed as following a random effects model, with underlying parameters estimated using a Bayesian approach. Specifically, these methods allow for a multi-level structure in which individual-level parameters are viewed as drawn from a distribution governed by hyperparameters. Various levels of uncertainty can then be quantified and described through posterior distributions. The modelling workflow is shown in Figure 1.

2.2.1 Bayesian concentration-response modeling for each chemical

The first step in the workflow (Fig. 1) is specifying the statistical and concentration-response model that will be applied for each dataset. We analyzed each chemical separately, combining all



Model Specification
Prior Distributions for Hyperparameters
m_{θ_0}, m_0
m_1, sd_{θ_0}, sd_0, sd_1, σ_{batch}
~Normal (N)
~Half-Normal (>0)
Normal Human Population Model
θ_0
β_0
β_1
~N(m_{θ_0}, sd_{θ_0})
~N (m_0,sd_0)
~N(m_1,sd_1)
y(conc) = θ_0+(θ_1−θ_0)inv.logit[β_0+$\beta_1$ln(conc.)]+ε
θ_1 = −100
ε ~ Student-t_5(0,σ_{batch})
n cell lines
GM06984
GM06985
GM06986
GM06991
GM06993
GM06994
...
Normalized response
50
0
-50
-100
10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2}
Concentration (μM)
Concentration-Response Data, Logistic Model, Robust Errors
MCMC simulations
Model Evaluation
Convergence
m_{θ_0}
$\hat{R}$=1.00
m_0
$\hat{R}$=1.01
0 1600 3200 4800 6400 8000
Iterations
etc....
Model fits
All σ_{batch} < 10
Visual confirmation
σ_{batch} < 10
σ_{batch} > 10
Reliability of individual EC_{10} predictions
Central estimate of EC_{10} in tested concentration range.
Posterior uncertainty range in each EC_{10} < 1000-fold.
Reliability of population EC_{10} predictions
Unimodal , normal distributions of β_0 and β_1.
β_0 and β_1 correlation < 0.5.
beta_0
beta_1
beta_0 vs. beta_1
Adequate model for individual EC_{10} predictions
Adequate model for population EC_{10} predictions
Model Predictions
Individual EC_{10} estimates
Directly calculated from tested population using individual θ_0, β_0, and β_1 estimates.
Toxicodynamic Variability Factor (TDVF) = ratio of median to 1st percentile individual EC_{10}s can be estimated for n >> 100.
Used only for n=1086
Predicted population distribution of EC_{10}s
Use estimated population parameters m_0 , m_1, sd_0, and sd_1 to generate predicted population of β_0 and β_1 via Monte Carlo sampling.
Toxicodynamic Variability Factor (TDVF) = ratio of median to 1st percentile sampled EC_{10}s can be estimated for any n.
Used for all n
EC_{10}(μM)

Fig. 1: Bayesian modeling, evaluation, and prediction workflow

the datasets that used the same chemical. Thus, for each chemical, each dataset *j* corresponds to a particular individual *i*[*j*] and batch *b*[*j*] . For each dataset *j*, the concentration-response data are assumed to follow a logistic model, as was assumed previously by Abdo et al. (2015b). Recognizing the issue of outliers, we assume that deviations between the data and the model follow a Student's t-distribution instead of a normal distribution (Bell and Huang, 2006). Specifically, the logistic model we used is (see bottom left panel, Fig. 1):

$$y_j(x_{jk}) = \theta_{0,j} + (\theta_{1,j} - \theta_{0,j})\ \text{inv.logit}(\beta_{0,i[j]} + \beta_{1,i[j]}\ x_{jk}) + \varepsilon_{jk}$$

$$x_{jk} = \ln(\text{concentration}_{jk})$$

$$\varepsilon_{jk} \sim T_5(0,\sigma_{b[j]})$$

$$\text{inv.logit}(u) = \exp(u)/[1+\exp(u)]$$

$\theta_{1,j}$ was assigned a fixed value of -100 because many chemicals did not reach a maximal response in the dose range used, and in these instances θ_1 could not be reliably estimated from the data. In addition, $T_5(0,\sigma)$ denotes a Student's t-distribution with 5 degrees of freedom, centered on 0, with scale parameter σ. Recognizing potential differences across batches, we allowed σ to vary across each batch *b*. We fixed the Student's t-degrees of freedom at 5 based on examining residuals from preliminary fits to the data. Additionally, the EC_{10} for an individual *i* is given by

$$EC_{10,I} = \exp([\ln(0.1/0.9) - \beta_{0,i}\ ]/\beta_{1,i}),$$

which depends only on $\beta_{0,i}$ and $\beta_{1,i}$.

Prior distributions were specified as follows (middle and top left panels, Fig. 1). Parameter $\theta_{0,j}$ was estimated separately for each dataset *j*, as there was some apparent drift in the normalization, with a normal prior distribution across datasets *j* of $\theta_{0,j} \sim N(m_{\theta 0}, sd_{\theta 0})$. We assumed a normal population distribution across individuals *i* for $\beta_{0,i} \sim N(m_0,sd_0)$ and $\beta_{1,i} \sim N(m_1,sd_1)$, with respective means m_0 and m_1 and standard deviations sd_0 and sd_1. We used normal priors with wide variances for m_0 and $m_{\theta 0}$, and half-normal priors with wide variances for m_1, sd_0, sd_1, $sd_{\theta 0}$, and σ (restricted to being positive).

The joint posterior distribution of the parameters φ given the data D is equal to

$$P(\varphi|D) = P(\varphi)\ P(D|\varphi)/P(D)$$

Here, P(φ) is the prior distribution of the individual model parameters and hyperparameters, P(D|φ) is the likelihood, and P(D) can be treated as a normalization factor.

2.2.2 Model computation, convergence, and evaluation

Although our model is straightforward, the fitting and computation of posteriors cannot be feasibly performed deterministically, and so the posterior distribution was sampled using the Markov Chain Monte Carlo (MCMC) algorithm (large left arrow, Fig. 1) implemented in the software package Stan version 2.6.2 (Gelman et al., 2015). Computations were performed in the Texas A&M University high performance computing cluster with four MCMC chains run per chemical. Evaluation of the model performance had several components as follows (middle panels, Fig. 1).

Convergence was assessed using the potential scale reduction factor R (Gelman and Rubin, 1992), which compares inter- and intra-chain variability. Values >> 1 indicate poor convergence, and asymptotically approach 1 as the MCMC chain converges. Parameters with values of $R \leq 1.2$ are considered converged.

The model fit was evaluated in three ways. First, because some chemicals showed very little response at the concentrations tested, the model could not confidently estimate an EC_{10}. A large value for the scale parameter for the error term σ is an indicator of poor model fit, so chemicals were dropped if the median estimate for any of the $\sigma \geq 10$. Additionally, chemicals were dropped if (a) the EC_{10} for the median individual was outside the tested concentration range or (b) more than 1% of the individual EC_{10} estimates had a 90% confidence range ≥ 1000-fold.

The model also estimates the overall population distribution of EC_{10} values, so it is necessary to check the fit at the population and not just the individual level. Specifically, we checked the assumption that β_0 and β_1 are unimodal, normally distributed, and independent. For unimodality, we used Hartigans' dip test (Hartigan and Hartigan, 1985); for normality, we visually examined quantile-quantile plots; and for independence, we required that the correlation coefficient among posterior samples be < 0.5 (i.e., an R^2 of < 0.25).

2.2.3 Model predictions

At the individual level, the model predicts posterior distributions of β_0 and β_1 for each individual, which can be used to estimate uncertainty in each individual's EC_{10} (upper right panel, Fig. 1). Note that these estimates are already corrected for measurement errors. Thus, these EC_{10} values can be used to derive an individual-based estimate for $TDVF_{01}$ as long as the number of individuals (n_{indiv}) ≥ 100, because we are using the 1st percentile. This approach is not feasible for $n_{\text{indiv}} < 100$ because the 1st percentile is not part of the sample. However, in the hierarchical Bayesian model, population predictions can still be made using the estimated values of the population-level parameters rather than the individual-level parameters (lower right panel, Fig. 1). In this case, the model predicts a posterior distribution for the population parameters m_0, m_1, sd_0, and sd_1, from which a virtual population of β_0 and β_1 can be generated via Monte Carlo sampling. Because the posterior distributions of m_0, m_1, sd_0, and sd_1 are also sampled (via MCMC), this is in essence a two-dimensional Monte Carlo, separately evaluating uncertainty and variability. The specific procedure to derive this population-based estimate for $TDVF_{01}$ is as follows:

(1) Uncertainty loop – randomly sample populations $l = 1\ldots10^3$ from the posterior distribution of m_0, m_1, sd_0, and sd_1. Thus, the distribution across l represents the uncertainty in the population means and standard deviations.

(2) Variability loop – for a given set $m_{0,l}$, $m_{1,k}$, $sd_{0,l}$, and $sd_{1,l}$. draw $i = 1\ldots10^5$ individual pairs of $\beta_{0,l,i} \sim N(m_{0,l},sd_{0,l})$ and $\beta_{1,l,i} \sim N(m_{1,l},sd_{1,l})$. Thus each l,i pair represents an individual i (representing variability) drawn from the population l (representing uncertainty).

(3) For each individual i, calculate the predicted $EC_{10,l,i} = \exp([\ln(0.1/0.9) - \beta_{0,l,i}]/\beta_{1,l,i})$. The distribution of $EC_{10,l,i}$ over i (for fixed l) is the variability in the EC_{10} for population l.

(4) For each population l, calculate the $TDVF_{01,l}$, which is the ratio of the median to the 1% quantile of the $EC_{10,l,i}$.

(5) The distribution of $TDVF_{01,l}$ over 1 reflects the uncertainty in the degree of variability in the population.

2.3 Data-derived prior distribution for population variability in toxicodynamics (default distribution)

2.3.1 Coverage of chemical space

The first element of the tiered approach is the development of a data-derived prior distribution for population variability. The principle behind such a default distribution is the assumption that the chemicals for which data were previously collected are sufficiently representative of chemical space so that a new chemical can be reasonably considered a random draw from the same distribution. To check this assumption, the chemicals examined by Abdo et al. (2015b) were compared with the over 32,000 chemicals in the CERAPP dataset (Mansouri et al., 2016), a virtual chemical library that has undergone stringent chemical structure processing and normalization for use in QSAR modeling. Chemical structures were mapped to chemical property space using DRAGON descriptors (DRAGON 6, http://www.talete.mi.it/help/dragon_help/), as implemented in ChemBench (Walker et al., 2010).

2.3.2 Deriving the default distribution

The default distribution was estimated using the individual-based estimates for $TDVF_{01}$. Specifically, for each of the chemicals for which the individual EC_{10} estimates for the individual cell lines tested were considered reliable, the median EC_{10} estimate of each of the 1086 individuals was used to construct the population variability distribution for that chemical. The $TDVF_{01}$ for each chemical is the ratio between the median and 1% quantile of the 1086 individual EC_{10} estimates. The individual-based estimate of $TDVF_{01}$ was chosen to represent the default distribution because it is less dependent on model assumptions, and thus was reliably estimated for more chemicals. Additionally, it is most similar to the approach used by Abdo et al. (2015b), the only difference being the method of estimating the individual EC_{10} values.

The default distribution across chemical-specific $TDVF_{01}$ values was fit to a lognormal distribution in $TDVM_{01} = \log_{10}$

$TDVF_{01}$ (i.e., ln ($\log_{10}$ $TDVF_{01}$) is fit to a normal distribution). This choice of distribution was motivated by several considerations. First, because $TDVF_{01}$ is restricted to be > 1, this implies $TDVM_{01}$ is restricted to be > 0, and the lognormal distribution is a natural choice for strictly positive values. Additionally, previous analyses of *in vivo* human data found toxicokinetic and toxicodynamic variability to be consistent with such a distribution (Hattis et al., 2002; WHO/IPCS, 2014). This choice was further checked using the Shapiro-Wilk test for normality (Royston, 1995) with a p-value threshold of 0.05.

The sensitivity of the resulting default distribution to the above choices was assessed in three ways: (1) using individual-based estimates for the chemicals that passed the model fit test at both the population level as well as the individual level; (2) using population-based estimates instead of individual-based estimates; and (3) leaving one chemical out at a time.

2.4 Computational experiments with smaller sample sizes

In order to characterize the added value as a function of sample size, sub-samples of individuals with $n_{\text{indiv}} = 5, 10, 20, 50$, and 100 were drawn for each chemical, and the $TDVF_{01}$ was re-estimated using the smaller sample. Ten different replicate sub-samples were drawn for each value of n_{indiv}. Because only population-based estimates of $TDVF_{01}$ are feasible for these sample sizes, the computational experiments were restricted to the chemicals that had reliable population-based predictions.

Additionally, two estimates of $TDVF_{01}$ were derived for each experiment. The first estimate used the same Bayesian modeling workflow used to derive the data-derived prior distribution, including the same prior distributions for the model parameters. This posterior distribution is denoted data distribution because it is based largely on the chemical-specific data, as the priors are broad enough to be unrestrictive as to the value of $TDVF_{01}$. A second default+data distribution estimate is derived based on combining the data distribution with the default distribution derived from the full dataset. This approach essentially treats the default distribution as a Bayesian prior for $TDVF_{01}$, in which case the default+data distribution is the appropriate Bayesian posterior for $TDVF_{01}$.

The accuracy and precision of the default, data, and default+data distributions were evaluated in two illustrative types of prediction: classification and estimation. "Classification" involves separating chemicals into two bins of high or low variability, defined as having $TDVF_{01}$ > or < than the median value from the default distribution. Different percentiles of each distribution were used as estimators of $TDVF_{01}$ (e.g., 5th percentile, median, 95th percentile) to reflect different tolerances for false positives and negatives. The rates of true/false positives and negatives were compiled as a function of sample size n_{indiv}, assuming the estimate based on 1086 individuals was the "true" value. The results were summarized in a Receiver Operating Characteristic (ROC) curve, balanced accuracy, and AUC. "Estimation" involves providing a numerical value for a chemical's $TDVF_{01}$. The accuracy of each distribution as a function of sample size *n* was evaluated by comparing each median prediction with the true value assumed to be the median estimate based on 1086 individuals, and quantified in terms of the slope and intercept of a linear regression. Because the uncertainty in $TDVM_{01} = \log_{10}(TDVF_{01})$ was found to be approximately lognormally distributed, the linear regression was performed on ln($TDVM_{01}$). The precision of each distribution was quantified in terms of the R^2 of the linear regression as well as the geometric standard deviation of $TDVM_{01}$. The degree of uncertainty was compared using the corresponding log-transformed variance var(ln $TDVM_{01}$) = (ln GSD_{TDVM})^2.

2.5 Software

MCMC computations and analyses of the convergence diagnostic R were performed with Stan version 2.6.2. The Stan statistical model code is included in the supplementary file[1]. All other statistical analyses were performed using R version 3.1.1.

3 Results

3.1 Estimating population variability for each chemical

For most chemicals, convergence was reached for all parameters with a chain length of 8000, where the first 4000 "warmup" samples of each chain were discarded, and the final 4000 samples were used for evaluation of convergence, model fit, and inference. If a chemical had not achieved convergence for all parameters after chain lengths of 128,000 (64,000 warmup), it was dropped due to poor convergence. 138 of the original 170 chemicals passed both convergence checks as well as checks related to model fit. For these chemicals, individual EC_{10} estimates for the individual cell lines tested were considered reliable. The 32 chemicals that failed these checks are listed in Table S1[1], along with the rationale for their exclusion.

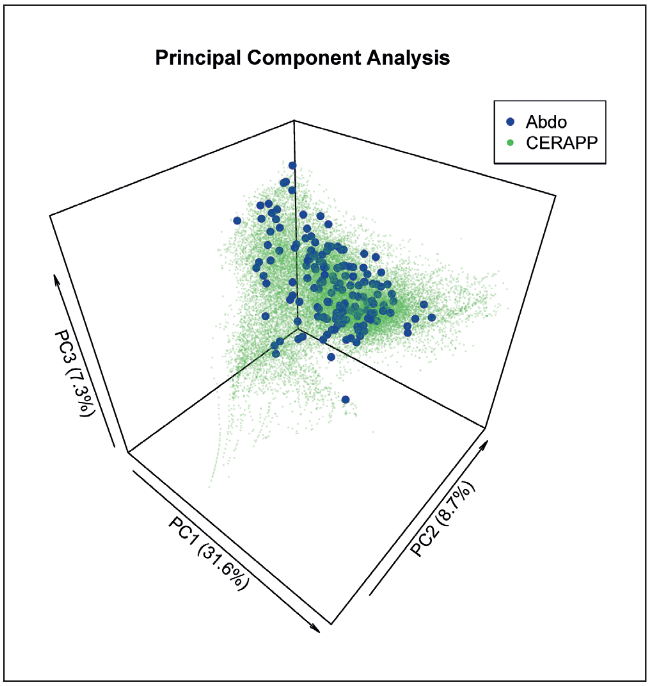
119 of the original 170 chemicals also passed these additional checks related to normality and unimodality of the population distribution. For these chemicals, population EC_{10} estimates (e.g., individuals generated via Monte Carlo) were also considered reliable. The 19 chemicals with reliable individual-based estimates but less than reliable population-based estimates are listed in Table S2[1], along with the rationale for their exclusion and the individual-based $TDVF_{01}$ estimate. The 119 chemicals with reliable individual- and population-based estimates are listed in Table S3[1], along with both $TDVF_{01}$ estimates.

3.2 Default distribution for population variability in toxicodynamics

3.2.1 Coverage of chemical space

Figure 2 shows a visualization of the overlap between the Abdo and CERRAP chemicals, using the first three principal components in chemical property space (which account for 48% of the variance). Quantitatively, using Euclidean distance in chemical

[1] doi:10.14573/altex.1608251s



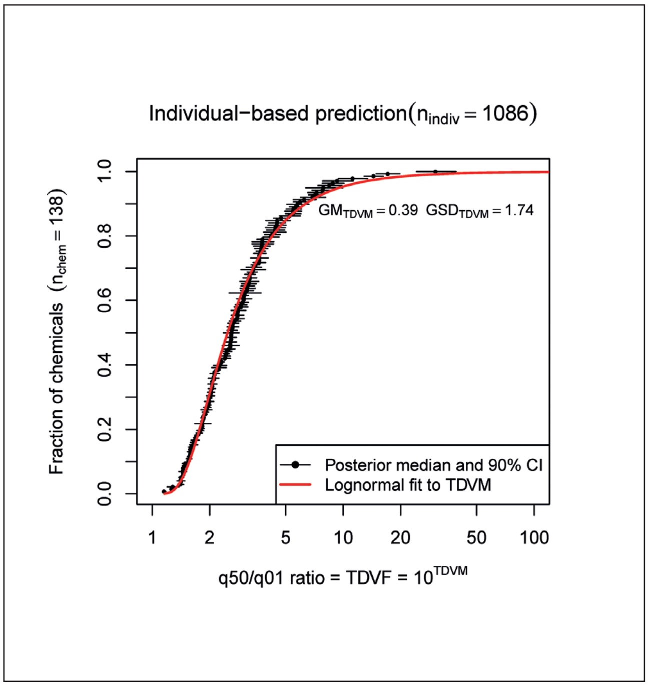


Fig. 2: Chemical space coverage of 170 chemicals from Abdo et al. (2015b) as compared to > 32,000 chemicals in the CERAPP chemical library (Mansouri et al., 2016), based on principal component analysis of chemical descriptors

Fig. 3: Default distribution for toxicodynamic variability factor $TDVF_{01}$ based on cytotoxicity profiling, contrasted with a lognormal fit to $TDVM_{01} = \log_{10}(TDVF_{01})$

Tab. 1: Default distribution for population variability in toxicodynamics based on cytotoxicity profiling and Bayesian modeling

Analysis	***n_{chem}***	**Lognormal distribution of $TDVM_{01}$**	**Distribution for $TDVF_{01}$** Median (90% CI)

Individual-based estimates, larger dataset	138	GM = 0.39
		GSD = 1.74
			2.48 (1.44, 9.57)

Alternative with population-based estimates, smaller dataset	119	GM = 0.37
		GSD = 1.74
			2.36 (1.41, 8.52)

Alternative with individual-based estimates, smaller dataset	119	GM = 0.37
		GSD = 1.73
			2.35 (1.41, 8.15)

Range of alternatives with individual-based estimates, leave-one-out, full dataset	137	GM = 0.39-0.40
		GSD = 1.70-1.74
			2.46-2.51 (1.44-1.47, 9.04-9.67)

Note: *The toxicodynamic variability factor at 1%* ($TDVF_{01}$) is defined as ratio of the EC_{10} for the median person ($EC_{10,50}$) to the EC_{10} for the more sensitive 1st percentile person ($EC_{10,01}$): $TDVF_{01} = EC_{10,50} / EC_{10,01}$. The toxicodynamic variability magnitude (TDVM) is the base 10 logarithm of the TDVF: TDVM = $\log_{10}$(TDVF), or TDVF = 10^{TDVM}.

space as a measure of similarity (Zhu et al., 2009), greater than 97% of the CERAPP chemicals (Mansouri et al., 2016) are within 3 standard deviations of the nearest neighbor distances across the Abdo chemicals. Thus, the Abdo et al. (2015b) chemicals represent a highly representative dataset from which to derive a data-derived prior distribution for population variability. For shorthand, this distribution is denoted the default distribution.

3.2.2 Default distribution

Using the 138 chemicals with reliable individual-based EC_{10} estimates, the distribution of $TDVF_{01}$ estimates ranged from 1.15 to 30.4, with a median of 2.6. As hypothesized, the distribution across chemicals of $TDVM_{01} = \log_{10}(TDVF_{01})$ was consistent with a lognormal distribution by the Shapiro-Wilk test ($p = 0.32$). The distribution of $TDVF_{01}$ estimates, along with the lognormal fit to $TDVM_{01}$, are shown in Figure 3. The parameters for the fit distribution, as well as their sensitivity to alternative methods, are shown in Table 1. As is evident from these results, alternative analyses lead to very small changes in the resulting default distribution of toxicodynamic variability. Therefore, the default distribution using individual-based estimates for the larger dataset of 138 chemicals was considered robust and was used in subsequent analyses.

3.3 Computational experiments with smaller sample sizes

A total of 5950 computational experiments were run, comprising 119 chemicals, five values of n_{indiv} (5, 10, 20, 50, and 100), and 10 replicates each. Figure 4 illustrates two typical results from these computational experiments. In each panel, the curves represent the Bayesian distributions for $TDVF_{01}$ based on (1) only the data, (2) only the default, and (3) the combined data+default. The chemical in the left panel has high variability, and the chemical in the right panel has low variability. Three key results are as follows:

- At small values of n_{indiv}, the data distribution is wider than the default distributions, indicating that the chemical-specific data provide a less precise estimate of toxicodynamic variability than do data on other chemicals. In the case of a high variability chemical, the precision of the estimate based on $n_{indiv} = 5$ is orders of magnitude worse than the precision of the default. This is to be expected in a Bayesian context where informative prior information (here derived from large experiments with many chemicals) can outweigh a small amount of new data. Only at n_{indiv}~20 does the data begin to have comparable precision to the default.
- The Bayesian approach of combining the data and default leads to estimates that are both more accurate (with less bias) and more precise (with narrower confidence intervals), even at small values of n_{indiv}. Even for n_{indiv} as small as 5, the median of the data+default distribution is closer to the true value estimated for $n_{indiv} = 1086$ than the data distribution. Additionally, by combining the two distributions, the resulting estimate is also more precise, as is evident from the narrower width of the data+default distributions.
- In the case of a low variability chemical, the concordance between data and data+default is higher, presumably because it is closer to the median across all chemicals (i.e., more similar to the prior).

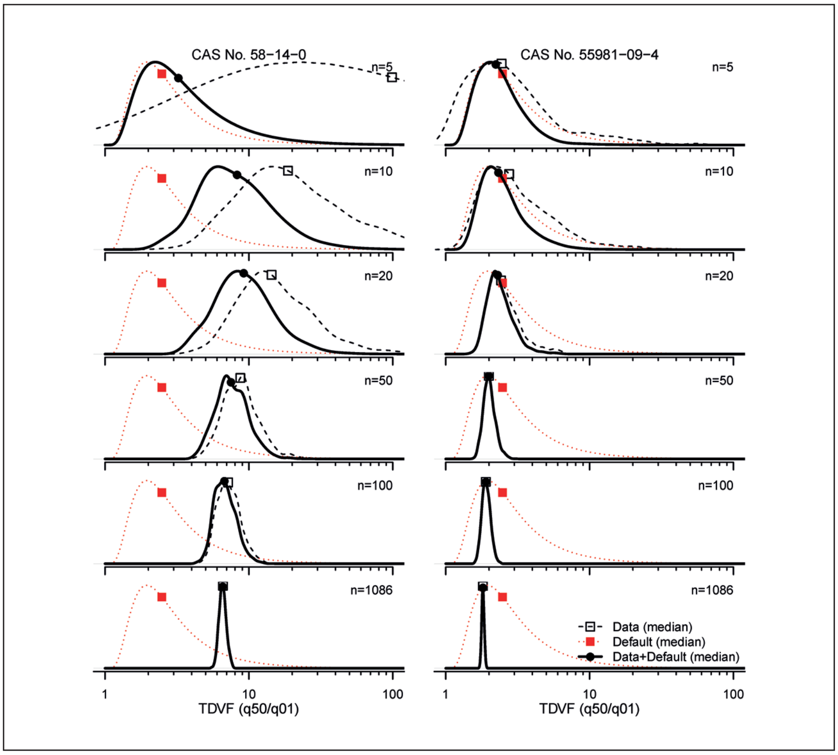


Fig. 4: Illustration of the effect of sample sizes on Bayesian estimates of the toxicodynamic variability factor $TDVF_{01}$ based on cytotoxicity profiling

The chemical on the left has high variability and the chemical on the right has low variability. For each chemical and sample size (n = 5 … 1086), three distributions reflecting uncertainty in the value of $TDVF_{01}$ are shown, along with the median estimate. The data distribution is the estimate based only on the chemical-specific data using the Bayesian workflow illustrated in Figure 1. The default distribution is the estimate without using any chemical-specific data, but assuming the chemical is randomly drawn from the distribution of chemicals as shown in Figure 3. The data+default distribution is the result of combining these distributions, i.e., the Bayesian posterior distribution treating the default as a prior.

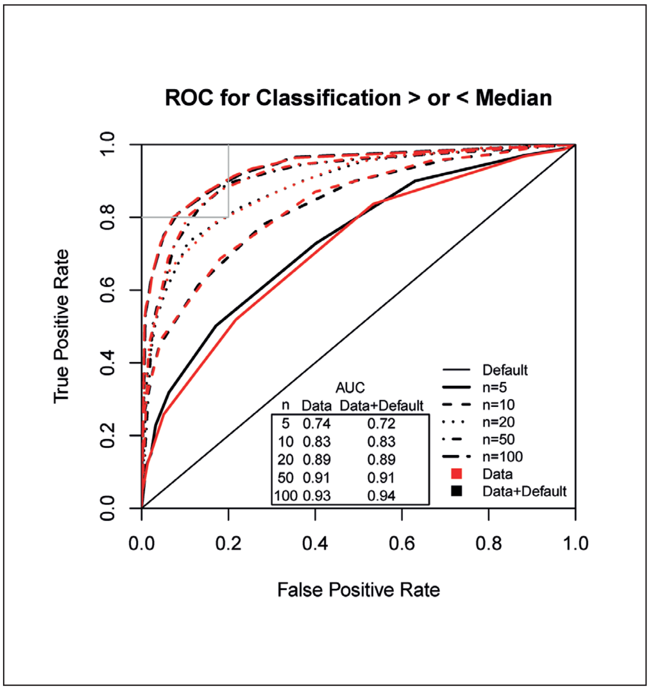


Fig. 5: Receiver Operative Characteristic (ROC) curve and AUCs for classifying a chemical as having high or low population variability, defined by > or < median of the default

These results hold generally across all the chemicals analyzed. Their implications are illustrated through two representative types of predictions: (i) classification of high/low variability chemicals and (ii) estimation of a chemical-specific $TDVF_{01}$.

3.3.1 Classification

The purpose of classification is to place one or more chemicals into bins of high and low population variability, and the key question is characterizing the rates of true/false positives and negatives. For illustration, we assume that the threshold is the median of the default distribution = 2.48, with the implication that without any chemical-specific information, there is a 50-50 chance of a correct classification. Figure 5 shows the ROC and corresponding AUC for different values of n_{indiv}. We find that a sample size of at least $n_{indiv} = 20$ is required to achieve both 80% specificity and 80% sensitivity, and even $n_{indiv} = 100$ cannot achieve 90% balanced accuracy.

The results for classification are similar whether the data or data+default distribution is used. This can be explained by noting that the result of data+default is simply to shrink the estimates toward the median, in comparison to using data alone. Because the classification is based on > or < the median, this shrinkage, while producing more accurate predictions, does not change the classification, leading to similar ROC curves.

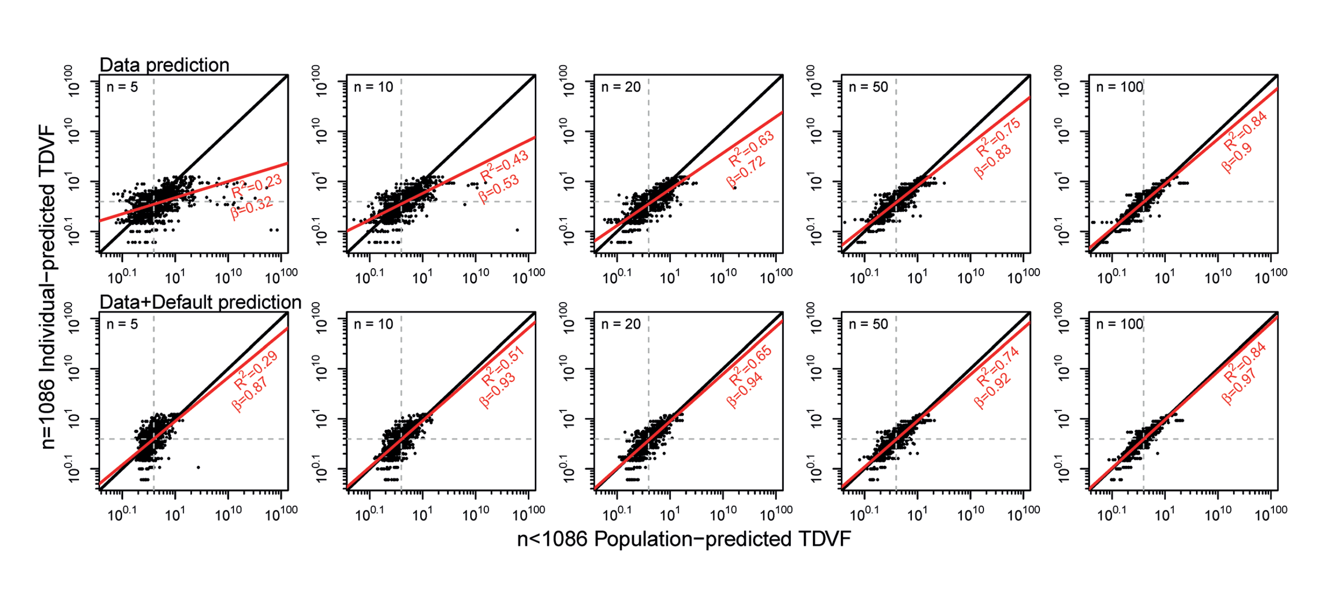


Fig. 6: Scatter plots (dots) and linear regression (red line) of predictions for n_{indiv} = 5…100 as compared to the predictions for n_{indiv} = 1086

The R^2 and slope β of the linear regression are also shown. The black line is the slope = 1 line, and the dotted lines separate the high and low values of variability with the median across chemicals as the cut-point. Note that the axis scales are double log transformed.

3.3.2 Chemical-specific toxicodynamic variability estimation

Estimates of a chemical-specific $TDVF_{01}$ can be used in calculating a human health toxicity value such as a reference dose. The key questions here are the accuracy and precision of the estimate as a function of sample size. These are illustrated graphically in Figure 6, which shows scatter plots of the predictions for n_{indiv} = 5…100 as compared to the predictions for n_{indiv} = 1086. Also shown are the linear regression lines on $\ln(TDVM_{01})$, along with slope β and R^2. Several key results are noteworthy:

- At all sample sizes, the data predictions have less accuracy (β further from 1) and less precision (smaller R^2) as compared to the data+default predictions. This bias tends to be positive for high variability chemicals and negative for low variability chemicals, as is evident from the regression line intersecting the β = 1 line at approximately the median value. This explains the similar ROC curves for data and data+default in Figure 5.
- In some cases, the values of the data predictions were extremely high, with absurdly unrealistic estimates of population variability, whereas the data+default predictions were much more reasonable due to the influence of the prior. For instance, at n_{indiv} = 10, across the 1190 computational experiments, the top 1% of the data predictions for $TDVF_{01}$ ranged from 1800 to 10^{61} (!), whereas the top 1% of the data+default predictions ranged from 16 to 31. This shows how unstable estimates of population variability are with small sample sizes in the absence of accounting for prior information.
- In all cases, for the same value of n_{indiv}, the data+default prediction had better precision, as evidenced by the higher R^2, as compared to the data prediction. This is further illustrated in Figure 7, which shows how the uncertainty in the $TDVF_{01}$ estimate decreases with increasing sample size. Only for $n_{indiv} \geq 20$ does the data prediction have an uncertainty smaller than the default at least 95% of the time. By contrast, at $n_{indiv} \geq 20$, in more than 99% of the data+default predictions the uncertainty is reduced 2-fold compared to the default, with a reduction in uncertainty of at least 5-fold 75% of the time.

Overall, for estimating chemical-specific toxicodynamic variability, the data+default predictions combining chemical-specific data with the default distribution as the prior are more accurate and more precise than the data predictions based on chemical-specific data alone. Additionally, the data+default predictions begin to provide substantial improvement over the default distribution alone at $n_{indiv} \geq 20$.

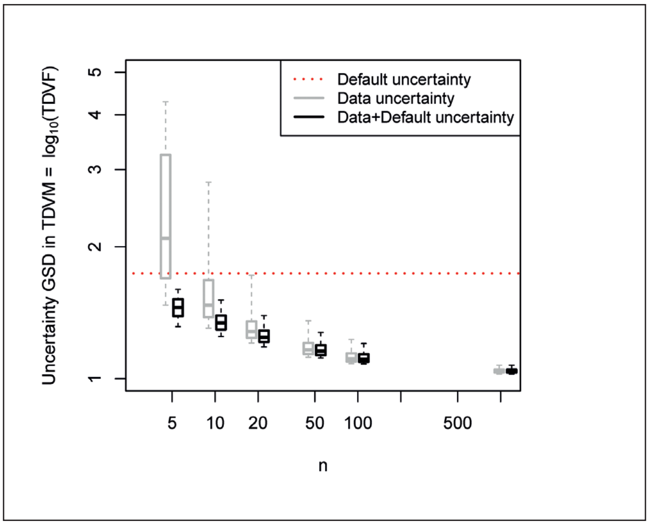


Fig. 7: Uncertainty in estimate of toxicodynamic variability as a function of sample size, comparing default (red dotted line), data (grey box plot), and data+default predictions (black box plot)

The box plots represent the median (horizontal bar), interquartile range (box), and 95%ile range (whiskers) across the 1190 computational experiments for each value of n_{indiv}. The measure of uncertainty shown is the posterior GSD of the $TDVM_{01} = \log_{10} TDVF_{01}$. For example, if the central estimate of $TDVF_{01} = 10^{½}$ = 3.16, then the central estimate of $TDVM_{01}$ = ½. If GSD of $TDVM_{01}$ = 1.5, then its 90% CI is ¼ to 1, which in turn implies the 90% CI of $TDVF_{01}$ is $10^{¼ \text{ to } 1}$ = 1.8 to 10.

4 Discussion

Our results provide scientific justification for a tiered experimental strategy applicable to fit-for-purpose population variability estimation in *in vitro* screening, as illustrated in Figure 8. The first tier relies on the default distribution derived from the large scale study of > 100 chemicals in > 1000 individual cell lines (Abdo et al., 2015b). We have demonstrated that the chemicals used to derive this distribution provide wide coverage of the chemical space occupied by the environmental and industrial compounds (Fig. 1), and that this default distribution is robust to multiple sensitivity analyses (Tab. 1). For many risk assessment applications and regulatory decisions, this default distribution may be deemed adequate, for example if margins of exposure are high or estimated health risks are low, even assuming a worst case of high variability. Moreover, although the default distribution is based on data from a single cell type, the resulting distribution is very similar to that based on available *in vivo* human data on toxicodynamic variability across a range of endpoints (Abdo et al., 2015b; WHO/IPCS, 2014). Therefore, as a default, this distribution is likely to be adequate regardless of the endpoint of interest.

In addition, further refinement of the prior distribution may be possible through chemo-informatic approaches – using chemical structure information to give greater weight to chemicals in the database that are more similar to the chemical of interest, such as incorporating the models reported in Eduati et al. (2015). Indeed, we found that the distance between chemicals in chemical property space (e.g., as in Fig. 1) has a small, but statistically significant, correlation (*r* = 0.12, *p* = 0.03, by the Mantel test (Mantel, 1967)) with the distance between chemicals in terms of their TDVF that is consistent with a small degree of clustering in chemical properties among chemicals with similar TDVF values.

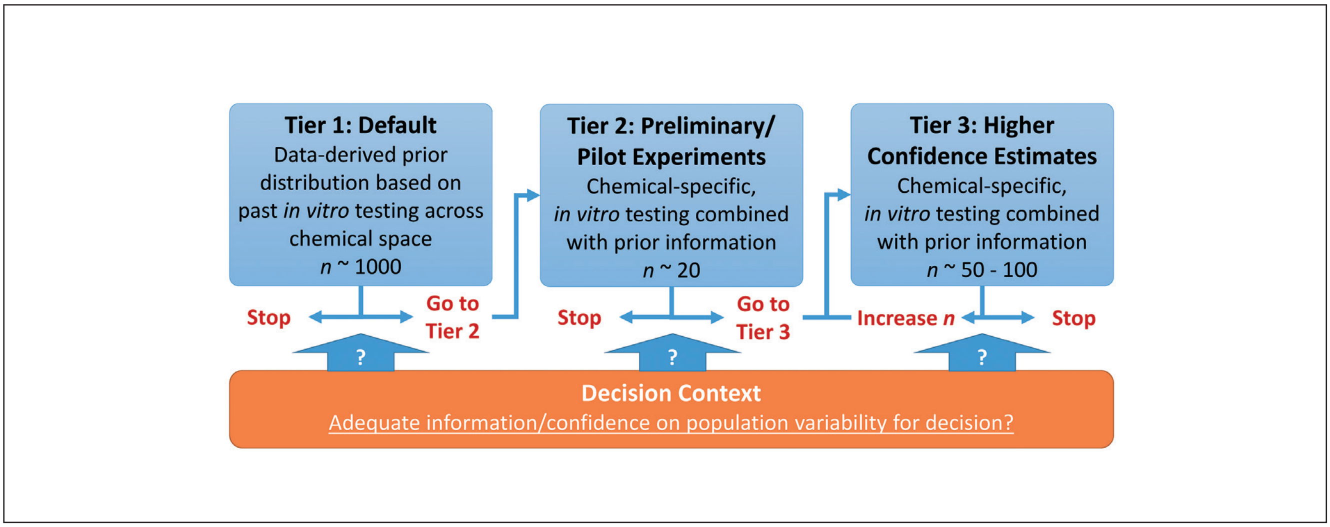

Tier 1: Default
Data-derived prior distribution based on past *in vitro* testing across chemical space
n ~ 1000
Tier 2: Preliminary/ Pilot Experiments
Chemical-specific, *in vitro* testing combined with prior information
n ~ 20
Tier 3: Higher Confidence Estimates
Chemical-specific, *in vitro* testing combined with prior information
n ~ 50 - 100
Stop
Go to Tier 2
Stop
Go to Tier 3
Increase *n*
Stop
?
?
?
Decision Context
Adequate information/confidence on population variability for decision?

Fig. 8: Tiered workflow using a Bayesian approach to estimate toxicodynamic population variability
After each tier, a decision is made as to whether the information from that tier is adequate to make the risk assessment decision.

The second tier focuses on preliminary or pilot experiments that provide a first level of refinement to the question of population variability. The results of this experiment could be used, for instance, to classify a large group of chemicals into bins of high and low population variability, or to provide a chemical-specific estimate of population variability for a particular substance or multiple agents at a reasonable cost. Based on the results with respect to accuracy and precision, sample sizes of ~20 individuals have > 80% balanced accuracy for classification and reduce prior uncertainty by > 50% for estimation. For classification, similar results are obtained from the estimates based on the data alone (using only chemical-specific data) or based on combining the data and default in a Bayesian manner. However, for quantitatively estimating a chemical-specific population variability, combining the data and default in a Bayesian manner results in a much more accurate and precise estimate. It is anticipated that this tier will address the needs of most of the risk assessment applications and regulatory decisions for which the default distribution alone is deemed to be inadequate.

The third tier is for deriving high confidence, chemical-specific estimates of toxicodynamic variability using sample sizes $n_{indiv} > 20$. While the difference between the data and data+default results are more similar for these sample sizes, the latter would be more consistent with the overall Bayesian framework. This tier may be repeated iteratively with progressively larger sample sizes until the information is considered adequate for the regulatory decision at hand.

One uncertainty in the experimental (i.e., second and third) tiers of this approach is possible differences among cell types. The experiments reported in (Abdo et al., 2015b), which form the basis of these analyses, were in lymphoblastoid cells, and the extent to which the degree of variability correlates across different cell types is not clear. Therefore, while the default distribution based on lymphoblastoid cells is likely to be adequate across endpoints, given the similarity with the distribution based on *in vivo* data across multiple endpoints, it is less clear that chemical-specific variability can be assessed using only one cell type. However, it is anticipated that induced pluripotent stem cell (iPSC)-based technologies will enable tissue-type specific analyses of population variability, and such experiments are already underway for iPSC-derived cardiomyocytes from individuals with familial cardiovascular syndromes (Chen et al., 2016).

The Bayesian approach illustrated in these analyses also naturally interfaces with an overall probabilistic approach to dose-response assessment, as advocated by the National Academy of Sciences and the World Health Organization International Program on Chemical Safety (NAS, 2009; WHO/IPCS, 2014). Specifically, by providing a distribution reflecting uncertainty in the degree of variability, the Bayesian estimates of TDVF_{01} can be used directly in the recent probabilistic framework developed by WHO/IPCS (2014) and summarized by Chiu and Slob (2015). This framework was developed as an extension of the current approach for deriving toxicity values, using uncertainty and variability distributions based on historical *in vivo* data. In particular, with respect to the factor for human variation, WHO/IPCS (2014) and Chiu and Slob (2015) argued that this probabilistic approach provides substantial added value in comparison with the usual 10-fold factor by explicitly quantifying both a "level of conservatism" (e.g., 90%, 95%, or 99% confidence) as well as a "level of protection" in terms of what residual fraction of the population may experience effects (e.g., 0.1%, 1%, or 5%). Thus, one consequence of the current 10-fold factor approach is that risk management judgements, such as the levels of confidence and protection, are hidden in the risk assessment, whereas a probabilistic approach that requires estimates such as the TDVF_{01} and its uncertainty allow for such judgments to be made transparent and explicit.

More broadly, the approach proposed here suggests that an overall Bayesian framework can substantially reduce required sample sizes, particularly if there is an existing database from which to derive informed prior distributions. The same three-tiered approach may indeed be applicable not only to other studies of population variability, but also perhaps other *in vitro* assays and *in vivo* studies as well, leading to a reduction in the number of animals used per experiment. This generic approach would consist of the following:

– Tier 1. Developing prior distributions through re-analysis of existing data that can be used in the absence of chemical-specific data. These distributions could not only replace the current explicit defaults (such as 10-fold safety factors), but also implicit defaults such as "no data=no hazard=no risk." This approach is consistent with recommendations from the NAS (2009) to replace current defaults with those based on the best available science.

– Tier 2. Developing a suite of preliminary or pilot experimental designs with smaller sample sizes that could provide an improvement in precision and accuracy over the default but with smaller sample sizes than current testing regimes. The ability to use smaller sample sizes rests on the Bayesian approach of combining the prior information with the chemical-specific information, thereby increasing overall accuracy and precision at a lower cost. Additionally, other alterations in study design and statistical analyses could make more efficient use of samples (e.g., designing studies with benchmark dose modeling in mind, rather than pairwise statistical tests) (Slob, 2014a,b).

– Tier 3. Only as a last resort would larger sample sizes like those traditionally used in toxicity testing be required.

A limitation of this approach is that developing an informative prior relies on the existence of a large dataset across chemicals. Even if it were not feasible to newly generate such a large dataset, it may be possible to mine existing databases, such as EPA's ACToR System (Judson et al., 2012).

In sum, we have demonstrated that a Bayesian approach embedded in a tiered workflow enables one to reduce the number of individuals needed to estimate population variability. A key component of this approach is using the existing database of large sample size experiments across are large number of chemicals to develop an informed prior distribution for the extent of toxicodynamic population variability. For many applications, this prior distribution may well be adequate for decision-making, so no additional experiments may be needed. In cases where this default distribution is too uncertain, experiments with modest sample sizes of ~20 individuals can, if combined with the prior, provide a substantial increase in accuracy and precision. Only for the rare cases where a high confidence estimate is required would larger samples sizes of up to ~100 individuals be used. Based on these results, we suggest that a tiered Bayesian approach may be more broadly useful in toxicology and risk assessment.

References

Abdo, N., Wetmore, B. A., Chappell, G. A. et al. (2015a). In vitro screening for population variability in toxicity of pesticide-containing mixtures. *Environ Int 85*, 147-155. doi:10.1016/j.envint.2015.09.012

Abdo, N., Xia, M., Brown, C. C. et al. (2015b). Population-based in vitro hazard and concentration-response assessment of chemicals: The 1000 genomes high-throughput screening study. *Environ Health Perspect 123*, 458-466. doi:10.1289/ehp.1408775

Bell, W. R. and Huang, E. T. (2006). Using the t-distribution to deal with outliers in small area estimation. In *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*. Ottawa, ON, Canada: Statistics Canada.

Chen, I. Y., Matsa, E. and Wu, J. C. (2016). Induced pluripotent stem cells: At the heart of cardiovascular precision medicine. *Nat Rev Cardiol 13*, 333-349. doi:10.1038/nrcardio.2016.36

Chiu, W. A., Campbell, J. L., Jr., Clewell, H. J., 3rd et al. (2014). Physiologically based pharmacokinetic (PBPK) modeling of interstrain variability in trichloroethylene metabolism in the mouse. *Environ Health Perspect 122*, 456-463. doi:10.1289/ehp.1307623

Chiu, W. A. and Slob, W. (2015). A unified probabilistic framework for dose-response assessment of human health effects. *Environ Health Perspect 123*, 1241-1254. doi:10.1289/ehp.1409385

Eduati, F., Mangravite, L. M., Wang, T. et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nat Biotechnol 33*, 933-940. doi:10.1038/nbt.3299

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. 457-472. doi:10.1214/ss/1177011136

Gelman, A., Lee, D. and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics 40*, 530-543. doi:10.3102/1076998615606113

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *Ann Statist 13*, 70-84. doi:10.1214/aos/1176346577

Hattis, D., Baird, S. and Goble, R. (2002). A straw man proposal for a quantitative definition of the RfD. *Drug Chem Toxicol 25*, 403-436. doi:10.1081/dct-120014793

Judson, R. S., Martin, M. T., Egeghy, P. et al. (2012). Aggregating data for computational toxicology applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. *Int J Mol Sci 13*, 1805-1831. doi:10.3390/ijms13021805

Kavlock, R., Chandler, K., Houck, K. et al. (2012). Update on EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management. *Chem Res Toxicol 25*, 1287-1302. doi:10.1021/tx3000939

Keisler, J. and Linkov, I. (2014). Environment models and decisions. *Environ Syst Decis 34*, 369-372. doi:10.1007/s10669-014-9515-4

Lock, E. F., Abdo, N., Huang, R. et al. (2012). Quantitative high-throughput screening for chemical toxicity in a population-based in vitro model. *Toxicol Sci 126*, 578-588. doi:10.1093/toxsci/kfs023

Mansouri, K., Abdelaziz, A., Rybacka, A. et al. (2016). CERAPP: Collaborative estrogen receptor activity prediction project. *Environ Health Perspect 124*, 1023-1033. doi:10.1289/ehp.1510267

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res 27*, 209-220.

NAS (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: National Academies Press.

NAS (2009). *Science and Decisions: Advancing Risk Assessment*. Washington, DC: National Academies Press.

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 44*, 547-551. doi:10.2307/2986146

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. London: Methuen. http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc

Rusyn, I., Gatti, D. M., Wiltshire, T. et al. (2010). Toxicogenetics: Population-based testing of drug and chemical safety in mouse models. *Pharmacogenomics 11*, 1127-1136. doi:10.2217/pgs.10.100

Slob, W. (2014a). Benchmark dose and the three Rs. Part I. Getting more information from the same number of animals. *Crit Rev Toxicol 44*, 557-567. doi:10.3109/10408444.2014.925423

Slob, W. (2014b). Benchmark dose and the three Rs. Part II. Consequences for study design and animal use. *Crit Rev Toxicol 44*, 568-580. doi:10.3109/10408444.2014.925424

Tice, R. R., Austin, C. P., Kavlock, R. J. and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: A Tox21 update. *Environ Health Perspect 121*, 756-765. doi:10.1289/ehp.1205784

Walker, T., Grulke, C. M., Pozefsky, D. and Tropsha, A. (2010). Chembench: A cheminformatics workbench. *Bioinformatics 26*, 3000-3001. doi:10.1093/bioinformatics/btq556

WHO/IPCS – World Health Organization International Program on Chemical Safety (2005). Chemical-Specific Adjustment Factors for Interspecies Differences and Human Variability: Guidance Document for Use of Data in Dose/Concentration-Response Assessment. http://apps.who.int/iris/bitstream/10665/43294/1/9241546786_eng.pdf

WHO/IPCS (2014). Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization. http://www.who.int/ipcs/methods/harmonization/uncertainty_in_hazard_characterization.pdf?ua=1

Zhu, H., Ye, L., Richard, A. et al. (2009). A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ Health Perspect 117*, 1257-1264. doi:10.1289/ehp.0800471

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge R. S. Thomas (USEPA), D. M. Reif (NCSU), and A. Gelman (Columbia U) for useful feedback on the approach developed in this manuscript. This publication was made possible by grants from the USEPA (STAR RD 83516602) and NIH (P30 ES023512). Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA or NIH. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Correspondence to

Weihsueh A. Chiu
Department of Veterinary Integrative Biosciences
Texas A&M University
4458 TAMU
College Station, TX 77843, USA
Phone: +1 979 845 4106
Fax: +1 979 847 8981
e-mail: wchiu@tamu.edu

Ivan Rusyn
Department of Veterinary Integrative Biosciences
Texas A&M University
4458 TAMU
College Station, TX 77843, USA
Phone: +1 979 458 9866
Fax: +1 979 847 8981
e-mail: irusyn@tamu.edu