Statistical Interactions from a Growth Curve Perspective

Sean M. Devlin Jaya M. Satagopan

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Keywords

Statistical interaction · Growth curve

Abstract

Logistic regression is widely used to evaluate the association between risk factors and a binary outcome. The logistic curve is symmetric around its point of inflection. Alternative families of curves, such as the additive Gompertz or Guerre-ro-Johnson models, have been proposed in various scenarios due to their asymmetry: disease risk may initially increase rapidly and be followed by a longer period where the rate of growth slowly decreases. When modeling binary outcomes in relation to risk factors, an additive logistic model may not provide a good fit to the data. Suppose the outcome and an additive function of the risk factors are indeed related through an asymmetric function, but we model the relationship using a logistic function. We illustrate – both from a mathematical framework and through a simulation-based evaluation – that higher-order terms, such as pairwise interactions and quadratic terms, may be required in a logistic regression model to obtain a good fit to the data. Importantly, as significant higher-order terms may be a manifestation of model misspecification, these terms should be cautiously interpreted; a more pragmatic approach is to develop contrasts of disease risk coming from a good fitting model. We

illustrate these concepts in 2 cohort studies examining early death for late-stage colorectal and pancreatic cancer cases, and 2 case-control studies investigating NAT2 acetylation, smoking, and advanced colorectal adenoma and bladder cancer.

© 2017 S. Karger AG, Basel

Introduction

Epidemiology studies often examine risk factors in relation to binary outcomes. Researchers must make 2 choices when conducting a statistical analysis of these data: (1) choose a link function or a scale for the outcome and then (2) select a model or a function of the risk factors. A statistical model then relates the outcome, measured on the chosen scale, to the chosen function of the risk factors. The most common choice of scale for binary outcomes is the canonical logistic link function, and the resulting analysis is referred to as a “logistic regression model” [1]. A link function specifies the relative rate at which disease risk increases as a function of the risk factors. A model for the risk factors specifies the parameterization of the factors of interest; this typically includes additive effects of the individual risk factors, and may also include higher-order polynomials of a risk factor or prod-

uct terms of 2 or more risk factors. The additive effects are commonly referred to as “main effects.” The product terms are referred to as “statistical interactions” or “interaction terms.” When the model does not include a statistical interaction, it means that the effect of a risk factor on the outcome (under the chosen scale) does not depend upon the value of the other risk factors. By contrast, when a model does include a statistical interaction, it means that the effect of a risk factor depends upon the value of the other risk factors involved in the interaction.

Statistical interactions have received particular attention in genetics and health sciences because of the anticipation that such interactions play an important role in disease etiology and can provide insights into underlying biological mechanisms of action [2, 3]. However, a practical challenge is how to interpret interactions occurring in statistical models, notably logistic regression models, because the occurrence of a statistical interaction can depend upon the scale on which the outcome is measured [4, 5]. The overarching goal of this paper is to explore this issue and obtain insights into the need for interactions in a logistic regression model.

The logistic curve is an S-shaped curve that is symmetric around the point of inflection [6]. Studies of tumor growth and plant disease epidemics have noted that the symmetry assumption in the logistic curve does not fit these data well as the rate of increase of disease risk is initially rapid, followed by a slower rate after the inflection point than the rate postulated by a logistic curve [7, 8]. This can lead to a potential under- and over-estimation of the true disease risk at the lower and upper tails, respectively, of the risk distribution, when we fit an additive logistic regression model to the data. Alternative additive models, such as an additive model under the Gompertz scale of risk, have been shown to better fit these scenarios [7, 8]. Throughout this paper, we use the term “additive model” to refer to a model that contains additive effects of the risk factors, and does not contain higher-order polynomial terms such as power functions or product terms of the risk factors.

Since the true rate at which disease risk increases is unknown in practice, we almost always model binary data using the logistic link function; the estimation approach finds the best fitting curve within the logistic family to the data. The thesis of this paper is that, when a function of disease risk is additive in the risk factors and it increases more (or less) rapidly than that postulated by a logistic link function, an additive logistic regression model will not provide an adequate fit to the data, and higher-order terms such as interactions may be required

to obtain well-calibrated estimates of disease risk. These interaction terms must be interpreted cautiously in the event that their significance may simply be a manifestation of model misspecification, i.e., misspecification of the outcome’s scale. Instead of directly interpreting the interaction terms, clinically meaningful contrasts of disease risk can alternatively be estimated using the predicted values of the regression model as noted in the Discussion section.

Our objectives in this paper, therefore, are to obtain insights into the need for and interpretation of statistical interactions in logistic regression models for binary outcomes in the context of certain mathematical growth models of disease risk. We use the Gompertz and the Guerrero-Johnson (henceforth abbreviated as GJ) family of curves to describe these methodological concepts. In the remainder of this paper, we investigate these objectives, review some mathematical structures and simulations in support of these concepts, and provide illustrative examples.

Methods

For simplicity of exposition, we initially illustrate the concepts using a single risk factor x , which may be continuous, ordinal, or a set of indicators corresponding to a categorical variable. We denote disease probability as p . Under the law of uncontrolled growth of disease risk, the relative rate of growth of p with respect to x is a constant [6, 9]; this is given by the differential equation:

$$\frac{1}{p} \frac{dp}{dx} = \alpha. \quad (1)$$

Integrating Equation 1, we obtain an exponential model for disease risk given by: $p = \exp\{\mu_0 + \alpha x\}$, where μ_0 is a constant. This model is additive under the natural logarithm scale of risk, i.e., $\log(p) = \mu_0 + \alpha x$. Under this exponential model, risk increases monotonically in x . While such an increase may be theoretically feasible, it is rarely observed in practice. Empirical studies suggest that as risk continues to increase, the rate of increase slows down and the magnitude of risk reaches an upper asymptote. This leads to an S-shaped pattern for risk in relation to the risk factors.

The rate at which the increase slows down is a defining feature of disease risk curves. There are numerous S-shaped curves, the most notable among them being the logistic curve. Of other S-shaped curves discussed in the literature, the Gompertz curve has received particular attention in studies of cancer [8]. The GJ family of curves, which includes the logistic curve as a special case, has also been studied in the statistics literature [10]. Below we summarize additive models based on the logistic, Gompertz, and GJ curves, and the role of higher-order effects, such as interactions, occurring in a logistic model.

Additive Logistic Model for Risk

We denote c_m as the upper asymptote for the disease probability p , i.e., $p \rightarrow c_m$ as $x \rightarrow \infty$. The relative rate of increase of disease risk slows down as $p/c_m \rightarrow 1$, implying that the relative growth rate is no longer a constant. One way to quantitatively represent relative growth rate in this setting would be to specify the relative growth rate as:

$$\frac{1}{p} \frac{dp}{dx} = \alpha \left(1 - \frac{p}{c_m} \right), \quad (2)$$

thus, the relative growth rate approaches 0 as p/c_m approaches 1. This provides a logistic model for disease risk given by:

$$p = \frac{c_m}{1 + e^{-(\mu_0 + \alpha x)}}, \quad (3)$$

where μ_0 and α are interpreted as the intercept and the effect (on the log odds scale) of x . This model is additive on the logistic scale of disease risk, i.e.,

$$\log \left(\frac{p/c_m}{1 - p/c_m} \right) = \mu_0 + \alpha x.$$

Additive Gompertz Model for Risk

The Gompertz model is characterized by a relative growth rate that decreases in $\log(p/c_m)$ [11]. This relative growth rate is given by:

$$\frac{1}{p} \frac{dp}{dx} = \alpha \left\{ 1 - \log \left(\frac{p}{c_m} \right) \right\}. \quad (4)$$

The decrease in terms of $\log(p/c_m)$ results in a period of initial fast growth for small values of p and a subsequent rapid decline in the rate as p increases, providing the characteristic asymmetric shape of the Gompertz curve with disease risk given by:

$$p = c_m e^{-e^{-(\mu_0 + \alpha x)}}. \quad (5)$$

This model is additive on the complementary log-log scale of risk:

$$-\log \left\{ -\log \left(\frac{p}{c_m} \right) \right\} = \mu_0 + \alpha x. \quad (6)$$

Henceforth, we refer to Equation 5 as the additive Gompertz model.

Additive GJ Model for Risk

Guerrero and Johnson [12] proposed a family of power transformations of the odds ratio that included the logistic model as a special case. Under this family, the relative rate of growth of p , indexed by a power parameter λ , is given by:

$$\frac{1}{p} \frac{dp}{dx} = \alpha \left(1 - \frac{p}{c_m} \right) \left(\frac{\frac{p}{c_m}}{1 - \frac{p}{c_m}} \right)^{-\lambda}. \quad (7)$$

While the logistic and the Gompertz models are constrained by the same number of parameters (μ_0 and α), the GJ family of curves is also indexed by the transformation parameter λ . Different values

of λ provide different members of the GJ family of curves. It includes the logistic link function as a special case when $\lambda = 0$. The properties of the relative growth rate and, hence, the properties of the risk distribution depend upon λ . In general, the growth rate of the GJ curve is initially more rapid than the logistic family when $\lambda < 0$, and initially slower than postulated by the logistic family when $\lambda > 0$. While the maximum disease probability, c_m , was not included in the original formulation, it is included here for notational consistency. Disease probability is defined in this family as:

$$p = c_m \left\{ 1 + \left[1 + \lambda (\mu_0 + \alpha x) \right]^{-1/\lambda} \right\}^{-1}, \text{ for } \lambda \neq 0. \quad (8)$$

Henceforth, we refer to Equation 8 as an additive GJ model. In all models of risk, the upper asymptote, c_m , is generally taken to be 1. Further, note that under the GJ link, the risk p is in the desired range (0, 1) when $x \in (-1/\lambda, \infty)$.

Interaction Effects

Given a binary outcome and a risk factor x , we can postulate a binomial distribution for the outcome and, in principle, assume disease risk p to be additive in x under some family – for example, the logistic, the Gompertz, or the GJ family of curves. If the disease rate is asymmetric in relation to x , as shown for the Gompertz and the GJ family of curves in Figure 1, then an additive model based on a relevant asymmetric link function may provide a better fit to the data than an additive logistic model. Since it is generally unknown at the outset whether the rate is symmetric or asymmetric with respect to x , we may use a goodness of fit test to choose between additive logistic, additive Gompertz, and additive GJ models [13]. However, logistic models are more popular than other models for fitting binary outcomes, as the logistic function is the canonical link function for analyzing binary data, and logistic regression methods are available in most statistical software packages. Further, parameter estimates from logistic models can be readily interpreted in terms of odds ratios, and they are more widely understood by clinical investigators.

Suppose the true model for growth is an additive Gompertz model (Equ. 5). In this setting, the logistic function can be approximated as (see Appendix for derivation):

$$\log \left(\frac{p}{1-p} \right) \approx \theta_0 + \sum_{j \geq 1} \theta_j x^j. \quad (9)$$

This argument also applies when the true model is additive under the GJ family. To see this, note that, when the true model is an additive GJ model (Equ. 8), the logistic link function can be written as:

$$\begin{aligned} \log \left(\frac{p}{1-p} \right) &= \frac{1}{\lambda} \times \log \left\{ 1 + \lambda (\mu_0 + \alpha x) \right\} \\ &\approx \theta_0 + \sum_{j \geq 1} \theta_j x^j. \end{aligned} \quad (10)$$

The approximation in the last step follows from a Taylor series approximation to the natural logarithm term on the right hand side of the first step. It takes the same form as the right hand side of Equation 9 in the sense that the logistic link function will include higher-order terms when the true model is additive under the Gompertz or the GJ model.

When x is binary, the right hand side of Equation 9 is additive in x , and we cannot distinguish between additive logistic, additive Gompertz, and additive GJ models. To see this, note that when x

Fig. 1. The logistic curve, Gompertz curve, and 2 curves from the Guerrero-Johnson family. The intercept and slope for all curves are set at 0 and 1, respectively.

is binary, the plot of x versus p is essentially a straight line, and one cannot distinguish between curvatures of the logistic, Gompertz, or GJ type. When x is not binary, Equation 9 is not additive in x . If one were to fit a logistic regression model to the data, then including at least a quadratic term in the model (i.e., a model with a right hand side of the form $\theta_0 + \theta_1 x + \theta_2 x^2$) can provide a better fit to the data than an additive logistic model.

In practice, we often have more than one risk factor. Denoting $x_1 + x_2, \dots, x_M$ as M risk factors of interest, the additive Gompertz model (Equ. 5) and the additive GJ model (Equ. 8) can be written by replacing the term αx with the summation term

$$\sum_{j=1}^M \alpha_j x_j$$

corresponding to the additive effects of the risk factors. Following the above arguments, when we fit a logistic regression model to the data, product terms of the risk factors (i.e., statistical interaction terms) will appear in the model, and including at least pairwise interaction and quadratic terms in a logistic regression model will provide a better fit to the data than an additive logistic model. In this setting, a logistic regression model with pairwise first-order interaction terms takes the form:

$$\log\left(\frac{p}{1-p}\right) = \theta_0 + \sum_{j=1}^M \theta_j x_j + \sum_{j=1}^M \sum_{k \geq j}^M \delta_{jk} x_j x_k. \quad (11)$$

Thus, when the true model for disease risk is an additive Gompertz model or an additive GJ model, interaction terms can arise in a logistic model for risk.

However, the converse that disease risk is an additive Gompertz model when there are interactions in a logistic model is not true. Consider 2 risk factors x_1 and x_2 , such that at least one of them is not binary. Suppose we plot the risk curve as a function of x_1 for

each value of x_2 . When these curves cross each other, it means that an interaction between x_1 and x_2 is qualitative. When an interaction occurring under a logistic regression model is qualitative, there is no link function under which the model is additive [14]. In other words, the simple presence of higher-order terms in a logistic model does not by itself imply additivity under an alternative growth model.

In the following section, we conduct simulation studies to demonstrate that the inclusion of higher-order polynomial or interaction terms is required to obtain a better fit to the data when we fit a logistic regression model to data arising from an additive Gompertz model.

Simulations

To illustrate how misspecification of the growth curve may warrant higher-order terms to improve overall fit, we pursued simulations under an additive Gompertz model for risk with 2 covariates: a uniformly distributed continuous covariate $X_1 \sim U(0, 30)$, and a binary covariate X_2 with $P(X_2 = 1) = 1 - P(X_2 = 0) = 0.5$. The true disease risk follows an additive Gompertz model for risk:

$$p = c_m e^{-e^{-(\mu_0 + \alpha_1 X_1 + \alpha_2 X_2)}}. \quad (12)$$

This additive model was evaluated under the following configurations: $c_m = 0.5$, $\mu_0 = -2$, $\alpha_1 = 0.15$, and $\alpha_2 \in \{0.0, 0.5, 1.0, 1.5, 2\}$. These Gompertz curves are shown in the Appendix (Figure A1).

We generated data under 2 commonly used epidemiological designs: cohort sampling and case-control sampling. Under cohort sampling, the true underlying Gompertz model generated the disease probability p using X_1 and X_2 for N independent individuals. A Bernoulli distribution declared each person to have the disease with a probability of p . Under the case-control sampling, the underlying Gompertz model generated a large cohort consisting of individuals with and without the disease, and N_D cases and $N_{\bar{D}}$ controls are sampled from this cohort. Under both sampling schemes, 2 total sample sizes were considered: $N = 300$ and $N = 600$. For the case-control sampling, we set $N_D = N_{\bar{D}} = 150$ and 300, and for the total sample size, we set $N = N_D + N_{\bar{D}} = 300$ and 600, respectively. A total of 5,000 independent data sets were generated under the parametric configuration shown in Equation 12. In each data set, the following 4 logistic regression models were fitted:

- M1: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (additive model),
- M2: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2$ (quadratic model),
- M3: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_1 X_2$ (interaction model), and
- M4: $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2$ (quadratic and interaction model).

Table 1. Simulation results under the additive Gompertz model for a sample size of 300 individuals

	Cohort (N = 300)					Case-control (N = 300)				
N_D	66.7	74.6	82.1	89.6	95.3	150	150	150	150	150
α_2	0.00	0.50	1.00	1.50	2.00	0.00	0.50	1.00	1.50	2.00
<i>Relative efficiency (MSE)</i>										
M1/M2	1.67	1.60	1.33	1.11	1.00	2.52	2.08	1.52	1.18	1.03
M1/M3	0.91	0.94	1.05	1.26	1.55	0.89	0.94	1.09	1.31	1.69
M1/M4	1.39	1.47	1.56	1.68	1.81	2.02	1.99	2.01	2.03	2.16
<i>AUC</i>										
M1	0.78	0.76	0.74	0.72	0.71	0.78	0.76	0.74	0.72	0.71
M2	0.78	0.76	0.74	0.72	0.71	0.78	0.76	0.74	0.72	0.71
M3	0.78	0.76	0.74	0.72	0.71	0.78	0.76	0.74	0.72	0.71
M4	0.78	0.76	0.74	0.73	0.72	0.78	0.76	0.74	0.72	0.72
<i>AIC</i>										
M1	270	293	315	332	343	340	354	365	373	377
M2	265	289	311	330	343	331	346	360	371	376
M3	271	294	314	329	337	341	354	363	368	369
M4	266	289	309	325	335	332	346	357	364	366
<i>Reject $\beta_3 = 0$ (proportion)¹</i>										
M1										
M2	0.67	0.64	0.52	0.33	0.19	0.89	0.83	0.64	0.39	0.23
M3										
M4	0.69	0.68	0.62	0.50	0.38	0.90	0.86	0.76	0.60	0.47
<i>Reject $\beta_4 = 0$ (proportion)¹</i>										
M1										
M2										
M3	0.02	0.07	0.26	0.55	0.78	0.03	0.12	0.37	0.69	0.86
M4	0.04	0.13	0.38	0.68	0.85	0.04	0.19	0.54	0.82	0.93

All results are the average over 5,000 replicates. In all simulations the following parameter values were used: $\mu_0 = -2$, $\alpha_1 = 0.15$, and $c_m = 0.5$. ¹ Based on a two-sided test at $\alpha = 0.05$. Models M1, M2, M3, and M4 are defined in the Simulations section.

Thus, we model the data using the logistic growth curve when the true underlying model is an additive Gompertz curve given by Equation 12. For large values of α_2 , we would expect statistically significant interactions under models M2, M3, and M4; we examine this in the simulations described below.

Various metrics were used to compare the 4 fitted regression models (M1, M2, M3, and M4). The relative efficiency of each model compared to M1 was calculated as the ratio of the mean squared errors (MSE). MSE, a measure of calibration of risk, was calculated as the average difference between the true disease probability coming from the underlying Gompertz model and the estimated disease probability from M1, M2, M3, and M4. A small MSE is an indication of a well calibrated risk estimate. The area under the receiver characteristic operating curve

(AUC) and the Akaike information criterion (AIC) were also included as 2 commonly used metrics to assess discrimination and overall model fit. The AUC measures the degree of separation in the estimated disease probability between individuals with and those without the disease; values closer to 1 denote improved discrimination. The AIC, a function of the estimated log-likelihood penalized by twice the number of parameters, is used to contrast the goodness-of-fit of competing models; lower AIC values denote improved model fit. Lastly, the proportion of times the null hypothesis, $H_0: \beta_3 = 0$, was rejected against the alternative hypothesis, $H_A: \beta_3 \neq 0$, in M2 and M4, and the proportion of times the null hypothesis, $H_0: \beta_4 = 0$, was rejected against the alternative hypothesis, $H_A: \beta_4 \neq 0$, in M3 and M4 were calculated as the number of times, out of 5,000 simulations, the corresponding p values were

<0.05 using a two-sided Wald test. These proportions represent how frequently significant quadratic and interaction terms occur due to misspecification of the underlying growth curve.

Results

Table 1 provides the simulation results. The first column in Table 1 provides the estimated metrics for cohort sampling when $\alpha_2 = 0$, indicating no effect of X_2 , and $N = 300$. Across 5,000 simulations, on average 66.7 individuals had the disease. In terms of relative efficiency, model M2 with the quadratic term (X_1^2) performed better than M1, while M3 performed worse. This latter poor performance of M3 is expected, as the model includes an interaction between X_1 and X_2 , when X_2 is not associated with disease risk. The AUCs for the 4 models were the same with a value of 0.78. Further, in terms of AIC values, M2 had the lowest AIC value, while M3 had the highest. The null hypothesis, $H_0: \beta_3 = 0$, was rejected 67 and 69% of the time for M2 and M4, respectively. The null hypothesis, $H_0: \beta_4 = 0$, was rejected close to 5% of the time, the nominal type I error rate.

Comparing across the different columns of cohort sampling with $N = 300$, larger values of α_2 resulted in larger relative efficiency gains for models M3 and M4 compared to M1. In terms of the AIC, M4 provided the lowest values, although M3 had an AIC of similar magnitude for higher values of α_2 . The proportion of times $H_0: \beta_3 = 0$ was rejected decreased for higher values of α_2 , while $H_0: \beta_4 = 0$ was more frequently rejected for these higher values. As the effect of X_2 in the true Gompertz model in Equation 12 increases (i.e., as α_2 increases), the effect of the interaction term occurring under the misspecified logistic growth model increases, resulting in higher power to reject $H_0: \beta_4 = 0$. Since α_1 is held at a constant value in our simulations ($\alpha_1 = 0.15$), the effect of X_1 relative to X_2 diminishes as α_2 increases. Hence, the power to reject $H_0: \beta_3 = 0$ decreases as α_2 increases. For the largest value of α_2 ($= 2$), M4 rejected both $H_0: \beta_3 = 0$ and $H_0: \beta_4 = 0$ more often than the 2 models that individually included the separate higher-order terms (M2 and M3). Similar results were observed for simulations with $N = 600$, as shown in Appendix Table A1.

Comparing case-control sampling to cohort sampling, the gains in relative performance were noticeably greater for case-control sampling because of the considerably larger number of cases for a given N under the design. The proportions of times β_3 and β_4 were rejected increased as well, in a similar pattern to cohort sampling.

Taken together, these results illustrate that when the true model for risk is additive under a link function that is different from the logistic link (which is the additive Gompertz model here), interaction terms will be required to obtain a good fit when we model the data using logistic regression.

Data Analysis

This section presents 2 sets of examples illustrating that the inclusion of higher-order terms allows curves within the logistic family to more closely align with the underlying empirical estimates. The first set of examples includes 2 cohort studies using data from the Surveillance, Epidemiology, and End Results (SEER) registry [15]. The second set of examples includes 2 previously published case-control studies investigating the association of N-acetyltransferase 2 (NAT2) and smoking status with bladder cancer and advanced colorectal adenoma [16, 17]. When testing the null hypothesis that a model parameter is 0 against a 2-sided alternative in all these data sets, we report statistical significance at a 5% level.

Cohort Sampling: SEER Cancer Registry

Using SEER, we identified individuals who were diagnosed with AJCC stage 4 colorectal and pancreatic cancers from 2000 to 2010 [15, 18, 19]. Selected individuals were diagnosed between the ages of 40 and 90, had no prior diagnosis of cancer within the SEER registry, and were recorded in the registry as either White or Black race. The binary outcome for all analyses was early death, defined as a recorded death within 3 months following diagnosis. There was a small number of patients who were lost to follow-up within the first 3 months (0.2–0.3%); these patients were included in the analysis and considered not to have died within the first 3 months.

The 4 logistic regression models (M1–M4; see the Simulations section) were fitted to each data set. The covariate X_1 is the age at diagnosis, which was recoded as the age in 10-year increments starting from 40 years old. The covariate X_2 is the sex of the individual (0 = female, 1 = male). For both studies, models M1–M4 were separately fitted for the 2 selected race categories (White, Black). Unlike the simulated data in the Simulations section, the true disease probability is unknown; therefore, the MSE is replaced by the Brier score to assess model fit. The Brier score is a measure of risk calibration like the MSE, and is defined as the mean squared difference between the estimated probability and the binary disease state [20].

Table 2. Metrics of model assessment for the 4 models of early death among adults with colorectal and pancreatic cancer diagnosed between 40 and 90 years old under additive and higher-order logistic models and the additive Guerrero-Johnson (GJ) and Gompertz models for disease risk

	Logistic models				Additive models	
	M1	M2	M3	M4	GJ	Gompertz
<i>Brier score</i>						
Colorectal cancer						
Black race	0.1798	0.1795	0.1798	0.1795	0.1796	0.1797
White race	0.1694	0.1691	0.1692	0.1690	0.1691	0.1692
Pancreatic cancer						
Black race	0.2338	0.2337	0.2336	0.2336	0.2338	0.2338
White race	0.2349	0.2342	0.2349	0.2342	0.2341	0.2346
<i>AUC</i>						
Colorectal cancer						
Black race	0.6558	0.6560	0.6560	0.6559	0.6557	0.6557
White race	0.6893	0.6894	0.6901	0.6901	0.6893	0.6893
Pancreatic cancer						
Black race	0.6267	0.6276	0.6280	0.6282	0.6267	0.6267
White race	0.6414	0.6416	0.6416	0.6415	0.6413	0.6414
<i>AIC</i>						
Colorectal cancer						
Black race	5,870	5,864	5,870	5,865	5,867	5,866
White race	29,994	29,957	29,964	29,935	29,960	29,974
Pancreatic cancer						
Black race	4,495	4,495	4,494	4,495	4,497	4,495
White race	27,313	27,249	27,310	27,250	27,246	27,282

The estimated values of λ for the additive GJ model are: -0.38 (Black race, colorectal cancer); -0.35 (White race, colorectal cancer); 0 (Black race, pancreatic cancer; equivalent to an additive logistic model); -0.79 (White race, pancreatic cancer). The logistic models M1, M2, M3, and M4 are defined in the Simulations section.

AUC and AIC were additionally estimated to assess goodness-of-fit and discrimination. Additivity under the Gompertz and GJ link functions were also evaluated. To estimate λ in Equation 8, we fitted the additive GJ models by fixing λ at various values, and selected the λ corresponding to the model with the smallest AIC. All our additive GJ models in the results given below were optimized for λ in this manner.

Colorectal Cancer

There were a total of 34,495 colorectal cancer cases in the analysis, of which 8,532 (25%) died within 3 months following diagnosis. A total of 5,412 (16%) cases with colorectal cancer diagnosis were Black, and the remaining were White. Appendix Table A2 presents odds ratios, confidence intervals, and associated p values for the estimated colorectal logistic regression models M1–M4. For Black individuals, the quadratic term for age was significantly different from 0 in models M2 and M4, while there

was no evidence of an interaction between age and sex in models M3 and M4. Among White individuals, the higher-order terms in models M2, M3, and M4 were all significantly different from 0.

As shown in Table 2, the significant quadratic term in models M2 and M4 for Black individuals translated to improved measures of model assessment: the Brier score and AIC were lower for models M2 and M4 compared to models M1 and M3. For White individuals, there were improvements in Brier score and AIC for all higher-order models (M2, M3, and M4) compared to the additive model M1. There were no apparent differences in AUC in the analyses of Black and White individuals.

The estimated early death probabilities for White and Black individuals for models M1–M4 are shown in Figure 2. In addition to the model-based probabilities, the empirical early death rates along with 95% confidence intervals are included for select ages. For White males and females, the estimated curves for the higher-order models

Fig. 2. Empirical and estimated probabilities of early death from models M1–M4 in the colorectal cancer analysis using SEER data. **a** Colorectal cancer in White males; **b** colorectal cancer in White females; **c** colorectal cancer in Black males; **d** colorectal cancer in Black females.

(M2, M3, and M4) tend to more closely align with the empirical estimates compared to the additive model M1. As illustrated by the larger confidence intervals around the empirical estimates for Black males and females, the smaller sample size presents a challenge when evaluating which curve most closely follows the empirical estimates.

While models M2 and M4 may more closely align with the empirical estimates for the highest displayed age, it is difficult to ascertain the best fitting model for the remainder of the age range.

Results of the additive Gompertz and additive GJ models are also shown in Table 2. For the Black and White

Fig. 3. Empirical and estimated probabilities of early death from models M1–M4 in the pancreatic cancer analysis using SEER data. **a** Pancreatic cancer in White males; **b** pancreatic cancer in White females; **c** pancreatic cancer in Black males; **d** pancreatic cancer in Black females.

colorectal cancer cohorts, the additive GJ and the additive Gompertz models achieved lower AIC than the additive logistic models. However, for the White colorectal cancer cohort, model M4 that includes both quadratic and interaction terms achieved the best AIC among all of the models considered. To see why the additive Gompertz and

additive GJ models do not provide a better fit than the logistic model M4, note that the estimated curves for the early death probabilities for men and women cross around an age at diagnosis of 65, suggesting a qualitative interaction (Appendix Figure A2); therefore, additivity will not hold under any link function.

Table 3. Metrics of model assessment for the 2 case-controls studies of NAT acetylation and smoking under the additive (M1) and interaction (M3) logistic models and under the additive Guerrero-Johnson (GJ) and Gompertz models for disease risk

	Logistic models		Additive models	
	M1	M3	GJ	Gompertz
<i>Brier score</i>				
Bladder cancer	0.2393	0.2391	0.2392	0.2393
Advanced colorectal adenoma	0.2434	0.2419	0.2421	0.2433
<i>AUC</i>				
Bladder cancer	0.5925	0.5943	0.5938	0.5925
Advanced colorectal adenoma	0.5733	0.5781	0.5753	0.5733
<i>AIC</i>				
Bladder cancer	32,382	32,369	32,368	32,379
Advanced colorectal adenoma	1,767	1,763	1,762	1,766

The estimated values of λ for the additive GJ model are: -3.2 (bladder cancer); -4.62 (advanced colorectal adenoma). The logistic models M1 and M3 are defined in the Simulations section.

Pancreatic Cancer

The pancreatic cancer cohort includes 24,022 newly diagnosed cases, of which 12,478 (52%) died within 3 months of diagnosis. A total of 3,402 (14%) cases with this diagnosis were Black, and the remaining were White. The odds ratio estimates and associated p values for the 4 logistic models (M1–M4) are shown in Appendix Table A2. None of the higher-order terms were significantly different from 0 in the logistic models for Black pancreatic cancer cases. For White individuals, the higher-order terms in models M2 and M3 were significantly different from 0. However, in model M4, only the quadratic term was significantly different from 0.

In terms of model assessment, there was a small decrease in the Brier score for the higher-order models compared to the additive model for Black individuals (Table 2). However, there was no noticeable improvement in the AICs of models M2, M3, and M4 when compared to model M1. For White individuals, models M2 and M4 had the lowest Brier scores and AIC values. There were no apparent differences in the discriminatory ability (AUC) across the 4 models.

The estimated early death probabilities for the 4 models are shown in Figure 3. For White individuals, the quadratic term for age in models M2 and M4 allows the predicted disease risk to more closely fit the empirical estimates, particularly in the 2 tails of the age distribution. Similar to the colorectal example, the smaller sample size of Black pancreatic cancer cases hinders the analysis of the evaluation of which curve most closely matches the empirical estimates. In the White pancreatic cohort, the additive GJ model performed best in the goodness-of-link

evaluation across all models (Table 2). This model also had the lowest Brier score among all of the models considered. In the Black pancreatic cohort, the additive logistic and Gompertz models had similar AIC values, which were also similar to those of the higher-order logistic models.

Case-Control Sampling: NAT2 and Smoking

Two published case-control studies have investigated the interaction between NAT2 and smoking status among cases with advanced colorectal adenoma compared to disease-free controls and among cases with bladder cancer compared to disease-free controls [16, 17]. NAT2 is categorized as a binary variable – rapid/intermediate acetylators and slow acetylators; smoking status is categorized as never, past, or current smoker.

Bladder Cancer

There were a total of 13,415 cases and 10,699 controls in the bladder cancer study. Appendix Table A3 provides the odds ratio, confidence intervals, and p values for 2 logistic regression models: the additive model (M1) and a model including NAT2 $\times$ smoking status interaction terms (M3). Since NAT2 acetylation and smoking status are both categorical, there are no quadratic models (M2 and M4) to evaluate. There was a significant interaction between NAT2 acetylation and smoking status. Consequently, model M3 has lower AIC values, a higher AUC, and a slightly lower Brier score than model M1 (Table 3). The estimated logarithms of odds (log-odds) for the 2 models are shown in Figure 4.

The additive GJ model had the lowest AIC compared to the additive Gompertz model or any of the logistic models considered (Table 3). The logistic model M3, however, had a lower Brier score and AUC because it is a saturated model. These results are similar to a previous investigation of these data [14].

Advanced Colorectal Adenoma

The advanced colorectal adenoma study included 645 cases and 649 controls. As shown in Appendix Table A3, there was a significant interaction term between smoking and NAT2 acetylation in model M3. This model appeared to better fit the data in terms of Brier score, AUC, and AIC (Table 3). Similar to the bladder cancer example, the assumption of the additive logistic model does not appear to hold (Fig. 4). In the goodness-of-link evaluation, the additive GJ model once again achieved the lowest AIC compared to the additive Gompertz model and any of the logistic models considered. Therefore, among the models considered, the additive GJ model provides the best fit. In order to get a well-calibrated logistic model, an interaction term was needed.

Discussion

The widespread use of the logistic regression model is partly due to convenience, as its implementation is available in most software packages, and partly due to the general flexibility of the family of curves. However, while flexible, the additive logistic model may not fit the true (and unknown) underlying risk model well. As illustrated in the Simulations section, higher-order terms can be introduced to accommodate – at least in part – for this lack of fit.

While misspecification of the link function can lead to higher-order terms in a logistic model, which in return leads to improved model fit, the converse need not be true: the presence of higher-order terms does not imply additivity under an alternative link function. Interaction terms may at times be needed in a model regardless of the link function choice [14]. Therefore, instead of directly interpreting the interaction effect, a pragmatic approach would be to first obtain a model that provides the best fit to the data, and use this model to examine suitable contrasts of risk (or a desired function of risk) corresponding to one risk factor according to specific levels of another risk factor. As an example, using the SEER colorectal cancer analysis, White men have a slightly higher rate of early death due to colorectal cancer compared to White

Fig. 4. Log odds ratios obtained from an additive logistic regression model (M1) and a logistic model including interaction terms (M3). Model M3 is the saturated logistic regression model; therefore, the estimated log-odds are identical to the empirical estimates. **a** Bladder cancer; **b** advanced colorectal adenoma. Acet., acetylation; Interm., intermediate.

women at age 65 (men: 20.5%; women: 18.9%; difference: 1.6%). However, at age 85, White men have a lower early death rate than White women (men: 45.3%; women: 48.8%; difference: -3.5%). This interpretation is more

clinically meaningful than interpreting the interaction odds ratio of 0.89 as provided in Appendix Table A2.

While a case-control study design precludes the estimation of the baseline disease risk without external data, differences in the log-odds can be easily interpreted. For example, as shown in Figure 4, the log-odds of advanced colorectal adenoma is higher for slow acetylators compared to rapid/intermediate acetylators among current smokers (log-odds for slow: 1.01; rapid/intermediate: 0.14; difference: 0.86). The difference in the log-odds is, however, diminished among past smokers (log-odds for slow: -0.13; rapid/intermediate: -0.07; difference: -0.06). These interpretations are again more straightforward than interpreting the interaction odds ratios as presented in Appendix Table A3. These interpretations are also particularly more desirable in this data set because of the possibility that the underlying model is additive under the GJ link, resulting in a significant interaction term in the logistic model.

In cancer research, understanding the rate of increase in disease risk is paramount for understanding disease progression, providing patient prognostication, and delivering treatment decisions. These require well-calibrated models for risk, and higher-order terms can be introduced to improve model fit beyond an additive model. Thus, we recommend interpreting higher-order terms as simply that: parameters in a statistical model required to improve model fit. Instead of a clinical interpretation of these higher-order terms, the focus should instead be on meaningful contrasts of disease risk. These contrasts can provide better insights into the rate at which risk increases; this in return can help facilitate better patient care.

Acknowledgements

This work was supported through grants P30 CA008748, R01 CA137420, and R01 CA197402 from the National Cancer Institute, USA, and grant UL1RR024996 from the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA. We thank the Associate Editor of *Human Heredity* and 3 anonymous reviewers for their insightful comments.

Appendix

Reproducibility of the Tables and Figures

The data and R code used to generate the results of tables and figures can be obtained from www.github.com/sedevlin.

Derivation of Equation 9 of the Methods Section

Without loss of generality let c_m equal 1 in Equation 5, and denote

$$p_0 = e^{-e^{-\mu_0}}.$$

Hence Equation 5 can be written as $p = p_0^{e^{-\alpha x}}$. Note that p_0 does not approach 1 as long as μ_0 does not approach ∞ . The logistic link function can be written as:

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &= e^{-\alpha x} \log(p_0) - \log\left[1 - p_0^{e^{-\alpha x}}\right] \\ &\approx e^{-\alpha x} \log(p_0) - \log\left[1 - p_0^{1 - \alpha x + \frac{1}{2}\alpha^2 x^2}\right] \\ &= e^{-\alpha x} \log(p_0) - \log\left[1 - p_0 e^{-\alpha x\left(1 - \frac{1}{2}\alpha x\right)\log(p_0)}\right] \\ &\approx e^{-\alpha x} \log(p_0) - \log\left\{\left(1 - p_0\right)\left[1 + \frac{p_0 y (1 - y/2)}{1 - p_0}\right]\right\}, \end{aligned}$$

where $y = \alpha x (1 - \alpha x / 2) \log(p_0)$. Therefore,

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &\approx \left[1 - \alpha x + \frac{(\alpha x)^2}{2}\right] \log(p_0) - \log(1 - p_0) \\ &\quad - \frac{p_0 y (1 - y/2)}{1 - p_0} \left[1 - \frac{1}{2} \frac{p_0 y (1 - y/2)}{1 - p_0}\right] \\ &\approx \log\left(\frac{p_0}{1 - p_0}\right) - (\alpha x) \log(p_0) \left[1 + \frac{p_0}{1 - p_0}\right] \\ &\quad + (\alpha x)^2 \left\{ \frac{1}{2} \log(p_0) + \frac{1}{2} \log(p_0) \times \frac{p_0}{1 - p_0} + \frac{1}{2} [\log(p_0)]^2 \right. \\ &\quad \left. \times \frac{p_0}{1 - p_0} + \frac{1}{2} [\log(p_0)]^2 \times \left(\frac{p_0}{1 - p_0}\right)^2 \right\} \\ &= \log\left(\frac{p_0}{1 - p_0}\right) - (\alpha x) \frac{\log(p_0)}{1 - p_0} + \frac{(\alpha x)^2}{2} \frac{\log(p_0)}{1 - p_0} \left\{1 + \frac{p_0 \log(p_0)}{1 - p_0}\right\}. \end{aligned}$$

Fig. A1. The true underlying model under the additive Gompertz simulation scenario.

Table A1. Simulation results under the additive Gompertz model for a total sample size of 600 individuals

	Cohort ($N = 600$)					Case-control ($N = 600$)				
	N_D	α_2								
	133.5	149.1	164.5	178.8	190.4	300	300	300	300	300
	0.00	0.50	1.00	1.50	2.00	0.00	0.50	1.00	1.50	2.00
<i>Relative efficiency (MSE)</i>										
M1/M2	2.50	2.22	1.60	1.22	1.05	3.91	2.83	1.78	1.28	1.08
M1/M3	0.94	0.98	1.14	1.43	1.89	0.93	1.00	1.20	1.54	2.03
M1/M4	2.10	2.26	2.40	2.54	2.76	3.21	3.12	3.18	3.18	3.28
<i>AUC</i>										
M1	0.78	0.76	0.74	0.72	0.71	0.77	0.75	0.74	0.72	0.71
M2	0.78	0.75	0.73	0.72	0.71	0.77	0.75	0.73	0.72	0.71
M3	0.78	0.76	0.73	0.72	0.71	0.78	0.76	0.73	0.72	0.71
M4	0.78	0.76	0.74	0.72	0.71	0.78	0.76	0.74	0.72	0.71
<i>AIC</i>										
M1	538	584	626	660	683	677	702	725	741	750
M2	526	573	619	656	682	658	687	715	736	748
M3	539	584	623	652	670	678	702	720	730	733
M4	527	573	614	645	665	659	685	708	721	727
<i>Reject $\beta_3 = 0$ (proportion)¹</i>										
M1										
M2	0.95	0.93	0.83	0.60	0.37	1.00	0.99	0.91	0.68	0.41
M3										
M4	0.96	0.95	0.91	0.80	0.68	1.00	0.99	0.97	0.89	0.77
<i>Reject $\beta_4 = 0$ (proportion)¹</i>										
M1										
M2										
M3	0.03	0.14	0.53	0.87	0.98	0.04	0.20	0.66	0.94	0.99
M4	0.05	0.24	0.69	0.94	0.99	0.04	0.35	0.85	0.99	1.00

All results are the average over 5,000 replicates. In all simulations the following parameter values were used: $\mu_0 = -2$, $\alpha_1 = 0.15$, and $c_m = 0.5$. Models M1, M2, M3, and M4 are described in the Simulations section. ¹ Based on a 2-sided test at $\alpha = 0.05$.

Fig. A2. Smoothed kernel estimates of early death for colorectal and pancreatic cancer cases between 40 and 90 years old at diagnosis. Kernel estimates were separately fitted for males and females. **a** Colorectal cancer, White cohort; **b** colorectal cancer, Black cohort; **c** pancreatic cancer, White cohort; **d** pancreatic cancer, Black cohort.

Table A2. Early death among adults with colorectal and pancreatic cancer between 40 and 90 years old at diagnosis under the logistic model for disease risk

	Colorectal cancer				Pancreatic cancer			
	Black race		White race		Black race		White race	
	odds ratio (95% CI)	<i>p</i> value	odds ratio (95% CI)	<i>p</i> value	odds ratio (95% CI)	<i>p</i> value	odds ratio (95% CI)	<i>p</i> value
<i>Model M1</i>								
Age	1.61 (1.53–1.7)	<0.001	1.76 (1.72–1.81)	<0.001	1.51 (1.42–1.61)	<0.001	1.57 (1.53–1.61)	<0.001
Sex	1.12 (0.99–1.27)	0.082	1.02 (0.96–1.08)	0.509	1.45 (1.26–1.67)	<0.001	1.16 (1.09–1.23)	<0.001
<i>Model M2</i>								
Age	1.19 (0.96–1.48)	0.124	1.26 (1.13–1.4)	<0.001	1.30 (1.00–1.68)	0.047	1 (0.89–1.11)	0.956
Sex	1.13 (1–1.28)	0.057	1.04 (0.98–1.1)	0.21	1.46 (1.27–1.69)	<0.001	1.17 (1.11–1.24)	<0.001
Age ²	1.06 (1.02–1.1)	0.005	1.06 (1.04–1.08)	<0.001	1.03 (0.98–1.08)	0.235	1.09 (1.07–1.11)	<0.001
<i>Model M3</i>								
Age	1.67 (1.55–1.8)	<0.001	1.89 (1.83–1.96)	<0.001	1.59 (1.46–1.73)	<0.001	1.62 (1.56–1.68)	<0.001
Sex	1.36 (1–1.86)	0.049	1.57 (1.33–1.84)	<0.001	1.92 (1.36–2.72)	<0.001	1.38 (1.18–1.61)	<0.001
Age:Sex	0.93 (0.83–1.03)	0.168	0.87 (0.83–0.91)	<0.001	0.89 (0.78–1.01)	0.081	0.94 (0.89–0.99)	0.018
<i>Model M4</i>								
Age	1.24 (0.98–1.56)	0.073	1.38 (1.23–1.55)	<0.001	1.41 (1.07–1.87)	0.015	1.02 (0.91–1.15)	0.728
Sex	1.32 (0.97–1.79)	0.076	1.49 (1.27–1.74)	<0.001	1.88 (1.32–2.67)	<0.001	1.27 (1.09–1.48)	0.003
Age ²	1.06 (1.01–1.1)	0.008	1.06 (1.04–1.08)	<0.001	1.02 (0.97–1.08)	0.4	1.09 (1.06–1.11)	<0.001
Age:Sex	0.94 (0.85–1.05)	0.282	0.89 (0.85–0.93)	<0.001	0.9 (0.79–1.03)	0.126	0.97 (0.92–1.02)	0.284

Age was recorded in 10-year increments. Models M1, M2, M3, and M4 are defined in the Simulations section.

Table A3. Two case-control data sets evaluating NAT2 acetylation and smoking status in bladder cancer and advanced colorectal adenoma patients

	Bladder cancer		Advanced colorectal adenoma	
	odds ratio (95% CI)	<i>p</i> value	odds ratio (95% CI)	<i>p</i> value
<i>Model M1</i>				
NAT2 acetylation		<0.001		0.062
Slow	(reference)		(reference)	
Rapid/intermediate	0.86 (0.81–0.9)		0.81 (0.65–1.01)	
Smoking status		<0.001		<0.001
Never	(reference)		(reference)	
Past	2.05 (1.92–2.19)		1.1 (0.86–1.4)	
Current	2.48 (2.31–2.67)		2.27 (1.67–3.08)	
<i>Model M3</i>				
NAT2 acetylation		0.497		0.324
Slow	(reference)		(reference)	
Rapid/intermediate	1.04 (0.93–1.16)		0.84 (0.59–1.19)	
Smoking status		<0.001		<0.001
Never	(reference)		(reference)	
Past	2.23 (2.05–2.42)		0.99 (0.71–1.37)	
Current	2.78 (2.54–3.04)		3.09 (2.04–4.74)	
Interaction		<0.001		0.014
Never smoker × r/i acetylation	(reference)		(reference)	
Past smoker × r/i acetylation	0.80 (0.7–0.92)		1.27 (0.77–2.09)	
Current smoker × r/i acetylation	0.74 (0.64–0.86)		0.5 (0.27–0.94)	

r/i, rapid/intermediate.

References

- 1 Agresti A: Categorical Data Analysis, ed 2. Hoboken, Wiley, 2002.
- 2 Cordell HJ: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;10:392–404.
- 3 Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003;56:73–82.
- 4 Wang X, Elston RC, Zhu X: The meaning of interaction. *Hum Hered* 2010;70:269–277.
- 5 Wang X, Elston RC, Zhu X: Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nat Rev Genet* 2011;12:74.
- 6 Wright S: Book review of *The Biology of Population Growth* by Raymond Pearl and *The Natural Increase of Mankind* by J. Shirley Sweeney. *J Am Stat Assoc* 1926;21:493–512.
- 7 Berger RD: Comparison of the Gompertz and logistic equations to describe plant disease progress. *Phytopathology* 1981;71:716–719.
- 8 Norton L, Simon R, Brereton HD, Bogden AE: Predicting the course of Gompertzian growth. *Nature* 1976;264:542–544.
- 9 Panik MJ: Growth Curve Modeling: Theory and Applications. Hoboken, Wiley, 2014.
- 10 Stukel TA: Generalized logistic models. *J Am Stat Assoc* 1988;83:426–431.
- 11 Winsor CP: The Gompertz curve as a growth curve. *Proc Natl Acad Sci* 1932;18:1–8.
- 12 Guerrero V, Johnson R: Use of Box-Cox transformation with binary response models. *Biometrika* 1982;69:309–314.
- 13 Pregibon D: Goodness of link tests for generalized linear models. *J R Stat Soc Ser C Appl Stat* 1980;29:15–23.
- 14 Satagopan JM, Elston RC: Evaluation of removable statistical interaction for binary traits. *Stat Med* 2013;32:1164–1190.
- 15 Surveillance Research Program, National Cancer Institute, “SEER*Stat software.” www.seer.cancer.gov/seerstat, 2015.
- 16 Moslehi R, Chatterjee N, Church TR, et al: Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics* 2006;7:819–829.
- 17 Rothman N, Garcia-Closas M, Chatterjee N, et al: A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet* 2010;42:978–984.
- 18 Greene FL, Balch CM, Haller DG, Morrow M: AJCC Cancer Staging Manual, ed 6. Chicago, American Joint Committee on Cancer, 2002.
- 19 Surveillance Epidemiology and End Results (SEER) Program, SEER*Stat Database: Incidence – SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases Nov (Sub) (1973–2012) Total U.S., 1969–2013 Counties. www.seer.cancer.gov, November 2014.
- 20 Gail MH, Pfeier RM: On criteria for evaluating models of absolute risk. *Biostatistics* 2005;6:227–239.