An Information Theory Analysis of Gene-Environmental Interactions in Count/Rate Data

Jonathan Knights    Murali Ramanathan
Department of Pharmaceutical Sciences, State University of New York, Buffalo, N.Y., USA

Key Words
Information theory · Count data · Gene-gene interactions · Gene-environment interactions · K-way interaction information

Abstract
Objective: To develop and critically evaluate an information theory method for identifying gene-gene and gene-environment interactions in count and rate data. Methods: The entropy-based metric k-way interaction information (KWII) was critically assessed for utility in detecting interactions with count data and over-dispersed count data in three simulation studies of increasing complexity and in datasets from animal models of depression and colitis. The results were compared to Poisson regression. The power and effect size dependence of the KWII for detecting interactions was also assessed. Results: The KWII was capable of effectively identifying the genetic and environmental predictors and their interactions in all three simulated datasets. The results indicate that the KWII approach may produce more parsimonious results than regression. In a rat model of depression, we successfully identified a prominent gender effect as well as other published associations. Analysis of severity scores from an animal model of colitis identified markers from chromosome 3, as well as unique first- and second-order associations for the individual sections of the colon and cecum. Conclusions: The results demonstrate the utility and versatility of our entropy-based method for gene-environment interaction analysis of count and rate data with Poisson and over-dispersed distributions.

Introduction
Many common diseases, ranging from cardiovascular, neurological, and autoimmune diseases to cancer, have complex genetic etiologies involving interactions among multiple genetic variations and environmental factors [1–3]. Rheumatoid arthritis is a prototypical example of a disease where there is an unmet need for a better understanding of the role of gene-gene (GGI) and gene-environment interactions (GEI) to delineate the mechanism of the disease, for both the response to available treatments and the prediction of disease severity and progression [4].

Count data are a distinctive type of data that result when a disease phenotype, end point, or outcome can be described using rates or frequencies. Examples of count
outcomes are urinary incontinence measures, relapse rates in multiple sclerosis, seizure counts in epilepsy, and swollen joint counts in rheumatoid arthritis [5]. They are commonly encountered and clinically relevant across many diseases.

Count data are often modeled using regression methodologies based on the Poisson distribution [6–8]. Regression-based methods require model specification, which can be onerous for large-scale genome-wide data sets. Furthermore, the performance of saturated regression models is adversely affected by co-linearity among predictor variables caused by linkage disequilibrium and correlation models is adversely affected by co-linearity among predictor variables caused by linkage disequilibrium and correlations among environmental variables.

There is a critical need for algorithms capable of detecting interactions between variables in very large data sets such as those from genome-wide association studies and from large-scale next-generation sequencing studies. In previous work, we have demonstrated the utility of AMBIENCE [9] and CHORUS [10], our information theory, entropy-based algorithms for GGI and GEI analysis. AMBIENCE was developed for binary or categorical discrete phenotypes, whereas CHORUS was developed for normally distributed phenotypes. The purpose of this paper is to extend and critically evaluate an information theory-based method for GGI and GEI analysis to count data.

### Materials and Methods

The information theory metrics underlying our methods have been previously described [9, 10], and we recapitulate these succinctly.

**Entropy**

The entropy $H(X)$ of a discrete random variable $X$ is obtained from its probability mass function $p(x)$ using the Shannon entropy formula:

$$H(X) = -\sum_x p(x) \log p(x)$$  (1)

**K-Way Interaction Information**

For the three-variable case involving two genetic or environmental variables denoted by $A$ and $B$, and the Poisson-distributed phenotype denoted by $Y$, the $k$-way interaction information ($\text{KWII}$) is defined in terms of entropies of the individual variables $H(A)$, $H(B)$, and $H(Y)$, and the joint entropies $H(AB)$, $H(AY)$, $H(BY)$, and $H(ABY)$:

$$\text{KWII}(A;B;Y) = -H(ABY) + H(AB) + H(AY) + H(BY) - H(A) - H(B) - H(Y)$$  (2)

For the $K$-variable case on the set $\nu = \{X_1, X_2, \ldots, X_K, Y\}$, the KWII is an alternating sum over all possible subsets $T$ of $\nu$. Using difference operator notation [11]:

$$\text{KWII}(\nu) = -\sum_{T \subset \nu} (-1)^{|\nu| - |T|} H(T)$$  (3)

where $|\nu|$ and $|T|$ represent the number of variables in the set and subset, respectively, and $H(T)$ represents the Shannon entropy of the subset $T$. The number of genetic and environmental variables $K$ (not including the phenotype) in a combination is called the order of the interaction. The KWII quantifies interactions by representing the information that cannot be obtained without observing all $K$ variables and the phenotype $Y$ at the same time [12–15]. The KWII of a given combination of variables is a parsimonious interaction metric in that it does not contain contributions arising from the KWII of other lower-order combinations (i.e. the subsets of the $K$-way variable combination).

**Interaction Definition**

We operationally define that ‘for each variable combination containing the phenotype, a positive KWII value indicates the presence of an interaction, negative values of KWII indicate the presence of redundancy, and a KWII value of zero denotes the absence of $K$-way interactions’.

**Extension to Poisson Count Data**

The expression for entropy and KWII can be computed for Poisson-distributed count data. The probability density function $P(Y = k)$ of a Poisson-distributed random variable $Y$ with the rate parameter $\lambda$ is given by:

$$P(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$  (4)

The Shannon entropy $H(Y)$ of a Poisson-distributed variable $Y$ with the rate parameter $\lambda$ is defined as:

$$H(Y) = -\sum_k p(k) \log p(k)$$  (5)

Combining equations 4 and 5 gives the expression:

$$H(Y) = -\sum_k \frac{e^{-\lambda} \lambda^k}{k!} \ln \frac{e^{-\lambda} \lambda^k}{k!}$$  (6)

$$= -\sum_k \frac{e^{-\lambda} \lambda^k}{k!} (\ln \lambda - \lambda - \ln k! - \ln k!)$$

From equation 6, we easily expand to:

$$H(Y) = \lambda e^{-\lambda} \sum_k \frac{\lambda^k}{k!} - \lambda (\ln \lambda) e^{-\lambda} \sum_k \frac{\lambda^{k-1}}{(k-1)!} \frac{\lambda^k}{k!}$$

$$+ e^{-\lambda} \frac{\lambda^k}{k!} (\ln k!)$$  (7)

Using the identity

$$e^\lambda = \sum_{k=0}^\infty \frac{\lambda^k}{k!}$$

for the summations in the first two terms, we obtain our final expression for the entropy of a Poisson random variable:

$$H(Y) = f(\lambda) = \lambda (1 - \ln \lambda) + e^{-\lambda} \sum_{k=0}^\infty \frac{\lambda^k (\ln k!)}{k!}$$  (8)

For KWII calculations, we require $H(X,Y)$, the entropy of the joint distribution of the phenotype $Y$ and discrete variable(s) $X$, Kn2012:73:123–138

Knights/Ramanathan
representing, for example, genetic variants or environment variables of interest or their combinations. It can be shown that:

\[ H(X, Y) = H(X) + \sum_x p(X = x) H(Y | X = x) \]  

(9)

The \( H(X) \) term contains only discrete variables and, as a result, this entropy can be computed using the Shannon entropy formula highlighted in equation 5.

We assume that the phenotype \( Y \), given \( X = x \), is Poisson distributed with the rate parameter \( \lambda_x \). The maximum likelihood estimator for \( \lambda_x \) is the mean value of the phenotype \( Y \), given \( X = x \). Expanding, simplifying, and substituting equation 9, we obtain:

\[ H(X, Y) = H(X) + \sum_x p(X = x) f(\lambda_x) \]  

(10)

These equations for the joint entropy, the entropies of discrete variable combinations and the entropy of the phenotype, enable computation of the KWII.

**Extension to Zero-Inflated Poisson Count Data**

We also investigated the KWII-based method for zero-inflated Poisson (ZIP) data to assess its potential usefulness for over-dispersed count distributions. The ZIP density function \( p(k) \) is defined by:

\[ p(k = 0) = (1 - \psi) + \psi e^{-\lambda} \quad \text{for } k = 0 \]  

(11)

\[ p(k > 0) = \psi e^{-\lambda} \lambda^k / k! \quad \text{for } k = 1, 2, 3 \ldots \]  

(12)

The ZIP distribution is a mixture distribution consisting of two components: a Poisson density with rate parameter \( \lambda \) plus an extra 'zero-generating' density \( 1 - \psi \) at \( k = 0 \) [7, 16]. The parameters of the ZIP distribution were calculated using the moment estimators:

\[ \hat{\lambda}_p = \frac{s^2}{\bar{X}} - 1 + \bar{X} \]  

(13)

\[ \hat{\psi}_p = \frac{\bar{X}}{\lambda_p} \]  

(14)

where \( \bar{X} \) is the observed sample mean and \( s^2 \) is the unbiased estimate of the population variance from the sample:

\[ s^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \bar{X})^2 \]  

(15)

The entropy \( H(Y) \) of the ZIP distribution was calculated using the Shannon entropy formula and the probability density \( p(k) \) of the ZIP distribution:

\[ H(Y) = f(\hat{\lambda}_p, \hat{\psi}_p) \]

\[ = p(k = 0) \ln p(k = 0) + \sum_{k=1}^{\infty} p(k) \ln p(k) \]  

(16)

The \( f(\hat{\lambda}_p, \hat{\psi}_p) \) denotes that the entropy \( H(Y) \) for a ZIP-distributed random variable \( Y \) is a function of the \( \hat{\lambda}_p \) and \( \hat{\psi}_p \) parameters. For entropy calculations, the individual genotype and genotype combination substrata were treated as ZIP distributed if their \( \hat{\psi}_p \) satisfied the criteria \( 0 < \hat{\psi}_p < 1 \), and as Poisson distributed otherwise.

---

**Simulations**

Simulations were conducted using three published GEI models from our group [9, 10] that incorporate increasingly complex interaction patterns among variables. All single nucleotide polymorphisms (SNPs) were simulated with a minor allele frequency (MAF) of 0.5 under the assumption of Hardy-Weinberg equilibrium.

The count data were described by Poisson distributions with the rate parameter \( \lambda_0 \) in healthy individuals and the rate parameter \( \lambda_D \) in diseased individuals. We define effect size as the difference between the two rates, normalized to the rate in healthy individuals:

\[ \frac{\lambda_D - \lambda_0}{\lambda_0} \]

For simplicity, we hold \( \lambda_0 = 1 \) and vary effect size by changing \( \lambda_D \). Simulations were conducted for disease rate parameter \( \lambda_D \) values of 1.25, 1.5, 2, and 3. For each value of the disease rate parameter \( \lambda_D \), 100 sample populations, each containing 5,000 individuals, were independently simulated for Models 1, 2, and 3.

Permutation-based KWII p-values were obtained by randomly permuting the phenotype values; 10,000 independent permutations of the phenotype were conducted for each combination of interest. A type 1 error level of 0.01 was used for significance. We present the number of tests for all one- and two-way combinations for each dataset, as well as the top unadjusted p-values and corresponding KWII values for one- and two-way combinations. Given that there are many multiple testing methods, e.g. the SDMinP method of Obreiter et al. [17] or the Benjamini-Hochberg method [18], we felt adjustments would make it more difficult for the readers to make their own assessments. We therefore opted to present the rank unadjusted values for scrutiny. The effect of sample size on the power of the KWII was also explored.

All simulations were carried out with user-written Java code implemented in the NetBeans 7.0 M1 Integrated Development Environment. All work including simulations, KWII computations, and permutations was done on a MacBook Pro with a 1.83 GHz Intel Core Duo processor running Mac OS X version 10.6.8. The code is provided at http://pharmsci.buffalo.edu/computational_software/murali_1/download/.

**Poisson Regression**

Poisson regression is an established statistical approach for count data analysis. We therefore compared our KWII-based permutation p-values to those obtained with Poisson regression analyses. In addition, we conducted power studies to compare the performance of the two methods. Poisson regressions were conducted using PASW Statistics 18.0 (IBM Inc.). Poisson regression analyses for Models 1 and 2 employed a saturated interaction model. For Model 3, a modified approach was necessary because saturated fourth-order model runs with more than eight predictors did not complete. Saturated regression models were run for orders 1–3, and the top eight predictors associated with combinations having p-values ≤0.05 from those orders (i.e. E1, E2, E3, E4, S1, S2, S5, and S6) were used to obtain a saturated model for all four-way combinations of this subset. The final comparative regression model for Model 3 combined all combinations from each order with p ≤ 0.05. A type 1 error level of ≤0.01 was used to assess significance.
Model Simulations

Figures 1a, 2a and 3a summarize the GEIs simulated in Models 1–3, respectively.

Model 1
This model contains a two-SNP interaction that is influenced by a single (three-state) environmental variable E. The proportions of the population with low (L), moderate (M), or high (H) exposure levels of the environmental variable E were 0.25, 0.5, and 0.25, respectively. Five biallelic SNPs (S1–S5) with a MAF of 0.5 were simulated. SNPs S3–S5 were simulated to have no influence on the phenotype or other SNPs. Only the genotypes of S1 and S2 interact with each other or with the states of E to determine the phenotype Y.

Model 2
The allele frequencies, environmental factor exposure levels, and proportions of Model 2 were identical to those employed in Model 1. However, this model (fig. 2) incorporates a more complex interaction pattern between the environmental variable E and the two SNPs S1 and S2.

Model 3
Model 3 incorporates two binary (E1, E3) and two ternary (E2, E4) environmental variables with six biallelic SNPs (with a MAF of 0.5) for a total of ten predictor variables. A binary (E1) and ternary (E2) environmental variable interact with two SNPs, S1 and S2, to determine the phenotype. The proportions of the population with low, moderate, or high exposure levels of the environmental variables E2 and E4 were 0.25, 0.5, and 0.25, respectively. The proportions of the population with low and high exposure levels to E1 and E3 were both 0.5. All other variables were simulated to have no influence on the phenotype.

Power Calculations
The null distribution of the KWII was obtained by setting both λ0 and λD equal to unity, i.e. an effect size of zero. KWII calculations from 1,000 independent simulations, each containing 5,000 individuals, were done to obtain the null distribution of the KWII. The 95th percentile KWII value of the null dataset was determined for each order to obtain the threshold for power assessment. The power of the KWII-based method to detect a particular combination of interest was calculated as the percentage of the original 100 simulated datasets where the KWII for that combination exceeded the 95th percentile threshold of the null distribution for that order. For Poisson regression, power was calculated as the percentage of the 100 simulated datasets in which the particular combination of interest was detected as significant at p ≤ 0.05. The effect of sample size on the power of the KWII was also investigated for sample sizes of 1,000 and 2,000 using the same approach.

Analysis of Real Datasets
Defensive Burying Dataset
The shock-probe defensive burying test (DB test) is a rodent experimental model for behavioral responses to stress. The DB test is a 15-min test wherein a probe that delivers electric shock is introduced in the rodent’s cage, following which the animal’s behavior is videotaped and scored. The animals usually approach the prod (and may be shocked) a number of times before shoveling bedding material, present in the cage, at the probe as a defensive mechanism.

Animals with active coping tend to have fewer approaches to (and shocks from) the probe (along with other characteristics) than those animals with passive coping [19, 20]. Rodent coping strategies are believed to be composed of genetic, environmental, and learned components [21]. The DB test has been validated as capable of distinguishing different coping strategies [21, 22]. This dataset was obtained from http://churchill.jax.org/index.shtml.

The number of shocks was used as the phenotype of interest. For predictor variables, we included gender, strain of paternal grandmother, and all genetic markers. This dataset was filtered as follows: genotypes with frequency <0.05 were marked as missing, while genetic markers with >20% missing data and animals missing values for >10% of the predictors were eliminated. The final dataset contained 89 predictors and 350 animals.

Colitis Data Set
A dataset from a mouse model of colon inflammation, or colitis, was also obtained from http://churchill.jax.org/index.shtml. This mouse model is intended to mimic chronic inflammatory bowel disease (IBD), which includes Crohn’s disease and colitis in humans. IBD in humans is a disorder with immunologic, environmental, and genetic components [23–25]. Targeted mutations in mouse genes have been shown to produce colitis and serve as a tool for understanding human IBD [24]. The total cecum score and the scores of the proximal, medial, and distal sections of the colon were analyzed individually as phenotypes of interest.

Gender, fecal state (labeled ‘Diarrhea’ and representing soft, normal, or liquid stool), and all genetic markers were chosen as potential predictors. The data were filtered in the same fashion as the defensive burying dataset. The final dataset contained 47 predictors and 396 animals. A type I error level of ≤ 0.01 was used to assess significance.

Results

Results for Simulated Data
Three simulated case studies with increasingly complex GEI were conducted to obtain an initial assessment of the KWII-based method for Poisson-distributed phenotypes.

Model 1
The KWII values for one-SNP- and two-SNP-containing combinations for an effect size of 0.5 are shown in figure 1b. The most prominent peaks were seen for the {E, Y}, {S1, Y}, and {S2, Y} combinations, which contain one predictor. The {E, S1, Y}, {E, S2, Y}, and {S1, S2, Y} combinations had the highest KWII values among the two-predictor-containing combinations. Notably, the environmental and genetic predictors simulated to have interactions in the model had prominent KWII peaks, whereas combinations containing SNPs S3–S5, which...
were not associated with the phenotype, had small peaks in comparison. There were 42 total one-, two-, and three-way tests for Model 1. Table 1 compares Poisson regression p-values with KWII p-values for combinations detected by either method for Model 1.

The effect size dependence of KWII for the first-order combination \( \{S1, Y\} \) and the second-order combination \( \{E, S1, Y\} \) is shown in Figure 1c. The KWII values increase monotonically with increasing effect size, with the most rapid increases occurring at effect sizes of 1.0 or less. Figure 1d.
The power of the KWII to identify the {S1, Y} and {E, S1, Y} combinations as a function of effect size. The power of the {E, Y} combination was 98% for an effect size of 0.25. The power for the {E, S1, Y} combination at an effect size of 0.25 was approximately 35% and increased to 86% at an effect size of 0.5. The KWII and power for the other combinations associated with the phenotype in the model were qualitatively similar (data not shown).

Table 2 shows the effect of sample size on the power of the KWII for the Model 1 combinations. At an effect size of 0.5 and a sample size of 1,000, the KWII has excellent power for detecting all the first-order combinations for Model 1. At a sample size of 1,000, however, detection of second-order interactions at a power around 50% requires an effect size of around 1.

At an effect size of 0.5 and a sample size of 5,000, the power of the KWII to detect all combinations in table 2 was 61%. At an effect size of 1 and a sample size of 5,000, the power to detect all combinations was 99%.

At a sample size of 5,000, the average KWII computation time for orders 0–3 for Model 1 were 0.074, 1.6, 6.3, and 15 s, respectively, where order 0 represents the computation of the Shannon entropy for a single variable. At the n = 1,000 sample size level, the average computation times decreased approximately proportionally to the sample size, giving average computation times of 0.02, 0.35, 1.3, and 3 s for orders 0–3, respectively. Permutation testing times for a sample size of 5,000 and 10,000 permutations of the phenotype for each first-, second-, and third-order interaction were approximately 6, 11, and 22 min, respectively.

Model 2

Figure 2b plots the KWII values for different one- and two-predictor-containing combinations for an effect size of 0.5. The most prominent peak for Model 2 was for the {E, S1, S2, Y} combination. The KWII values of the lower-order {E, Y}, {E, S1, Y}, and {E, S2, Y} combinations were smaller than that of {E, S1, S2, Y}. The comparative absence of the {S1, S2, Y} peak suggests that the interaction of the SNPs S1 and S2 with E is critical for explaining the phenotype. This is not surprising as the phenotype rate in this model, given S1 and S2, is simulated to be different depending on the exposure to E. There were 42 total one-, two-, and three-way tests for Model 2. Table 1 compares Poisson regression p-values with KWII p-values for combinations detected by either method for Model 2.

Power analysis (fig. 2d) shows that the KWII had a power of almost 100% at an effect size of 0.5 to detect the most prominent first-order {E, Y} combination and third-order {E, S1, S2, Y} combination in this model. The second-order peak {E, S2, Y} had a power of 73% at an effect size of 0.5, which increased to nearly 100% at an effect size of 1. Table 2 shows the effect of sample size on the power of the KWII for Model 2 combinations. At an effect size of 0.5 and a sample size of 1,000, the KWII displays adequate power for detection of the relevant first-order combination {E, Y}. The KWII detects the full third-order combination with nearly 100% power at all sample sizes.

At an effect size of 0.5 and a sample size of 5,000, the power of the KWII to detect all Model 2 combinations in table 2 was 68%. The power to detect all combinations for an increased effect size of 1.0 was 100%.

At a sample size of 5,000, the average KWII computation time for orders 0–3 for Model 2 were 0.067, 1.6, 6.3, and 15 s, respectively. At the n = 1,000 sample size level, the average computation times again decreased approximately proportionally to the sample size, giving average computation times of 0.02, 0.35, 1.3, and 3 s for orders

Table 1. Comparison of the p-values of the KWII to those of Poisson regression

<table>
<thead>
<tr>
<th>Combination</th>
<th>KWII, p-value</th>
<th>Poisson, p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E, Y}</td>
<td>&lt;0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>{S1, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{S2, Y}</td>
<td>&lt;0.0001</td>
<td>0.16</td>
</tr>
<tr>
<td>{E, S1, Y}</td>
<td>0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>{E, S2, Y}</td>
<td>&lt;0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>{S1, S2, Y}</td>
<td>&lt;0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{E, S4, Y}</td>
<td>0.014</td>
<td>0.005</td>
</tr>
<tr>
<td>{E, S1, S2, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E2, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{S1, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{S2, Y}</td>
<td>&lt;0.0001</td>
<td>0.002</td>
</tr>
<tr>
<td>{E1, E2, Y}</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{S1, S2, Y}</td>
<td>0.0007</td>
<td>0.17</td>
</tr>
<tr>
<td>{E1, E2, S1, Y}</td>
<td>0.009</td>
<td>N/A</td>
</tr>
<tr>
<td>{E2, S1, S2, Y}</td>
<td>0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>{E2, E3, S1, Y}</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td>{E3, S2, S4, Y}</td>
<td>N/A</td>
<td>0.007</td>
</tr>
<tr>
<td>{E3, S1, S2, S3, Y}</td>
<td>N/A</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Only combinations significant at p ≤ 0.01 for either method are shown. N/A indicated if combination was not present in the final Poisson regression model or if the combination was not among the top ten combinations with the highest KWII values.
Permutation testing times for a sample size of 5,000 and 10,000 permutations of the phenotype for each first-, second-, and third-order interaction were approximately 6, 11, and 22 min, respectively.

Model 3

Figure 3b shows the KWII spectrum for the top six combinations of orders 1–4 at an effect size of 0.5. Peaks corresponding to combinations {E1, Y}, {E2, Y}, {S1, Y}, and {S2, Y}, which represent the first-order effects of all variables associated with the phenotype, were present in the spectrum. Second-order peaks were observed for {E1, E2, Y}, {S1, S2, Y}, {E2, S1, Y}, and {E2, S2, Y}. We attribute the absence of the {E1, S1, P} and {E1, S2, Y} peaks to ‘shielding’, i.e. the {E1, Y} interaction and interactions of the S1 and S2 variables with E2 sufficiently explain the effects on the phenotype. Peaks corresponding to the higher-order combinations {E2, S1, S2, Y} and {E1, E2,
S1, S2, Y} were also observed. There were 375 total one-, two-, three-, and four-way tests for Model 3. Table 1 compares Poisson regression p-values with KWII p-values for combinations detected by either method for Model 3.

The power (fig. 3d) was generally correlated with the values of the KWII. At an effect size of 0.5, the power for the first-order combination \{E, Y\} was 75%, whereas the power for the second-order combination \{E1, E2, Y\} was 78%, correlating with its higher KWII value. Further, the \{E2, S1, S2, Y\} combination, which had a greater KWII value than \{E1, E2, S1, S2, Y\} (fig. 3b), had correspondingly higher power (96%) at an effect size of 0.5. Similarly, the power of the \{E1, E2, S1, S2, Y\} combination increases from 11% for an effect size of 0.5 to 77% at an effect size of 1. Similarly, the power of the \{E1, E2, S1, S2, Y\} combination increases from 11% for an effect size of 0.5 to 77% at an effect size of 1. Table 2 shows the effect of sample size on the power of the KWII for Model 3 combinations. SNPs S1 and S2 are detected at 74% power at an effect size of 0.5 and a sample size of 1,000. Adequate power of higher-order interactions is seen with an effect size of 1.0 at all sample sizes tested.

At an effect size of 0.5 and a sample size of 5,000, the power of the KWII to detect all combinations in table 2 for Model 3 was 6%. At a sample size of 5,000, the power to detect all combinations at an increased effect size of 1.0 was 77%.

At a sample size of 5,000, the average KWII computation times for orders 0–3 for Model 3 were 0.10, 4.9, 12.6, and 32 s, respectively. At the n = 1,000 sample size level, the average computation times decreased approximately proportionally to the sample size, giving average computation times of 0.031, 0.98, 2.6, and 7.1 s for orders 0–3, respectively. Permutation testing times for a sample size of 5,000 and 10,000 permutations of the phenotype for each first-, second-, third-, and fourth-order interaction were approximately 6, 11, 22, and 41 min, respectively.

Comparisons of KWII and Poisson Regression

The results from the comparisons of the KWII-based GEI analyses (and permutation testing) and those from Poisson regression analyses for case studies 1–3 with a sample size of 5,000 and an effect size of 0.5 are shown in table 1.

Model 1

There was excellent concordance between the KWII and Poisson regression for Model 1. There was qualitative agreement between the KWII and Poisson regression p-values for combinations detected by either method for Model 3.

Table 2. Dependence of power of the KWII on sample sizes of 1,000, 2,000, and 5,000 for effect sizes of 0.5 and 1

<table>
<thead>
<tr>
<th>Combination</th>
<th>Effect size = 0.5</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 1,000</td>
<td>n = 2,000</td>
<td>n = 5,000</td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E, Y}</td>
<td>88</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>{S1, Y}</td>
<td>85</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>{S2, Y}</td>
<td>86</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>{E, S1, Y}</td>
<td>27</td>
<td>46</td>
<td>86</td>
</tr>
<tr>
<td>{E, S2, Y}</td>
<td>23</td>
<td>44</td>
<td>83</td>
</tr>
<tr>
<td>{S1, S2, Y}</td>
<td>22</td>
<td>47</td>
<td>87</td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E, Y}</td>
<td>55</td>
<td>89</td>
<td>100</td>
</tr>
<tr>
<td>{E, S1, Y}</td>
<td>10</td>
<td>27</td>
<td>68</td>
</tr>
<tr>
<td>{E, S1, S2, Y}</td>
<td>96</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Model 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{E2, Y}</td>
<td>17</td>
<td>45</td>
<td>75</td>
</tr>
<tr>
<td>{S1, Y}</td>
<td>74</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>{S2, Y}</td>
<td>74</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>{E1, E2, Y}</td>
<td>11</td>
<td>40</td>
<td>78</td>
</tr>
<tr>
<td>{S1, S2, Y}</td>
<td>18</td>
<td>30</td>
<td>71</td>
</tr>
<tr>
<td>{E1, E2, S1, Y}</td>
<td>0</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>{E2, S1, S2, Y}</td>
<td>39</td>
<td>75</td>
<td>96</td>
</tr>
<tr>
<td>{E1, E2, S1, S2, Y}</td>
<td>1</td>
<td>2</td>
<td>28</td>
</tr>
</tbody>
</table>
agreement between permutation-based p-values of the KWII and the corresponding regression-derived p-values from Poisson regression. Both methods detected the $\{E, Y\}$, $\{S1, Y\}$, $\{E, S1, Y\}$, $\{E, S2, Y\}$, and $\{S1, S2, Y\}$ combinations as significant.

The KWII identified the $\{S2, Y\}$ as significant but Poisson regression did not. However, a trend ($p = 0.045$) was found for the $\{E, S1, S2, Y\}$ combination in Poisson regression that was not present in the KWII analysis (data not shown). This suggests that the KWII approach identifies
The KWII and Poisson regression methods both correctly identified the \( E, Y \) and \( E, S_1, S_2, Y \) combinations. The spurious \( E, S_4, Y \) combination was significant (\( p = 0.005 \)) in Poisson regression. The spurious \( \{ E, S_4, S_1, Y \} \) combination exhibited trends in both methods (\( p = 0.029 \) for the KWII and \( p = 0.017 \) for Poisson regression; data not shown).

Table 3 shows the power analysis for the two methods. For Model 2, both methods detect the environmental variable \( E \) as the only informative first-order interaction, with the KWII showing a distinct advantage at an effect size of 0.25. For the second-order, the KWII detects interactions between \( E \) and both informative SNPs \( S_1 \) and \( S_2 \) (only the \( \{ E, S_1, Y \} \) case is shown in table 2). Poisson regression does not detect any second-order interaction with \( E \), reporting only the \( \{ S_1, S_2, Y \} \) interaction with a power of 34% at an effect size of 0.25 (data not shown). Both methods detect the full third-order interactions at 100% for all the effect sizes.

Model 3

Both methods detected the \( \{ E_2, Y \}, \{ S_1, Y \}, \{ S_2, Y \}, \{ E_1, E_2, Y \}, \{ S_1, S_2, Y \}, \{ E_1, E_2, S_1, Y \} \), \( \{ E_2, S_1, S_2, Y \} \) combinations. A spurious association, \( \{ E_2, E_3, S_1, Y \} \) (\( p = 0.011 \) for the KWII and \( p = 0.009 \) for Poisson regression), was detected by both methods. Further, the spurious associations \( \{ E_3, S_2, S_4, Y \} \) and \( \{ E_3, S_1, S_2, S_3, Y \} \) were identified by Poisson regression but were not present among the top-ranked KWII values for this model.

As noted in the Materials and Methods section, the saturated Poisson model with fourth-order interactions did not execute and was terminated after 24 h. This suggests that the KWII-based method may be useful in situations where the combinatorial explosion in the number of interactions limits Poisson regression.

Table 3 shows the power analysis for the two methods. For Model 3, both methods correctly detect all four predictors as informative first-order interactions, with the KWII showing a distinct advantage at an effect size of 0.25 for the three combinations shown. Poisson regression detected the \( E_1 \) predictor at higher power until both methods converged at an effect size of 1.0 (data not shown). In line with the first-order results, Poisson regression detects the second-order combinations containing \( E_1 \) at higher power than the KWII, while the KWII reports interactions between \( E_2 \) and both informative SNPs at higher power (data not shown). For the third-order interactions, both methods displayed similar power for the \( \{ E_2, S_1, S_2, Y \} \) combination, which was the only combination the two methods had in common at this order. Poisson regression was unable to detect the full fourth-order interaction at sufficient strength to make it to the final model.
The high level of concordance between the KWII and the established Poisson regression method across three diverse models demonstrates the validity and usefulness of the KWII as a complementary analysis tool for GGI and GEI analysis. Additionally, our power analysis suggests that the KWII may outperform Poisson regression at small effect sizes.

**Analysis of Real Datasets**

Defensive Burying Dataset

Figure 4a shows a probability-probability (P-P) plot of the observed distribution of the number of shocks to the probability predicted by the Poisson distribution. The linearity and approach to the line of identity (slope 0.99) demonstrate that the Poisson distribution is a reasonable approximation for the data. After permutation testing of the phenotype to attain significance thresholds, five first-order combinations remained significant at p < 0.05: DXWox30, D14Rat71, D5Rat51, D9Rat75, and sex. Figure 5c shows bar graphs for the top four significant combinations, determined by p-value. These bar graphs clearly show that the KWII identifies combinations with differences in the phenotype that can be found on direct graphical inspection of the experimental data. In particular, the presence of both the DXWox30 and sex as significant predictor values suggests a significant sex effect, mediated by the X chromosome.
Fig. 5. a P-P plot for the total cecum inflammation score phenotype. The predicted cumulative probability from a Poisson distribution is plotted against the observed cumulative probability. The solid line is the regression line through the origin. b The top five first- and second-order markers with the highest KWII values for the total cecum inflammation score; the p-values for each combination are shown against the bars. c P-P plot for the distal colon (filled circles) and proximal colon (open circles) inflammation score phenotypes. The predicted cumulative probability from a ZIP distribution is plotted against the observed cumulative probability. d, e The top five first- and second-order markers with the highest KWII values for the distal and proximal colon inflammation scores, respectively.
We compared the findings from the information theory approach to findings reported by previous studies of this dataset [21, 26, 27]. After full X chromosome genotyping, Ahmadiyeh et al. [21] found that the largest association of any X-chromosomal marker with shock-probe approaches was DXWox30. All of the autosomal chromosomes found to have significant predictors of shock-probe approaches in the defensive burying test have also been implicated in other studies. Through genome-wide scans, Solberg et al. [27] found associations with depression-like behavior on rat autosomes 5, 9, and 14 as well, albeit with different markers. Further, Baum et al. [26] also implicate rat autosome 9 in 'hypoactivity'. Although different studies use different measurable variables, experimental approaches, and specific inbred strains, the concordance of our results across multiple studies indicates utility of the KWII to detect significant interactions in real datasets.

Additionally, we performed Poisson regression on the five first-order combinations detected as significant by the KWII at the p < 0.05 threshold and included the full results as online supplementary material (for all online suppl. material, see www.karger.com/doi/10.1159/000337934).

Main-effects models were run for these combinations. For the five first-order combinations, only DXWox30 emerged as significant in a main-effect model for Poisson regression (online suppl. table S1), suggesting that the KWII may outperform Poisson regression.

Colitis Dataset

For the colitis dataset, we analyzed the total cecum as well as the proximal and distal inflammation scores as phenotypes.

The linearity (slope = 0.995) of the P-P plot of the observed distribution of the total cecum score (fig. 5a) demonstrates that the Poisson distribution provides a satisfactory approximation for the total cecum score phenotype. The corresponding P-P plots for the proximal, medial, and distal colon scores indicated that these scores were not Poisson distributed. We investigated whether the distribution of these scores could be approximated by the ZIP distribution. The linearity of the P-P plot of the observed distribution of the distal (slope = 0.993) and proximal colon (slope = 0.999) inflammation scores in figure 5c demonstrates that the ZIP distribution provides a suitable approximation. The P-P plot of the medial colon inflammation scores was similar to that of the proximal and distal colon inflammation scores; however, in the interest of brevity, we chose to analyze only the proximal and distal colon inflammation scores, as previous results indicated the potential to detect site-specific signals with these two phenotypes. Many real count datasets are often over-dispersed relative to a Poisson, and this provided a context wherein the effectiveness of the KWII approach for over-dispersed distributions could be assessed.

The predictors diarrhea, D3Mit348, D3Mit189, D3Mit212, and D3Mit19, corresponding to the top five first-order interactions for the total cecum score (fig. 5b), were concordant with the first-order predictors associated with inflammation scores for the distal (fig. 5d) and proximal colon (fig. 5e). The five predictors with the highest KWII values for the medial colon KWII (data not shown) also overlapped completely with the proximal and distal colon predictors.

Additionally, we performed Poisson regression on the five first- and second-order combinations with the highest permutation-based KWII p-values (online suppl. material). Main-effects Poisson regression models were run for all first-order combinations. For second-order combinations, the Poisson regression models included main effects from each predictor in the combination, as well as an interaction term. All combinations tested for this dataset were significant for both methods, further indicating the KWII as a complementary tool for searching combinatorial spaces for informative GGI and GEI interactions.

As diarrhea is known to be a common symptom of IBD, the presence of this predictor is to be expected. We also compared our results to the original analysis of the data done by Farmer et al. [24]. When analyzing the scores of the individual sections of the colon, they listed only one predictor significant at the genome-wide level, D3Mit348. Furthermore, D3Mit348 was the marker that gave the maximum LOD score for the colon and was also significant in the analysis of the cecum. The top three chromosome 3 markers identified by the KWII were also found to be significant at the genome-wide level in the study by Farmer et al. [24], whereas D3Mit19 was identified as suggestive by Farmer et al. as it was just shy of the significance threshold.

For the total cecum score, D3Mit348 was a significant predictor and D8Mit191 had a suggestive association in the analysis by Farmer et al. [24]. Our method also identified D3Mit348 (p < 0.0001) and D8Mit191 (p = 0.0031). Additionally, Farmer et al. [24] also found the second-order interaction between D8Mit94 and D18Mit124 identified by our analysis.

For the distal colon, D2Mit62 was identified as significant, whereas D1Mit156 was identified as suggestive by
Farmer et al. [24]. The KWII approach in our study (data not shown) also identified both markers for the distal section of the colon alone (p = 0.0039 and p = 0.0093, respectively).

The concordance of our findings with the conclusions from Farmer et al. [24], especially the confirmation of D1Mit156 and D2Mit62 as predictors of distal colon inflammation scores, and of D8Mit191 and D8Mit94*D18Mit124 as unique first- and second-order predictors of the total cecum score, demonstrates the effectiveness of our entropy-based methods as a complementary tool for searching combinatorial spaces for informative GGI and GEI interactions.

Discussion

This report focuses on extending an entropy-based methodology to count and rate data phenotypes. We demonstrated that the KWII was capable of characterizing interactions in count/rate data in three challenging simulations and in experimental datasets from two models of different human diseases. We also extended and applied the method for over-dispersed count data. The results complement our previous work with discrete and normally distributed quantitative traits [9, 10] demonstrating the use of the methodology for GGI/GEI search and detection.

The Poisson distribution is fundamental to count and rate data modeling. In Poisson regression analysis, a predefined Poisson likelihood function is maximized to obtain parameter estimates that measure the contribution of each variable to the outcome of interest. In three models with increasingly complex GGI/GEI, the KWII approach performed comparably to a saturated Poisson regression model, outperforming or identifying a more parsimonious set of predictors than the regression model in some cases. At the smaller effect sizes examined, i.e. 0.25 and 0.5, the KWII displayed higher power for combinations detected by both methods for all three models. The Poisson method becomes computationally onerous for large datasets because it requires model specification and because the number of interaction terms increases at a combinatorial rate. For example, we used a modified approach for Model 3 because saturated models with more than eight predictor variables did complete within 24 h on the hardware used. Our entropy-based approach requires no structural model specification and removes the contribution of lower-order interactions from higher-order interactions [9]. As a result, our methodology provides a platform capable of efficiently searching large datasets for interactions without the need for repetitive model specification.

The Poisson distribution is sometimes limiting because it is a one-parameter distribution whose variance is equal to its mean. Often in real count data, the variance is larger than the mean – such data is considered ‘over-dispersed’ relative to the Poisson. A common cause of over-dispersion is simply the presence of increased density at zero. We therefore selected the ZIP distribution as a prototypical distribution to demonstrate the generalizability of our method to over-dispersed count data. Integration of the ZIP distribution to our method was straightforward and greatly increases the scope of real data that could be successfully analyzed using our methodology. Moment estimators were used because simple, closed-form expressions exist for their direct calculation from the data. The maximum likelihood estimators for the ZIP distributions do not have a closed-form solution and require obtaining the roots of a pair of non-linear equations. The moment estimators were deemed adequate for displaying proof of concept for our methodology; however, we will incorporate the maximum likelihood estimators in the future. The moment estimates could be used as initial values for a non-linear root finding algorithm, e.g. the Newton-Raphson algorithm [28].

We analyzed four different outcome variables (phenotypes) from a single rodent dataset of colitis and found an excellent concordance of our results with the published literature. Three phenotypes (the proximal, medial, and distal colon inflammation scores) were adequately described by a ZIP distribution, whereas the fourth phenotype (total cecum score) was Poisson distributed. Previous studies have indicated overlapping genetic associations with severity in these phenotypes across gastrointestinal sections, as well as some unique interactions within specific sections, and this dataset provided a good opportunity to investigate both the Poisson and ZIP methods. KWII analysis of the total cecum score identified significant first-order interactions with fecal state and multiple markers from chromosome 3 that were common across the different sections of the colon, as well as the first-order {D8Mit191, Y} interaction and the second-order {D8Mit94, D18Mit124, Y} interaction that were unique to the cecum. Farmer et al. [24] also reported these findings. The D2Mit62 and D1Mit156 markers for the distal colon inflammation score reported by these authors were also uniquely detected by the KWII analysis of the distal colon in this study.
The limitations of our approach result in part because it is built on entropy. Entropy is an aggregate characteristic of a distribution. When compared to Poisson regression and other exhaustive search methods, the KWII does not identify the precise genotype and genotype combinations causing an association. However, the risk-increasing or protective effects associated with each genotype and genotype combination can be easily identified by direct interrogation of the significant predictors found by the KWII using summary tables and graphical inspection. The differences between the KWII and regression methods arise because regression methods fit the model containing interactions to the data simultaneously.

Permutation testing, which enables calculation of the null distribution of any statistic of interest, was used to assess the significance of the KWII values by randomly permuting the phenotype values. The number of permutations necessary increases inversely with the type 1 error threshold. The computational burden of this approach for permutation testing can be particularly onerous for genome-wide dataset analyses, which have large numbers of predictor combinations and use very low type 1 thresholds to avoid false positives. However, many count datasets can be expressed as contingency tables, and in such cases, a fast exact permutation testing method based on an efficient table generation algorithm due to Patefield [29] can be used. We have implemented the fast permutation testing algorithms for discrete data [30] but have not yet integrated the algorithm for count data. Furthermore, in our approach, permutation analysis can be limited to a smaller subset of combinations with the highest KWII values.

In conclusion, we have successfully extended the KWII to characterize interactions in count/rate data through simulations and the analysis of two real datasets. The findings suggest that information theory methods may provide a versatile, model-independent framework for GEI analysis appropriate for a range of different classes of phenotypes.

Acknowledgements

Funding from Pfizer fellowship to Jonathan Knights 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.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


11 Han TS: Multiple mutual informations and multiple interactions in frequency data. Information and Control 1980;46:26–45.


