Citation: CPT: Pharmacometrics & Systems Pharmacology (2013) 2, e25; doi:10.1038/psp.2012.25

www.nature.com/psp

ORIGINAL ARTICLE

Vertical Integration of Pharmacogenetics in Population PK/PD Modeling: A Novel Information Theoretic Method

J Knights[1,2], P Chanda[3], Y Sato[4], N Kaniwa[5], Y Saito[5], H Ueno[6], A Zhang[2] and M Ramanathan[1,2]

To critically evaluate an information-theoretic method for identifying gene–environmental interactions (GEI) associated with pharmacokinetic (PK), pharmacodynamic (PD), and clinical outcomes from genome-wide pharmacogenetic data. Our approach, which is built on the *K*-way interaction information (KWII) metric, was challenged with simulated data and clinical PK/PD data sets from the International Warfarin Pharmacogenetics Consortium (IWPC) and a gemcitabine clinical trial. The KWII efficiently identified both novel and known interactions for warfarin and gemcitabine. Interactions between herbal supplementation and *VKORC1* genotype were associated with warfarin response. For gemcitabine-associated neutropenia, combination treatment with carboplatin and cytidine deaminase (*CDA*) 208G→A genotypes were identified as risk factors. Gemcitabine disposition was associated with drug metabolism–transporter interactions between deoxycytidine kinase (DCK) and the equilibrative nucleoside transporter (ENT). This novel approach is effective for detecting GEI involved in drug exposure and response and could enable integration of genome-wide pharmacogenetic data into the population PK/PD analysis paradigm.

CPT: Pharmacometrics & Systems Pharmacology (2013) **2,** e25; doi:10.1038/psp.2012.25

Pharmacokinetic/pharmacodynamic (PK/PD) modeling is the dominant approach for dose and dosing regimen selection in drug discovery and development. It has the unique capability of incorporating constraints imposed by the underlying pharmacology/pathophysiology and offers a rich body of behaviors for modeling the relationship between drug dose and the time course of drug disposition and effect. The Food and Drug Administration requires submission of PK/PD data that frequently includes population PK/PD analyses during the approval process for new drugs.

The role of genetic and environmental factors for individualizing dosing is well established for some drugs. The Food and Drug Administration has approved diagnostic tests and labeling changes when evidentiary support for genetic testing has been defined in clinical trials, e.g., *CYP2C9* and *VKORC1* diagnostic tests for warfarin and *HLA-B*5701* testing for abacavir.[1–3] When the roles of specific candidate genetic variations in drug disposition and response are known, randomized clinical trials to test for pharmacogenetic effects can be easily designed.

However, for many drugs, the genetic factors may be complex and not characterized in advance. Pharmacogenetic and pharmacogenomic data obtained via high-throughput microarray and sequencing platforms could potentially provide critical insights into drug action and response variability in these situations. Nowadays, DNA is commonly collected from subjects enrolled in clinical trials for use in these analyses, and the Food and Drug Administration has issued widely disseminated white papers to encourage better utilization of both model-based approaches and pharmacogenomic data during drug development.

The clinical pharmacology and PK/PD research communities are still struggling to handle genome-wide genetic variation data effectively as there is a dearth of systematic methods for vertically integrating and leveraging such data into PK/PD modeling. The reasons are manifold and include interrelated contributions from the size and dimensionality of the data; lack of effective modeling strategies; and the high level of user intervention required by the existing tools.

In previous reports from our group, we have demonstrated the usefulness of the *K*-way interaction information (KWII) and phenotype-associated information for gene–environment interaction (GEI) analysis of discrete phenotypes and quantitative traits.[4–6] We developed efficient algorithms that leverage the computational properties of the phenotype-associated information metric to search and identify variable combinations involved in the strongest interactions.[4,5] The GEI identified using these methods can be leveraged in modeling efforts to identify the mechanisms underlying experimental and clinical outcomes; however, the use of these information-theoretic methods for GEI analysis of PK/PD and clinical outcomes has not been investigated.

The purpose of this paper was to critically evaluate our information-theoretic framework for identifying the key genetic variations, gene–gene interactions, and GEI contributing to PK/PD, and clinical outcomes of drugs. We demonstrate that these methods possess key capabilities for identifying covariates and risk factors from genome-wide pharmacogenomic data, which can be used to drive mechanistic modeling in clinical systems and population PK/PD.

[1]Department of Pharmaceutical Sciences, State University of New York, Buffalo, New York, USA; [2]Department of Computer Science and Engineering, State University of New York, Buffalo, New York, USA; [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA; [4]Division of Genetics, National Cancer Center Research Institute, Tokyo, Japan; [5]Division of Medicinal Safety Science, National Institute of Health Sciences, Tokyo, Japan; [6]Hepatobiliary and Pancreatic Oncology Division, National Cancer Center Hospital, Tokyo, Japan. Correspondence: M Ramanathan (Murali@Buffalo.Edu)
Received 12 November 2012; accepted 19 December 2012; advance online publication 6 February 2013. doi:10.1038/psp.2012.25

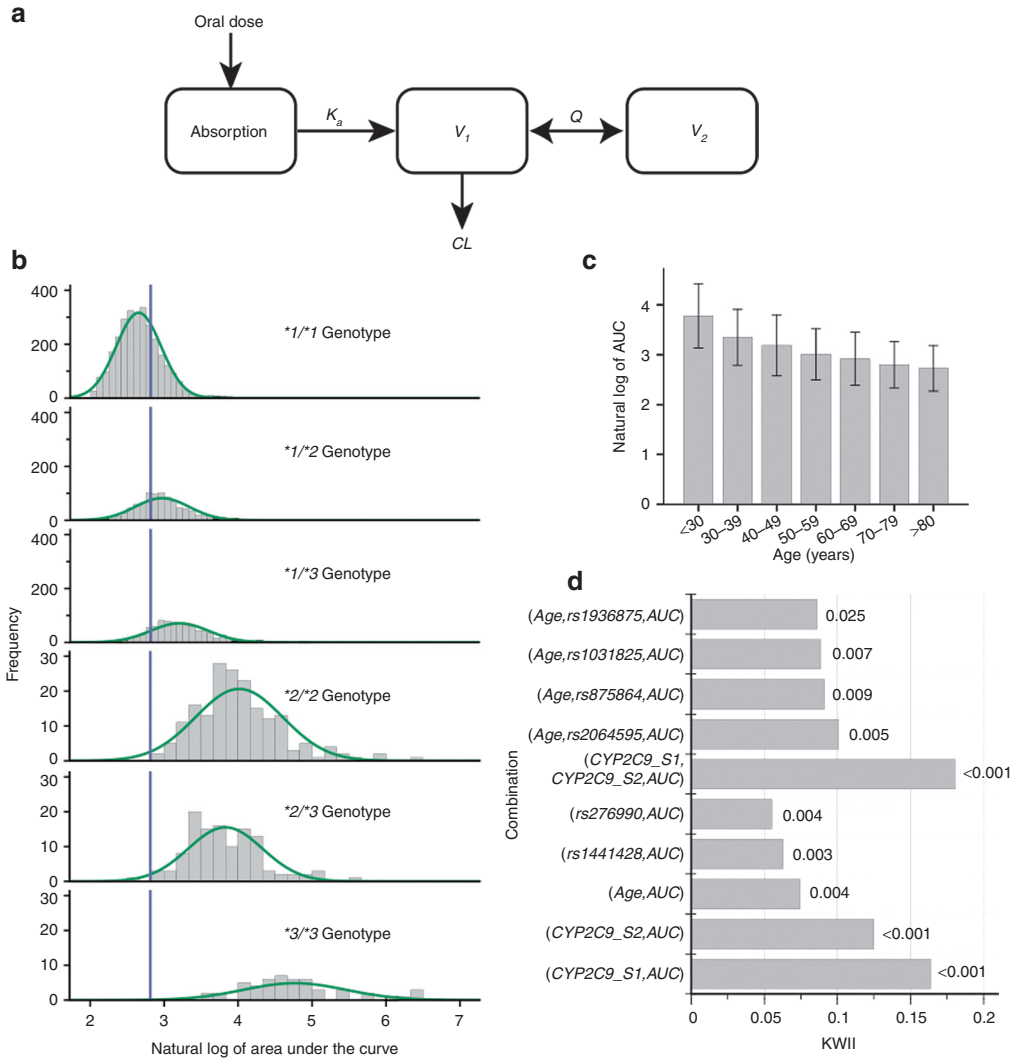

Figure 1 Simulation data. (**a**) The model that was used for simulations. The panels in (**b**) show the histograms for the natural logarithm of the area under the curve for the different *CYP2C9* genotypes. The simulations used a sample size of 500. However, the bottom three histograms in (**b**) were generated using a sample size of 5,000 for to enhance visual clarity of the lower frequency genotypes. Note that the scales for the less frequent **2/*2*, **2/*3*, and **3/*3* genotypes are different from those for the other genotypes. (**c**) The dependence of clearance on age in the simulations. (**d**) The top five first- and second-order combinations with the highest KWII values. The permutation-based *P* values are shown against each bar: "rs" labeled combinations represent the noisy SNPs added for complexity. AUC, area under the concentration–time curve; CL, clearance; K_a, first-order elimination rate constant; KWII, *K*-way interaction information; Q, first-order intercompartmental transfer rate constant; SNP, single-nucleotide polymorphism ; V_1, volume of distribution of the central compartment; V_2, volume of distribution of the peripheral compartment.

RESULTS

Warfarin PKs

Figure 1 shows the single-nucleotide polymorphism (SNP) interactions associated with warfarin area under the concentration–time curve (AUC) for the simulated warfarin PK data set. The largest two KWII peaks in **Figure 1d** (*CYP2C9_S1* and *CYP2C9_S2*, $P < 0.001$ for both) are the two SNPs known to correspond with reduced *CYP2C9* activities. The high KWII values represent the associations between these two SNPs and warfarin AUC. The *Age* variable had the third-highest KWII value ($P < 0.0001$).

The high KWI value for the second-order combination {*CYP2C9_S1*, *CYP2C9_S2*, *AUC*} indicates synergistic interactions between the SNPs. This is in line with our simulation as these two SNPs fully define the *CYP2C9* effect. This example demonstrates that our KWII analysis successfully identified all the informative factors included in the simulation for warfarin systemic exposure from among 1,004 potential predictors, including the second-order interaction between both *CYP2C9* SNPs.

We also analyzed individual clearance estimates from nonlinear mixed effect model program (NONMEM) using CHORUS. The results were concordant with the analysis of AUC. CHORUS successfully identified both *CYP2C9* SNPs and age as informative covariates of CL (data not shown), indicating that covariates for model parameters can also be detected using the approaches proposed. In addition, the signal-to-noise ratio for the CL analysis was qualitatively superior to the

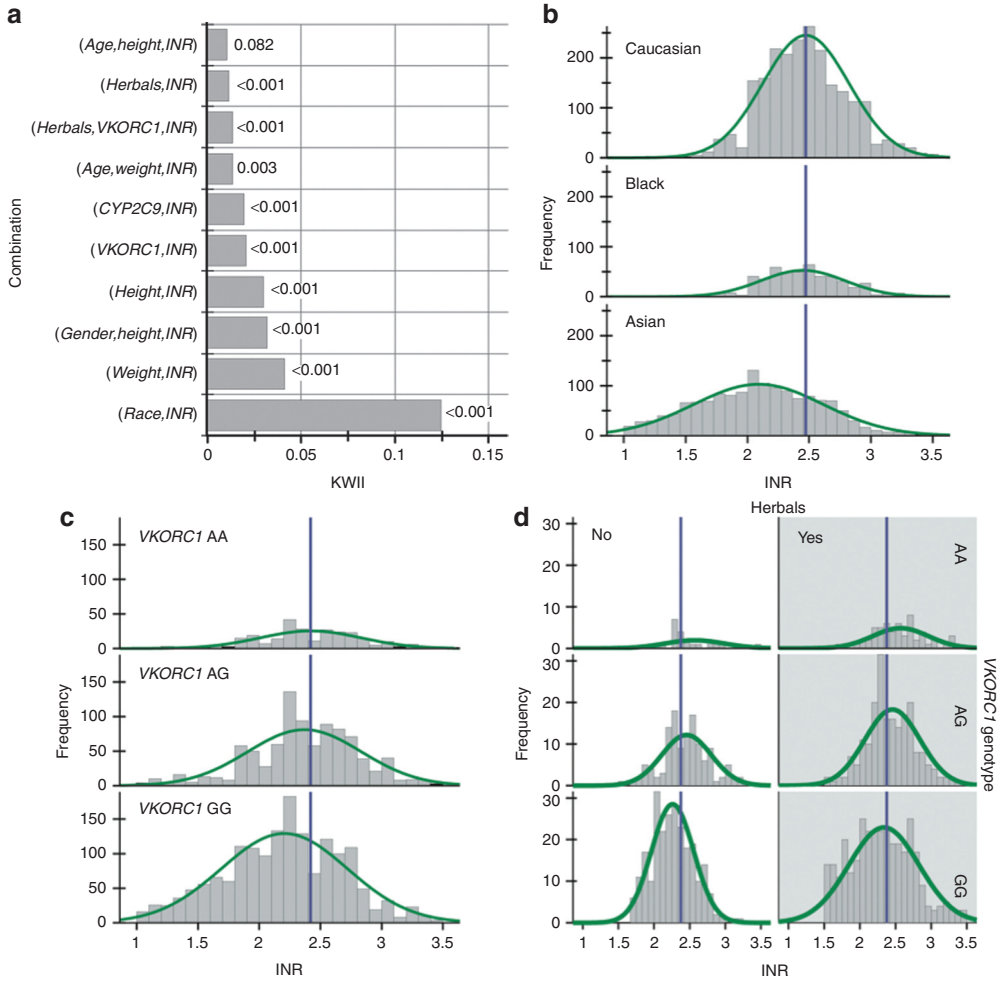

Figure 2 IWPC INR data. (**a**) Summarizes the top 10 combinations with the highest KWII values with the permutation-based *P* values against each bar. The histograms in (**b**,**c**) summarize the dependence of the INR distribution on the *Race* variable and the *VKORC1* (*rs7294*) genotype variable, respectively. The panel of histograms in (**d**) summarizes the second-order interaction of the INR distribution for different combinations of use of *Herbals* and *VKORC1* (*rs7294*) genotype. The corresponding normal distribution is overlaid on each histogram. The vertical lines in (**b**,**c**) correspond to the mean value of the histogram on the top panel; in (**d**), the vertical lines correspond to the sample mean. INR, international normalized ratio; KWII, *K*-way interaction information.

AUC analysis. For example, the KWII for the first-order combination {*Age, CL*} was roughly fourfold higher than the highest noisy SNP, whereas the KWII for the {*Age, AUC*} combination was ~18% higher than the highest noisy SNP.

Warfarin PDs

Warfarin response was assessed using international normalized ratio (INR) values from the International Warfarin Pharmacogenetics Consortium (IWPC) data set as the phenotype. Inspection of the data indicated that the normal distribution was a reasonable approximation for the distribution of INR in the IWPC data set (data not shown).

Figure 2a summarizes the top 10 first- and second-order combinations with the highest KWII values and their permutation-based *P* values. The KWII analysis indicates strong first-order associations with *Race*, *VKORC1* genotype, *CYP2C9* genotype, and use of *Herbals*. Strong second-order associations were found for *Gender, Height, and INR* and for *Herbals, VKORC1, and INR*.

We critically assessed the interactions identified by the KWII through direct examination of the data. **Figure 2b** highlights the shifting median INR values across the *Race* variable. The leftward shift in the bottom panel of **Figure 2b** shows that the Asian racial group (n = 1,505) has lower mean INR values than the Caucasian (n = 2,366) and Black (n = 497) racial groups. The histograms in **Figure 2c** demonstrate lower mean INR values in the group with *VKORC1* GG genotype. A representative second-order interaction, {*Herbals*, *VKORC1*, *INR*}, is highlighted in **Figure 2d**. The mean INR in the *VKORC1* GG group was modestly lower in the group that did not use herbals.

Comparisons to regression results. To further assess the results from the KWII analysis, we compared our findings to those from multiple linear regression. The regression model included the top nine first- and second-order combinations as shown in **Figure 2b**. In the regression analysis, the first-order interactions corresponding to *Race* (partial correlation coefficient r_p = –0.27, P < 0.001), *Height* (r_p = 0.05, P = 0.001),

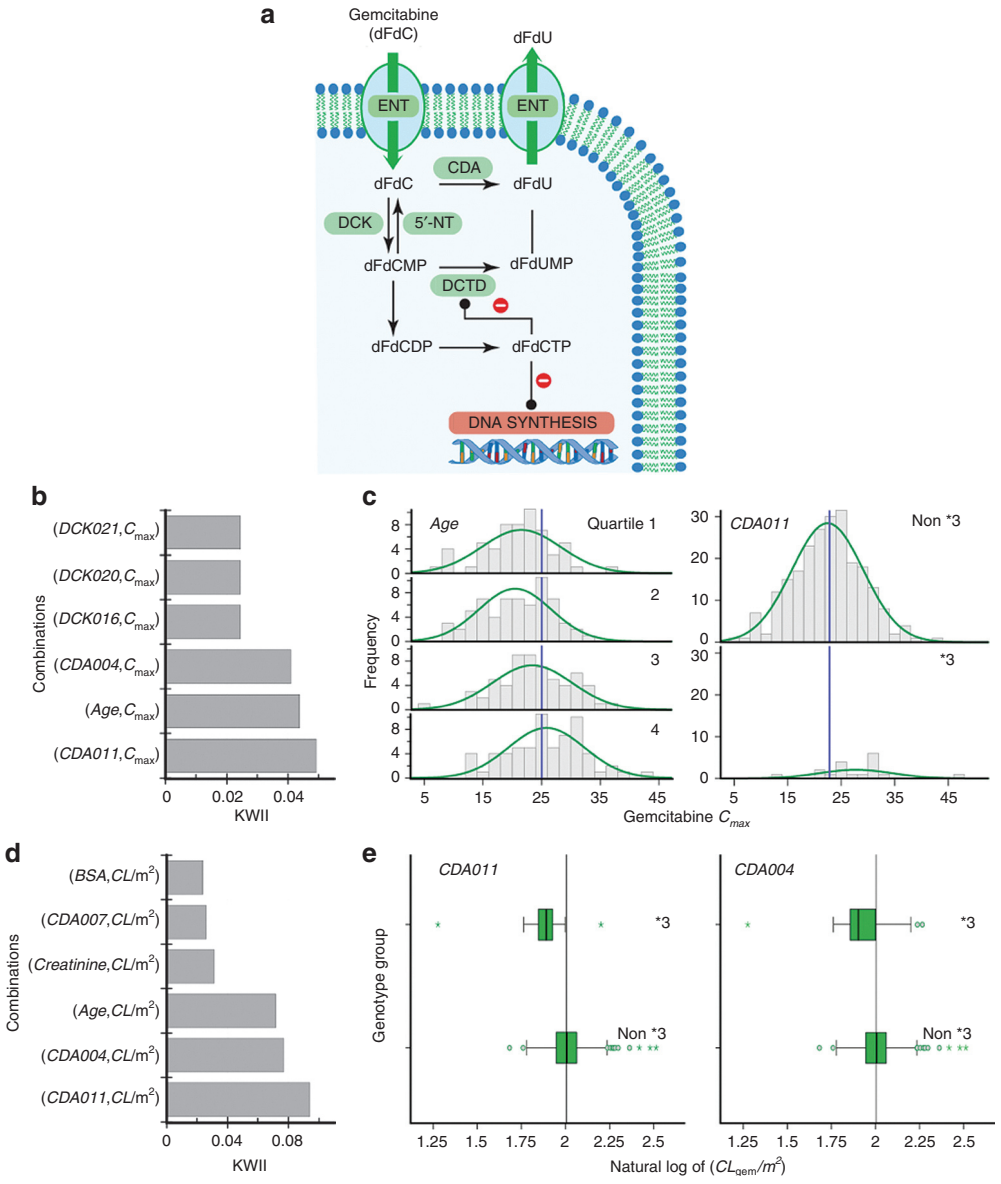

Figure 3 Gemcitabine PK. (**a**) The metabolic pathway and direct mechanism of DNA synthesis inhibition for gemcitabine (dFdC) and the primary metabolite studied (dFdU). (**b**) The KWII values for the top six predictors using maximum observed gemcitabine concentration (mg/l). The histograms in (**c**) display the observed distribution of concentrations across quartiles of observed age (left) and genotypes of the *CDA* 208G→A (*CDA011*) SNP (right). (**d**) The KWII values for the top six predictors using the natural logarithm of the individual gemcitabine clearance values normalized to body surface area. The box-plots in (**e**) summarize the natural-log transformed values of clearance for the individual genotypes of the *CDA011* (left) and *CDA-116G→A* (*CDA004*, right) SNPs. The reference lines represent the observed population mean for the individual predictors. 5′ NT, 5′ nucleotidase; CDA, cytidine deaminase; CL, clearance; CL_{gem}, gemcitabine clearance; C_{max}, peak plasma concentration; DCK, deoxycytidine kinase; DCTD, deoxycytidylate deaminase; dFdCTP, 2',2'-difluorodeoxycytidine triphosphate; ENT, equilibrative nucleoside transporter; KWII, *k*-way interaction information; PK, pharmacokinetic; SNP, single-nucleotide polymorphism.

Herbals (r_p = 0.03, P = 0.033), and *CYP2C9* genotype (r_p = 0.029, P = 0.042) were significant. The second-order interactions between Gender and *Height* (r_p = –0.046, P = 0.001) and between *Weight* and *Age* (r_p = –0.032, P = 0.026) were also significant. We did not find evidence for associations with *Weight* (P = 0.084), *VKORC1* genotype (P = 0.44), or the interaction between *Herbals* and *VKORC1* genotype (P = 0.65). This comparison demonstrates that the KWII method is broadly concordant with multiple linear regression. The discrepancies indicate that the information-theoretic framework identifies novel candidate interactions that are not detected by regression.

Gemcitabine PKs

Figure 3a is a schematic of the metabolism and transport pathways and the mechanism of action for gemcitabine (dFdC) and its metabolite (dFdU). **Figure 3b,d** show the KWII spectra for the top first-order interactions associated with the maximum observed concentration of gemcitabine (C_{max} mg/l), and with gemcitabine clearance (CL_{gem} l/h/m²). The corresponding

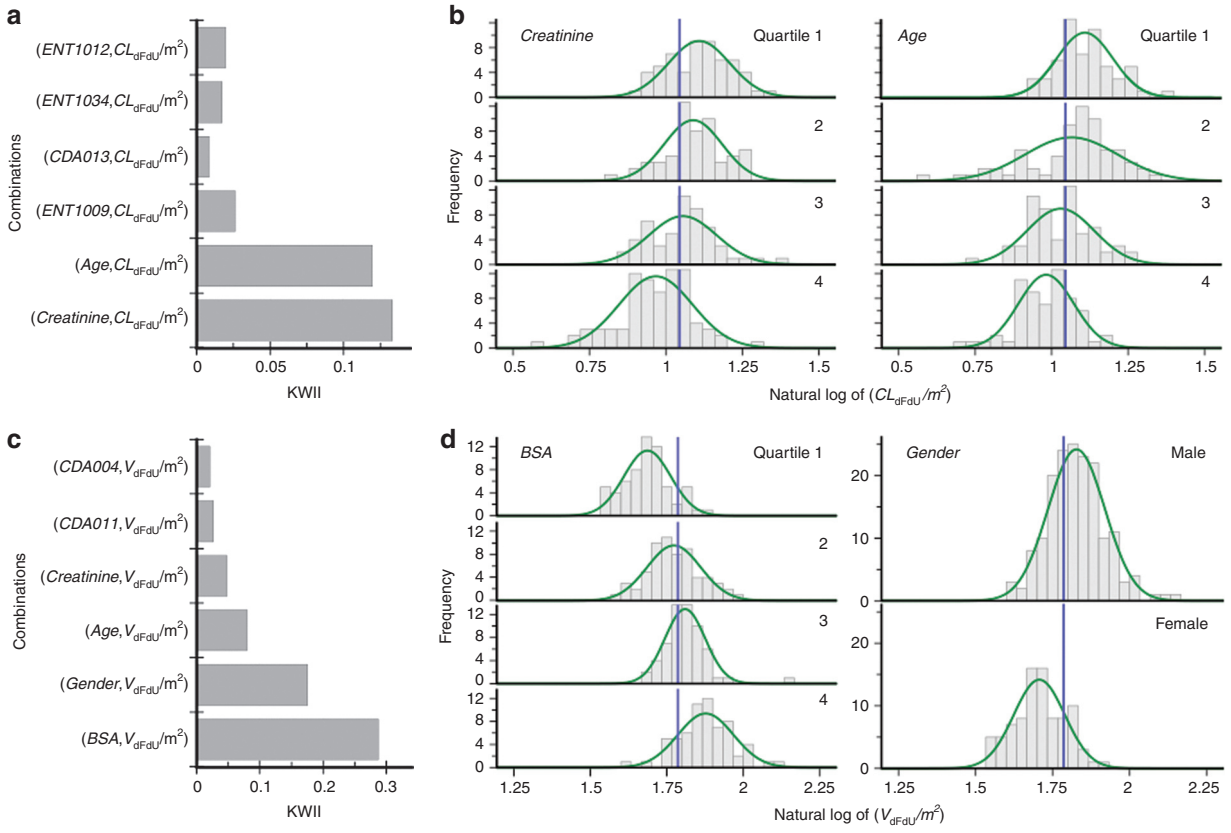

Figure 4 Gemcitabine metabolite PK. (**a**) The KWII values for the top six predictors using the natural logarithm of the clearance of the gemcitabine metabolite dFdU normalized by body-surface area (BSA). The histograms in (**b**) display the distribution of values of the natural logarithm transformed values of dFdU clearance (l/h) across pretreatment creatinine clearance quartiles (left) and age quartiles (right). (**c**) The KWII values for the top six predictors using the natural log transformed values of the volume of distribution for the gemcitabine metabolite dFdU normalized by BSA. The histograms in (**d**) display the distribution of values of the natural logarithm transformed values of the volume of distribution for the gemcitabine metabolite dFdU normalized by BSA across BSA quartiles (left) and gender (right). The vertical reference lines represent the observed overall mean of the phenotype across all values of the predictor. CL_{dFdU}, clearance of the gemcitabine metabolite dFdU; KWII, *k*-way interaction information; PK, pharmacokinetic; V_{dFdU}, metabolite volume of distribution.

plots for clearance of the gemcitabine metabolite dFdU (CL_{dFdU} l/hr/m^2) and its theoretical volume of distribution (V_{dFdU} l/m^2) are shown in **Figure 4a,c**, respectively.

First-order interactions with the cytidine deaminase (*CDA*) *208*G→A (*CDA011*) and *CDA*-116G→A (*CDA004*) SNPs, as well as *Age* (**Figure 3c**, all *P* values ≤0.01) were observed for C_{max}. **Figure 3c** shows the distribution of C_{max} values across age groups and *CDA011* genotypes in the study population and highlights the shifting median values present upon examination. In addition, trends towards association with C_{max} were seen with three of the equilibrative nucleoside transporter (ENT) SNPs – *ENT1 IVS8 + 97T→C* (*ENT1024*, *P* = 0.028), *ENT1 1861C→T* (*ENT1037*, *P* = 0.031), and *ENT1*-7789T→C (*ENT1004*, *P* = 0.041). For CL_{gem}, the *CDA011* and *CDA004* SNPs, along with *Age*, exhibited significant first-order interactions (**Figure 3d**, all *P* values ≤0.01). **Figure 3e** highlights the shifting median values across genotype groups for the *CDA011* and *CDA004* SNPs in the study population. Furthermore, a trend (*P* = 0.012) towards association was found for *ENT1004* (data not shown) with the CL_{gem} phenotype. No significant second-order interactions were found for CL_{gem} or C_{max}.

Figure 4 shows KWII outputs for CL_{dFdU} and V_{dFdu}. For CL_{dFdu}, significant first-order interactions were found for pretreatment levels of creatinine and *Age*. Significant second-order interactions for CL_{dFdU} were found between *ENT1*-3268_-3249del20bp (*ENT1011*) and seven *CDA* SNPs, as well as between *ENT1011* and four deoxycytidine kinase (*DCK*) SNPs.

Significant first-order interactions with V_{dFdU} were seen for body-surface area (BSA), *Gender*, *Age*, pretreatment creatinine level, *CDA011*, and *CDA004*, whereas a trend towards association with V_{dFdU} was observed with *DCK 1736G→A* (*DCK025*). The presence of BSA as a significant first-order predictor even after normalization suggests a disproportional effect from BSA on V_{dFdU}. In addition, for V_{dFdU}, significant second-order interactions involving BSA and two ENT SNPs, *ENT1-5851G→A* (*ENT1005*) and *ENT1 IVS7-121C→T* (*ENT1021*), as well as between BSA and two CDA SNPs, *CDA-182G→A* (CDA003) and *CDA IVS2 + 242A→G* (CDA016), were seen. The significant second-order interactions for V_{dFdU} involve BSA, which is also the first-order interaction with the highest KWII value.

Comparisons to population modeling results. **Table 1** compares the results from the KWII analyses of CL_{gem}, CL_{dFdU}, and V_{dFdU} to the corresponding covariate analyses reported

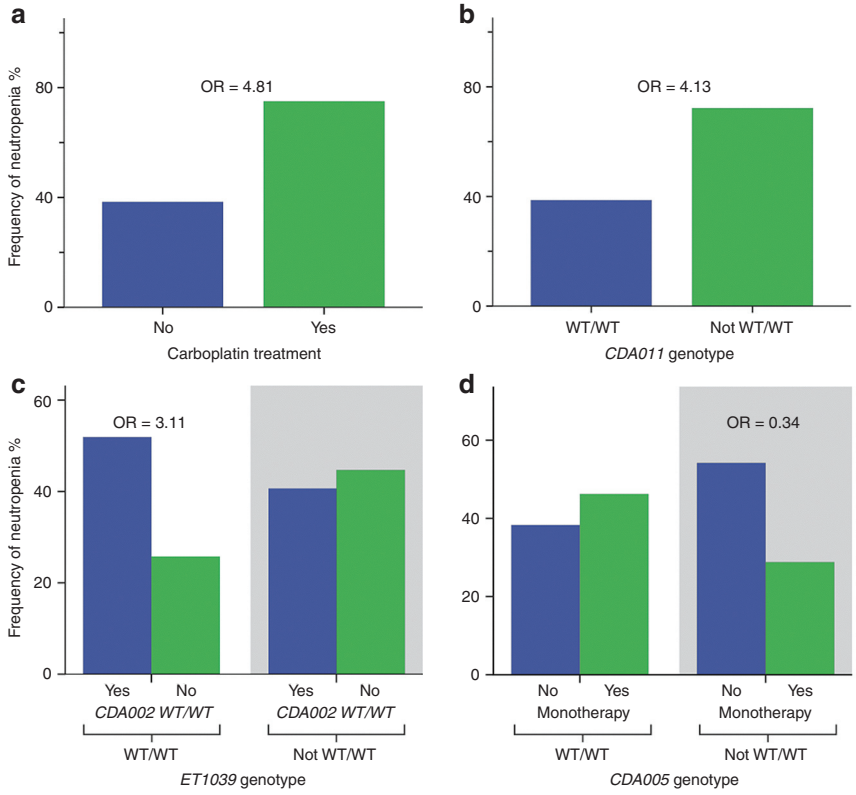

Figure 5 Gemcitabine toxicity. (**a**) The percentage of patients with severe neutropenia (≥grade 3) among cancer patients treated with gemcitabine in the presence (right bar) and absence of carboplatin (left bar). (**b**) The percentage of patients with severe neutropenia for the wild-type homozygous (denoted WT/WT, left bar) and not-WT/WT genotypes of *CDA* 208G→A (*CDA011*). (**c,d**) The percentage of patients with observed neutropenia toxicities for patients with or without mutations at either gene for the (*CDA-205C→G* (*CDA002*), *ENT1* 1984 + 69A→C (*ENT1039*), *Neutropenia*) combination and between those patients with or without *CDA*-92A→G (*CDA005*) mutations across treatment regimens in the {*CDA005*, *Combination therapy*, *Neutropenia*} combination, respectively. ORs are displayed above the relevant variable combinations. CDA, cytidine deaminase; ENT, equilibrative nucleoside transporter; OR, odds ratio.

by Sugiyama *et al.*[7] using NONMEM for the nonlinear mixed effects modeling.

The KWII analyses identified pretreatment creatinine and age to be among the top five predictors for all three PK parameters whereas the NONMEM-based analysis identified creatinine as a covariate only for CL_{dFdU}. Sugiyama *et al.* found the homozygous *CDA**3 (208G→A, A70T (*CDA011*)) and heterozygous *CDA**3 genotypes, as well as the number of *CDA-31delC* deletions to be covariates associated with CL_{gem}, whereas the KWII approach identified *CDA011* and *CDA004*, which are SNPs defining the *CDA*3* haplotype status and are in strong linkage disequilibrium with one another.[8]

In contrast to both the clearance terms in the original analysis, BSA was listed as a covariate for V_{dFdU} without normalization. Normalization of V_{dFdU} by BSA was done for our KWII analysis to better approximate a normal distribution, and BSA was still detected as a significant predictor of V_{dFdU} by the KWII suggesting a nonproportional effect of BSA on V_{dFdU}.

Overall, the comparisons in **Table 1** indicate extensive concordance between the results of study by Sugiyama *et al.* and those from the KWII. For each PK parameter, two or more of the covariates identified via nonlinear mixed effects modeling were among the top five predictors with the highest KWII values.

Gemcitabine-associated toxicities

We also investigated GEI associated with gemcitabine treatment–related neutropenia, a dose-limiting toxicity. Neutropenia was defined as nadir grade of neutrophil counts ≥grade 3.

The significant first-order predictors of serious gemcitabine-associated neutropenia with the highest KWII values were concomitant carboplatin administration (P = 0.01), *CDA011* (P = 0.008), *CDA-205C→G* (*CDA002*, P = 0.018), *CDA004* (P = 0.016), and *ENT-7947G→A* (*ENT1003*, P = 0.04) genotypes. **Figure 5a,b** show the percentage of patients with observed neutropenia toxicities for patients receiving carboplatin (or not), and for patients with mutated (or wild-type) *CDA011* genotypes, respectively. Further examination of the data revealed that the odds ratio for toxicity associated with concomitant carboplatin administration was 4.81 (95% confidence interval = 1.5–15.4). Carrying at least one *CDA011* mutation had an odds ratio of 4.13 (95% confidence interval = 1.5–12.0).

The significant second-order combinations with the highest KWII values were {*CDA*-92A→G (*CDA005*), *Combination therapy*, *Neutropenia*} (P = 0.006), and {*CDA002*, *ENT1* 1984 + 69A→C (*ENT1039*), *Neutropenia*} (P = 0.008). **Figure 5c,d** show the percentage of patients with observed neutropenia toxicities for the {*CDA002*, *ENT1039*, *Neutropenia*} and

Table 1 Comparison of KWII to the population modeling results using NONMEM from Sugiyama *et al.*[7]

Parameter	NONMEM[a]	Top 5 KWII
CL_{gem} (l/h/m^2)	*CDA*3* (*CDA208G→A*) *Homozygous*	*CDA011* (*CDA208G→A*)
	*CDA*3* (*CDA208G→A*) Heterozygous	*CDA004* (*CDA-116G→A*)
	Number CDA-31delC	Age
	S-1 coadministration	Pretreatment creatinine
		CDA007 (CDA-31delC)
CL_{dFdU} (l/h/m^2)	Age	Pretreatment creatinine
	Pretreatment creatinine	Age
		ENT1009 (*ENT1-3548G→C*)
		CDA003 (*CDA 182G→A*)
		ENT1012 (*ENT-1355T→C*)
V_{dFdU} (l)	BSA	BSA
	Age	Age
	Gender	Gender
		Pretreatment creatinine
		CDA011 (*CDA208G→A*)

BSA, body-surface area; CDA, cytidine deaminase; CL_{dFdU}, clearance of the gemcitabine metabolite dFdU; CL_{gem}, gemcitabine clearance; ENT, equilibrative nucleoside transporter; KWII, *k*-way interaction information; V_{dFdU}, volume of distribution.

[a]For CL_{gem} and CL_{dFdU}, the BSA term from the NONMEM model was not included in the table as it is implied necessary for individual prediction from the units listed.

{*CDA005*, *Combination therapy*, *Neutropenia*} combinations, respectively.

Further inspection of the data revealed that the *CDA002* wt/wt genotype group had greater risk of neutropenia than those containing *CDA002* non-wt genotypes in individuals with *ENT1039* wt/wt genotypes (**Figure 5c** left, odds ratio: 3.11, 95% confidence interval: 2.6–3.8). In addition, the data suggest that gemcitabine monotherapy was associated with a reduced risk of neutropenia in the presence of one or more *CDA005* non-wt alleles as compared with combination therapy (**Figure 5d** right, odds ratio: 0.34, 95% confidence interval: 0.32–0.37).

The results indicate that the KWII approach is able to detect genetic and environmental factors associated with drug-related toxicities.

DISCUSSION

The objective of this work was to critically evaluate an innovative information-theoretic approach for GEI analysis capable of integrating genome-wide pharmacogenomic data into the population PK/PD analysis paradigm. We tested the method with several challenging data sets and the results provided novel findings while also demonstrating concordance with published literature.

The inclusion of pharmacogenetic data by Gage *et al.*[9] into algorithms for predicting warfarin therapeutic dose improved the explained variability to ~54% from the previous 17–22% using clinical factors alone. The KWII analysis of INR indicates strong first-order associations with *Race*, *VKORC1* genotype, *CYP2C9* genotype, and use of *Herbals*. The detection of concomitant herbal use as an important predictor of warfarin INR in the IWPC data set is notable. The potential for herbals to interact with warfarin treatment is documented in the literature.[10] Herbals such as St. John's wort, ginseng, coenzyme Q10, danshen, devil's claw, green tea, and papain have shown evidence of interacting with warfarin in *in vitro* studies and the emergence of *Herbals* in our KWII analysis highlights its potential clinical importance. Information on the exact herbals used by subjects was unfortunately not available in the IWPC data set.

We also analyzed the PKs of gemcitabine and its metabolite dFdU. dFdU is eliminated mainly by renal excretion; however, the efflux of dFdU by the ENT transporters from the cell into the circulation, or from blood to the renal tubule, could represent a potential rate-limiting step in its elimination. Although evidence is emerging supporting a role for *ENT1* genotypes in the disposition of gemcitabine,[11] few studies have confirmed these findings. We did not find *ENT1* genotypes among the top combinations for CL_{gem}. However, for V_{dFdU}, significant second-order interactions for two ENT SNPs, *ENT1-5851G→A* (*ENT1005*) and *ENT1 IVS7-121C→T* (*ENT1021*) with BSA were found. In addition, trends towards association with gemcitabine C_{max} were found for three ENT SNPs, *ENT1 IVS8 + 97T→C* (*ENT1024*), *ENT1 1861C→T* (*ENT1037*), and *ENT1-7789T→C* (*ENT1004*). A role for *ENT1039* in combination with *CDA002* was also detected with gemcitabine-associated neutropenia.

From the gemcitabine toxicity data, we detected the *CDA* 208G→A (*CDA011*) and *CDA-116G→A* (*CDA004*) SNPs, defining the *CDA*3* haplotype, which has previously been linked to nucleoside-analog treatment sensitivity and increased rates of severe neutropenia during gemcitabine monotherapy,[12] as well as when gemcitabine is administered in combination with platinum-based chemotherapeutics.[13]

Genetic, environmental, and demographic data are useful for building covariate models in population PK/PD analyses, which seek to identify sources of variability involved in drug disposition and effect. Typically, stepwise selection procedures are used for covariate modeling, which requires iterative refitting, and higher-order interactions are onerous to detect even in small data sets. Previous work from our group suggests that testing saturated parametric interaction models containing as few as 10 predictors may not be feasible on desktop computers.[14] Minimizing bias in covariate effect estimates and model parameters has been previously assessed in population modeling.[15–19] Algorithms for building covariate models have also been proposed.[15,17] For example, Jonsson and Karlsson developed an automated covariate model building strategy within NONMEM[17] that helps to evaluate the effects of adding a covariate on unrelated parameters and tests both linear and nonlinear effects within each run.

Computational issues inherent with genome-scale data have not been examined in population PK/PD covariate modeling. Identifying interactions in large pharmacogenetic data sets presents computational challenges because the number of interactions grows explosively as the number of predictors increases. This combinatorial growth makes it computationally difficult to exhaustively search the full range of genetic and environmental variables for potential interactions associated

with PK/PD and clinical outcomes. Our search algorithms leverage information theory to find the most prominent interactions. Another strength of our approach is that it is nonparametric and does not require a structural interaction model: covariate(s) and covariate combinations (interactions) can be detected with a single unified analysis method. Our approach handles large numbers of predictors, collinearity between predictors, as well as interactions among predictors effectively.

We have shown the potential utility of our information-theoretic analysis method in pharmaceutical applications using point estimates (C_{max}), calculated and individual empirical Bayes estimates (NONMEM), as well as overall exposure (AUC). We have developed a model-building algorithm that selects the parsimonious informative set of predictors.[6] The explicit quantitative relationships between covariates needed for nonlinear mixed effects modeling have to be obtained by visual inspection or other regression methods. This becomes tractable because number of predictors is reduced. Analysis of full longitudinal data is currently a limitation of our method. These entropy expressions require consideration of the correlation structure between time points and are an active area of ongoing research.

It is important to emphasize that our method is not limited to PK/PD data. It can also be used for analysis of non–time course data from early drug development or from the clinical trial and the postmarketing settings. For example, the method could be used to identify drug responder/nonresponder status or for analyzing drug–drug interactions in large databases such as the Food and Drug Administration Adverse Event Reporting System. The information-theoretic approach is versatile for a diverse range of PD response data types as it can handle binary, discrete, rate/count, and continuous variables.[4,5,14] Because the KWII approach does not require model specification, it may allow drug response to be critically analyzed earlier in the development process and could therefore, have a significant impact on the overall modeling strategy for population analysis.

In conclusion, our information-theoretic approach provides novel and effective analytical capabilities for population analyses and for the integration of pharmacogenetic data during the analysis of PK/PD and clinical outcomes. It provides a systematic framework for identifying and incorporating GEI to better assess drug disposition, effect, and outcomes in populations. Critical analysis earlier in the development process could have a significant impact on the overall drug-development paradigm.

METHODS

Definitions, terminology, and representation. Definitions of GEI, gene–gene interactions, and entropy are provided in

$$\begin{aligned}\mathrm{KWII}\,(A,B,C) = &-H(A) - H(B) - H(C) + H(A,B)\\ &+ H(A,C) + H(B,C) - H(A,B,C)\end{aligned} \quad (1)$$

Supplementary Methods online.

KWII. The underlying terminology and representation for the KWII and phenotype-associated information[20] are concisely recapitulated here. For the three-variable case, the KWII is defined in terms of the entropies for the individual variables, $H(A)$, $H(B)$, and $H(C)$ and the joint entropies, $H(A,B)$, $H(A,C)$, $H(B,C)$, and $H(A,B,C)$:

For the K-variable case on the set $v = \{X_1, X_2,\ldots,X_k\}$, the KWII can be written succinctly as an alternating sum over all possible subsets T of v using the difference operator notation of Han:[21]

$$\mathrm{KWII}(v) = -\sum_{T\subseteq v}(-1)^{|v|-|T|}H(T) \quad (2)$$

The number of variables K in a combination is called the order of the combination. The KWII represents the gain or loss of information due to the inclusion of additional variables; it quantifies interactions by representing the information that cannot be obtained without observing all K variables at the same time.[22–25]

The KWII of a given combination is a parsimonious interaction metric; it does not contain contributions arising from the KWII of lower-order combinations (subsets) of these variables. The KWII was employed as the principal measure of GEIs because it is resistant to confounding factors such as linkage disequilibrium and correlations among the variables.

In the bivariate case, the KWII is always nonnegative, but in the multivariate case, the KWII can be positive or negative. We define positive KWII values to indicate interactions (or net synergy) between the variables and negative KWII values to indicate net redundancy between variables. A value of zero indicates the net absence of K-way interactions.

Computational algorithms

CHORUS algorithm. CHORUS is an information-theoretic search algorithm for detecting GEI that is based on the KWII. The algorithm employs the phenotype-associated information to facilitate efficient searching of combinatorial space. The details of CHORUS are described in Chanda *et al*.[5]

Significance testing. For continuous phenotypes, the significance of KWII combinations was assessed using a permutation-derived P value from 10,000 random permutations of the phenotype.

Warfarin PKs. Simulation studies were conducted to assess the capability of our GEI analysis method to identify covariates in a warfarin population PK data analysis. Concentration profiles following a 10 mg oral dose of warfarin were simulated using the structural and variance model parameter values from the warfarin population PK model developed by Hamberg *et al*.[26] The model is shown in **Figure 1a**, with inter-individual variability terms for CL, V_1, and V_2.

Natural-log transformed values of the AUC and individual clearance estimates from NONMEM (CL, in ml/h) were used as phenotypes.

We used a P value ≤0.01 to determine significance. The simulations are described in detail in **Supplementary Methods** online.

Warfarin PDs. The IWPC Warfarin Data Set was obtained from PharmGKB.[27] The INR was used as the phenotype of interest. All patients in the reported IWPC cohort had a target INR of 2–3: the reported INR values in the data set represent

the treatment INR achieved over a period during which the dose of warfarin was stable.[28] The IWPC warfarin data set contains candidate genetic variations known to affect warfarin response including *CYP2C9* and *VKORC1* (*rs7294*); however, as it lacks sufficient size to adequately challenge our information-theoretic method, we created a hybrid data set in which 1,000 SNPs from the GAW15 Problem 1 data set were appended to each subject.

The results from the KWII analysis were compared with the findings from linear regression. Details of the methodology are provided in **Supplementary Methods** online.

Gemcitabine PKs. Genetic information and gemcitabine clinical and PK/PD data were collected from 256 Japanese patients with cancer receiving gemcitabine (dFdC). The data were collected in the National Cancer Center in Japan and the National Institute of Health Sciences in Japan. All methods and protocols for the original studies were approved by the ethics committees of the National Cancer Center and the National Institute of Health Sciences and have previously been described.[7,8,29] Part of the data is available at the Genome Medicine Database of Japan (http://gemdbj.nibio.go.jp).

We performed an information-theoretic analysis using our algorithm, CHORUS,[4] a set of 94 candidate genes (including polymorphisms from the genes for CDA, *DCK*, and *ENT1*), and patient characteristics including age, BSA, pretreatment creatinine levels, and gender. Continuous predictors were discretized into four groups using the observed quartiles as thresholds. Because the candidate gene set contained both existing and novel polymorphisms,[7,8,29,30] we introduced each polymorphism using specific notation and subsequently refer to it using the provided abbreviated notation.

The following calculated PK parameters were individually analyzed as phenotypes of interest: (i) systemic clearance of gemcitabine (CL_{gem}), (ii) the maximum plasma concentration (C_{max}) of gemcitabine, (iii) systemic clearance of the metabolite 2′,2′-difluorodeoxyuridine (dFdU; CL_{dFdU}), and (iv) apparent V_{dFdU}.

All parameters except C_{max} were normalized to BSA and log-transformed to reduce skew in the data and to better approximate a normal distribution.

We used a P value ≤0.01 to determine significance; a trend was defined as a P value ≤0.05.

Gemcitabine toxicities. The data described above for "Gemcitabine PKs" were used. The neutrophil count nadir grade during treatment was used to assess toxicity (evaluated according to the National Cancer Institute Common Toxicity Criteria, version 2). We defined the toxicity phenotype as a binary variable depending on whether or not the nadir grade was 3 or higher. The variable "monotherapy" was added to indicate whether or not gemcitabine was given alone. Complete patient demographics and baseline values are described elsewhere.[7]

Given the limited sample size and higher expected variability inherent in PD data sets, a P value ≤0.05 was used to determine significance.

Acknowledgments. Funding from Pfizer fellowship to J.K. is gratefully acknowledged. Support from the National Multiple Sclerosis Society (RG3743 and a Pediatric MS Center of Excellence Center Grant) and the Department of Defense Multiple Sclerosis Program (MS090122) is gratefully acknowledged. We thank Teruhiko Yoshida for his collaboration in the gemcitabine study. Grants from the National Institute of Biomedical Innovation, Japan, to Dr. Kaniwa and colleagues are gratefully acknowledged.

Author Contributions. J.K. wrote the manuscript. J.K. and M.R. designed the research. J.K., Y.Sato, N.K., Y.Saito, and H.U. performed the research. J.K. and M.R. analyzed the data. P.C. and A.Z. contributed new reagents/analytical tools.

Conflict of interest. The authors declared no conflict of interest.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
Computational constraints severely limit incorporation of gene–gene and gene–environment interactions in drug development. There are currently no effective or efficient methods capable of integrating genomewide-scale pharmacogenetic data into population PK/PD analyses.

WHAT QUESTION DID THIS STUDY ADDRESS?
Can an information-theoretic approach be leveraged to identify gene–gene and gene–environment interactions for drug development and population PK/PD applications?

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE
This work provides novel methodology capable of detecting gene–gene and gene–environment interactions and leveraging large-scale pharmacogenetic/genomic data sets into drug development and population PK/PD analyses.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS
The novel approach allows vertical integration of pharmacogenetic data, gene–gene, and gene–environment interactions during drug development. The ability to leverage large-scale pharmacogenetic data to inform population PK/PD model synthesis could improve predictions of disposition and response to drugs that target complex pathways. This could be translated into more effective treatments and clinical trial designs. Detection of gene–gene and gene–environment interactions from clinical trial data will also enable insight into the underlying mechanistic processes *in vivo*.

1. Lesko, L.J. The critical path of warfarin dosing: finding an optimal dosing strategy using pharmacogenetics. *Clin. Pharmacol. Ther.* **84**, 301–303 (2008).
2. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**, 568–579 (2008).
3. Woodcock, J. & Lesko, L.J. Pharmacogenetics–tailoring treatment for the outliers. *N. Engl. J. Med.* **360**, 811–813 (2009).

4. Chanda, P., Sucheston, L., Liu, S., Zhang, A. & Ramanathan, M. Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits. *BMC Genomics* **10**, 509 (2009).
5. Chanda, P. *et al.* AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* **180**, 1191–1210 (2008).
6. Chanda, P., Zhang, A. & Ramanathan, M. Modeling of environmental and genetic interactions with AMBROSIA, an information-theoretic model synthesis method. *Heredity* (*Edinb*). **107**, 320–327 (2011).
7. Sugiyama, E. *et al.* Population pharmacokinetics of gemcitabine and its metabolite in Japanese cancer patients: impact of genetic polymorphisms. *Clin. Pharmacokinet.* **49**, 549–558 (2010).
8. Sugiyama, E. *et al.* Pharmacokinetics of gemcitabine in Japanese cancer patients: the impact of a cytidine deaminase polymorphism. *J. Clin. Oncol.* **25**, 32–42 (2007).
9. Gage, B.F. *et al.* Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clin. Pharmacol. Ther.* **84**, 326–331 (2008).
10. Heck A.M., DeWitt B.A. & Lukes A.L. Potential interactions between alternative therapies and warfarin. *Am. J. Health Syst. Pharm.* **57**, 1221–1227; quiz 1228–1230 (2000).
11. Gusella, M. *et al.* Equilibrative nucleoside transporter 1 genotype, cytidine deaminase activity and age predict gemcitabine plasma clearance in patients with solid tumours. *Br. J. Clin. Pharmacol.* **71**, 437–444 (2011).
12. Ueno, H. *et al.* Homozygous CDA*3 is a major cause of life-threatening toxicities in gemcitabine-treated Japanese cancer patients. *Br. J. Cancer* **100**, 870–873 (2009).
13. Yonemori, K. *et al.* Severe drug toxicity associated with a single-nucleotide polymorphism of the cytidine deaminase gene in a Japanese cancer patient treated with gemcitabine plus cisplatin. *Clin. Cancer Res.* **11**, 2620–2624 (2005).
14. Knights, J. & Ramanathan, M. An information theory analysis of gene-environmental interactions in count/rate data. *Hum. Hered.* **73**, 123–138 (2012).
15. Ribbing, J., Nyberg, J., Caster, O. & Jonsson, E.N. The lasso–a novel method for predictive covariate model building in nonlinear mixed effects models. *J. Pharmacokinet. Pharmacodyn.* **34**, 485–517 (2007).
16. Wählby, U., Jonsson, E.N. & Karlsson, M.O. Comparison of stepwise covariate model building strategies in population pharmacokinetic-pharmacodynamic analysis. *AAPS PharmSci* **4**, E27 (2002).
17. Jonsson, E.N. & Karlsson, M.O. Automated covariate model building within NONMEM. *Pharm. Res.* **15**, 1463–1468 (1998).
18. Joerger, M. Covariate pharmacokinetic model building in oncology and its potential clinical relevance. *AAPS J.* **14**, 119–132 (2012).
19. Ribbing, J. & Jonsson, E.N. Power, selection bias and predictive performance of the Population Pharmacokinetic Covariate Model. *J. Pharmacokinet. Pharmacodyn.* **31**, 109–134 (2004).
20. Chanda, P. *et al.* Information-theoretic metrics for visualizing gene-environment interactions. *Am. J. Hum. Genet.* **81**, 939–963 (2007).
21. Han T.S. Multiple mutual informations and multiple interactions in frequency data. *Inform Contr* **46**, 26–45 (1980).
22. Jakulin A. Machine Learning Based on Attribute Interactions. Thesis, Univ Ljubljana (2005).
23. Jakulin A. & Bratko I. Testing the significance of attribute interactions. In (eds. Greiner, R. D. & Schuurmans, D.). Proceedings of the Twenty-first International Conference on Machine Learning (ICML-2004); 2004, Banff, Canada, 2004:409–416.
24. McGill W.J. Multivariate information transmission. *Psychometrika* **19**, 97–116 (1954).
25. Fano R.M. Transmission of Information: A Statistical Theory of Communications. (MIT Press, Cambridge, MA, 1961).
26. Hamberg, A.K. *et al.* A PK-PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. *Clin. Pharmacol. Ther.* **81**, 529–538 (2007).
27. PharmGKB. IWPC pharmacogenetic dosing algorithm. <http://www.pharmgkb.org/do/serve?objCls=Submission&objId=PS208895> (2009).
28. Klein T.E. *et al.* Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**, 753–764 (2009).
29. Kim, S.R. *et al.* Thirty novel genetic variations in the SLC29A1 gene encoding human equilibrative nucleoside transporter 1 (hENT1). *Drug Metab. Pharmacokinet.* **21**, 248–256 (2006).
30. Kim, S.R. *et al.* Twenty novel genetic variations and haplotype structures of the DCK gene encoding human deoxycytidine kinase (dCK). *Drug Metab. Pharmacokinet.* **23**, 379–384 (2008).

Supplementary Information accompanies this paper on the *Pharmacometrics & Systems Pharmacology* website (http://www.nature.com/psp)