epigenomes

Article

Statistical Models for High-Risk Intestinal Metaplasia with DNA Methylation Profiling

Tianmeng Wang , Yifei Huang and Jie Yang *

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA; twang218@uic.edu (T.W.); yhuan39@uic.edu (Y.H.)
* Correspondence: jyang06@uic.edu

Abstract: We consider the newly developed multinomial mixed-link models for a high-risk intestinal metaplasia (IM) study with DNA methylation data. Different from the traditional multinomial logistic models commonly used for categorical responses, the mixed-link models allow us to select the most appropriate link function for each category. We show that the selected multinomial mixed-link model (Model 1) using the total number of stem cell divisions (TNSC) based on DNA methylation data outperforms the traditional logistic models in terms of cross-entropy loss from ten-fold cross-validations with significant p-values 8.12×10^{-4} and 6.94×10^{-5}. Based on our selected model, the significance of TNSC's effect in predicting the risk of IM is justified with a p-value less than 10^{-6}. We also select the most appropriate mixed-link models (Models 2 and 3) when an additional covariate, the status of gastric atrophy, is available. When the status is negative, mild, or moderate, we recommend Model 2; otherwise, we prefer Model 3. Both Models 2 and 3 can predict the risk of IM significantly better than Model 1, which justifies that the status of gastric atrophy is informative in predicting the risk of IM.

Keywords: AIC; BIC; categorical response; cross-entropy loss; cross-validation; multinomial logistic model; multinomial mixed-link model

check for updates

Citation: Wang, T.; Huang, Y.; Yang, J. Statistical Models for High-Risk Intestinal Metaplasia with DNA Methylation Profiling. *Epigenomes* **2024**, *8*, 19. https://doi.org/10.3390/epigenomes8020019

Academic Editors: Ivana De la Serna and Che-Kun James Shen

Received: 26 March 2024
Revised: 28 April 2024
Accepted: 9 May 2024
Published: 11 May 2024

1. Introduction

Gastric intestinal metaplasia (IM) is a precancerous change in the mucosa of the stomach with intestinal epithelium [1], which increases the risk of gastric cancer [2], the third leading cause of cancer death worldwide and the fifth most common malignancy in the world [3]. Intestinal-type gastric cancer is more common and is associated with chronic inflammation, atrophy, and IM of the stomach, often relevant to *Helicobacter pylori* infection [4]. The exact mechanism of how IM leads to gastric cancer is not fully understood, but it may involve genetic and epigenetic alterations that affect the expression and function of key genes, including DNA methylations [5]. There has been increasing evidence that DNA methylation changes in normal tissue are correlated with cancer risk [6–12], including gastric cancer [5,13]. The DNA methylation levels observed in IM tissue samples are significantly higher than normal gastric samples, which indicates that the DNA methylation profiles may help with predicting IM and gastric cancer [5].

In this study, we utilize the DNA methylation data of 124 samples obtained from the Gastric Cancer Epidemiology Program (GCEP) and deposited in NCBI (GSE103186) by [5]. We aim to build the most appropriate statistical model to predict the risk level of IM, including Normal (normal gastric samples), MIM (mild IM or low-risk samples, type I), and IM (high-risk samples, type II or type III), using the total number of stem cell divisions per stem cell (TNSC) estimated by the epiTOC2 (Epigenetic Timer of Cancer-2, [12]) model from the measured DNA methylation profile, along with other clinical information such as the status of gastric atrophy [5].

For categorical responses with three or more categories, such as {Normal, MIM, IM} in this study, multinomial logistic models have been widely used in the statistical literature, including the baseline-category, cumulative, adjacent-categories, and continuation-ratio logit models [14–17]. Among the four classes of logit models, the baseline-category logit model, also known as the (multiclass) logistic regression model, has been extended with a probit link and is known as the multinomial probit model [18–20]; the cumulative logit model has been extended to cumulative link models [19,21,22]; and the continuation-ratio logit model has been extended with a complementary log-log link [23]. It should be noted that all these models assume the same link function for all categories.

In this study, we adopt the multinomial mixed-link model (see Section 2.2), proposed by [24] recently, because it not only covers all the models mentioned above but also allows us to choose different link functions across categories. By choosing the multinomial mixed-link model, we find out that the cumulative mixed-link model with proportional odds (po) assumption and g_1 = loglog, g_2 = logit link functions outperforms the traditional models, in terms of predicting the risk level of IM using DNA methylation profiles (see Section 3.1). Based on ten-fold cross-validations, the improvement is statistically significant. Our results also show that by incorporating the status of gastric atrophy can further improve the prediction accuracy significantly. Having run our model selection procedure again, we determine that an adjacent-categories logit model with po (see Section 3.2) is most appropriate when the status of gastric atrophy is negative, mild or moderate, whereas an adjacent-categories probit model with po (see Section 3.3) works the best when the status is marked or unknown. For readers' reference, we provide the predictive probabilities for each tissue sample in the Supplementary Materials, as well as the sample IDs and the corresponding covariates.

2. Materials and Methods

2.1. epiTOC2 Model and TNSC Covariate

The mitotic age of tissues is relevant to the total number of cell divisions, which can be estimated by the DNA methylation changes in the stem cell. Recent studies have shown the correlation between the mitotic age of tissue and the neoplastic transformation [25–27]. Many models for estimating mitotic age have been proposed based on DNA methylation data, including the epiTOC model [28], the solo-WCGWs model [29], and the epiTOC2 model [12]. In this study, we adopt the epiTOC2 model, which shows good robustness and is better for discriminating preneoplastic lesions [12]. The epiTOC2 model estimates the total number of stem cell divisions directly (TNSC) and is based on CpG sites marked by polycomb repressive complex-2 (PRC2). These sites are generally unmethylated across fetal tissues and become methylated during ontogeny and aging. The epiTOC2 model was fitted using the Illumina Infinium 450k data from [30], who selected $n_c = 163$ CpG sites in their model based on the rate of increase in DNA methylation rates. A simplified epiTOC2 model can be rewritten as a weighted average of DNA methylation beta values over the n_c CpGs in a sample s as follows:

$$\mathrm{TNSC}(s) = \frac{1}{n_c}\sum_{i=1}^{n_c} w_i\beta_{is} = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{2\beta_{is}}{\delta_i}$$

where δ_i is a model parameter representing the probability of de novo methylation of parent and daughter strands (see [12] for more details).

In this study, we first use TNSC as the only covariate representing the DNA methylation profile to predict the risk level of IM (see Section 3.1).

2.2. Multinomial Mixed-Link Models

In general, we consider d covariates or predictors with m distinct settings $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$, for $i = 1, \ldots, m$. At the ith setting, n_i categorical responses are collected and summarized into a multinomial response $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iJ})^T \sim \text{Multinomial}(n_i; \pi_{i1}, \cdots, \pi_{iJ})$, where Y_{ij} is the number of observations with the jth response category, and π_{ij} is the probability that the response falls into the jth category, $j = 1, \ldots, J$. Assuming all

$\pi_{ij} \in (0,1)$, there are four classes of multinomial logit models that have ever been used in the literature (see [16] and the references therein):

$$\begin{aligned}
\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) &= \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i, \text{ baseline-category} && (1)\\
\log\left(\frac{\pi_{i1} + \cdots + \pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}}\right) &= \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i \text{ , cumulative} && (2)\\
\log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) &= \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i \text{ , adjacent-categories} && (3)\\
\log\left(\frac{\pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}}\right) &= \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i, \text{ continuation-ratio} && (4)
\end{aligned}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^T$, $i = 1, \ldots, m$, and $j = 1, \ldots, J-1$. In the statistical literature (see, for example, [16]), the four logit models, (1)–(4), are also called nonproportional odds (npo) models, which allow $\boldsymbol{\beta}_j$'s to be different across $j = 1, \ldots, J-1$. If we further assume $\boldsymbol{\beta}_j \equiv \boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$, then the four models are known as proportional odds (po) models. For more general odds structures for multinomial logistic models, that is, partial proportional odds (ppo) models, please see [16,17].

In the form of npo models, the multinomial mixed-link model [24] can be written as follows

$$g_j(\rho_{ij}) = \beta_{0j} + \boldsymbol{\beta}_j^T \boldsymbol{x}_i \tag{5}$$

where

$$\rho_{ij} = \begin{cases} \frac{\pi_{ij}}{\pi_{ij}+\pi_{iJ}} & \text{, for baseline-category mixed-link models} \\ \pi_{i1} + \cdots + \pi_{ij} & \text{, for cumulative mixed-link models} \\ \frac{\pi_{ij}}{\pi_{ij}+\pi_{i,j+1}} & \text{, for adjacent-categories mixed-link models} \\ \frac{\pi_{ij}}{\pi_{ij}+\cdots+\pi_{iJ}} & \text{, for continuation-ratio mixed-link models} \end{cases} \tag{6}$$

where g_j is a predetermined link function, $i = 1, \ldots, m$, and $j = 1, \ldots, J-1$. It can be verified that if $g_1(\rho_{ij}) \equiv \cdots \equiv g_{J-1}(\rho_{ij}) = \log(\rho_{ij}/(1-\rho_{ij}))$, that is, the logit link, then the multinomial mixed-link model (5) plus (6) leads to the four multinomial logit models (1)–(4). In this study, we also consider some other link functions that have been commonly used in the literature, namely, probit ($g_j(\rho_{ij}) = \Phi^{-1}(\rho_{ij})$, where Φ is the cumulative distribution function of standard normal distribution), log-log (or loglog, $g_j(\rho_{ij}) = -\log(-\log(\rho_{ij}))$, and complementary log-log (or cloglog, $g_j(\rho_{ij}) = \log(-\log(1-\rho_{ij}))$. For more options of link functions, please see Table 1 in [24].

Following the notation in [24], the multinomial mixed-link model (5) plus (6) can be written into its matrix form:

$$\mathbf{g}\left(\frac{\mathbf{L}\boldsymbol{\pi}_i}{\mathbf{R}\boldsymbol{\pi}_i + \pi_{iJ}\mathbf{b}}\right) = \boldsymbol{\beta}_0 + \mathbf{B}^T \mathbf{x}_i \tag{7}$$

where $\mathbf{g} = (g_1, \ldots, g_{J-1})^T$, $\mathbf{L}$ and $\mathbf{R}$ are $(J-1) \times (J-1)$ constant matrices, $\mathbf{b}$ is a constant vector of length $J-1$, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{i,J-1})^T$, $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0,J-1})^T$, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1})$ is a $d \times (J-1)$ matrix of parameters. Note that the vector $\mathbf{g}$ of link functions in (7) applies to the ratio of two vectors component-wise. That is, if we denote $\mathbf{L} = (\mathbf{L}_1, \ldots, \mathbf{L}_{J-1})^T$, $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_{J-1})^T$ and $\mathbf{b} = (b_1, \ldots, b_{J-1})^T$, then the multinomial mixed-link model (7) can be written in its equation form:

$$g_j\left(\frac{\mathbf{L}_j^T \boldsymbol{\pi}_i}{\mathbf{R}_j^T \boldsymbol{\pi}_i + \pi_{iJ} b_j}\right) = \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i \text{ ,} \quad j = 1, \ldots, J-1$$

In other words, ρ_{ij} in (5) and (6) can be written as

$$\rho_{ij} = \frac{\mathbf{L}_j^T \boldsymbol{\pi}_i}{\mathbf{R}_j^T \boldsymbol{\pi}_i + \pi_{iJ} b_j}, \quad j = 1, \ldots, J-1$$

In this study, we consider the four classes of mixed-link models listed in (6). For baseline-category mixed-link models, $\mathbf{L} = \mathbf{R} = I_{J-1}$, the identity matrix of order $J-1$, and $\mathbf{b} = \mathbf{1}_{J-1}$, the vector of ones with length $J-1$; for cumulative mixed-link models,

$$\mathbf{L} = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{(J-1)\times(J-1)}$$

$\mathbf{R} = \mathbf{1}_{J-1}\mathbf{1}_{J-1}^T$, and $\mathbf{b} = \mathbf{1}_{J-1}$; for adjacent-categories mixed-link models, $\mathbf{L} = I_{J-1}$,

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & 1 & \ddots & \\ & & & \ddots & 1 \\ & & & & 1 \end{bmatrix} \in \mathbb{R}^{(J-1)\times(J-1)}, \text{ and } \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{J-1}$$

and for continuation-ratio mixed-link models, $\mathbf{L} = I_{J-1}$,

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ & 1 & \cdots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} \in \mathbb{R}^{(J-1)\times(J-1)}$$

and $\mathbf{b} = \mathbf{1}_{J-1}$.

In this study, we implement the algorithms described in Section 4 of [24] to find the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}$ for either the npo model's parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$ of length $p = (d+1) \times (J-1)$, or the po model's $\boldsymbol{\theta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}^T)^T$ of length $p = d + J - 1$.

2.3. Model Selection and Evaluation

In this study, we use the multinomial mixed-link model (5) plus (6) to predict the risk level of IM in three ordered categories, namely, Normal, MIM, and IM. In terms of the structure of ρ_{ij} as defined in (6), we have four options, namely, baseline-category, cumulative, adjacent-categories, and continuation-ratio mixed-link models. In this study, the number of response categories is $J = 3$. For each $j = 1, \ldots, J-1$, we consider four possible link functions, namely, logit, probit, loglog, and cloglog. From the right-hand side of (5), we still have two options, an npo model ($\beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i$) or a po model ($\beta_{0j} + \boldsymbol{\beta}^T \mathbf{x}_i$). As a summary, we have $4 \times 4^{J-1} \times 2$ candidate models.

In the statistical literature, the Akaike Information Criterion (AIC, [31,32]) and Bayesian Information Criterion (BIC, [33]) have been widely used for model selection, given that a statistical model is assumed. In our case, the maximized likelihood $l(\hat{\boldsymbol{\theta}})$ is obtained along with the MLE $\hat{\boldsymbol{\theta}}$ after fitting the model. In our notation,

$$\begin{aligned} \text{AIC} &= -2 \cdot l(\hat{\boldsymbol{\theta}}) + 2 \cdot\ p \\ \text{BIC} &= -2 \cdot l(\hat{\boldsymbol{\theta}}) + \log(n) \cdot p \end{aligned}$$

where $n = \sum_{i=1}^{m} n_i$ stands for the total number of observations or the sample size, $p = (d+1) \times (J-1)$ for npo models or $d + J - 1$ for po models in our study. Smaller AIC or BIC values imply better models. Since in this study the sample size $n = 124$ (see Section 3) is not large, we recommend AIC against BIC if their results of model selection are not consistent (see, for example, [34], for more discussions on AIC and BIC).

To show if the selected model is significantly better than commonly used models in the literature, we use a ten-fold cross-validation to estimate the prediction errors of the models under comparison. Different from five-fold cross-validations chosen by [17], we choose ten-fold cross-validations in this study because our sample size $n = 124$ is relatively smaller (for more discussion on ten-fold versus five-fold cross-validations, see [34]).

Different from many machine learning techniques, the multinomial mixed-link model provides a stochastic classification answer [35] to each tissue sample. That is, given the covariate or predictor setting $\mathbf{x}_i$, we obtain by the fitted multinomial mixed-link model predictive probabilities $\hat{\pi}_{ij}$ for Normal ($j = 1$), MIM ($j = 2$), and IM ($j = 3$), respectively, which is much more informative than a deterministic classification answer [35]. Following [17], we use the cross-entropy loss to evaluate the performance of statistical models under comparison. Given a random partition B of the index set $[n] = \{1, \ldots, n\}$, which divides $[n]$ into ten non-overlapped subsets (called blocks) of roughly the same size, the (average) cross-entropy (CE) loss for a specified model is

$$\mathrm{CE}(B) = -\frac{1}{n} \sum_{i=1}^{n} \log\left(\hat{\pi}_{i,y_i}^{k(i)}\right)$$

where $n = 124$ is the sample size, y_i is the observed response label of the ith tissue sample, and $k(i)$ is the block label to which the ith sample belongs. More details about calculating CE can be found in Section 2.4 of [17] except that we use a ten-fold instead of five-fold cross-validation.

A smaller CE value implies a better model. To check whether the improvement of one model against another is statistically significant, in this study we randomly generate partitions and use a one-sided paired t-test to check whether the improvement is significant.

3. Results

3.1. Statistical Model Selection for Predicting IM Based on TNSC

In this study, we first match the DNA methylation data downloaded from NCBI (https://www.ncbi.nlm.nih.gov/geo/, GSE103186, accessed on 23 January 2024) with the tissue samples listed in Table S3 in [5] (https://www.cell.com/cancer-cell/, accessed on 18 January 2024). Among the 134 tissue samples collected at the antrum site [5], there are 10 samples lacking DNA methylation profiles. We use the remaining 124 samples for our analysis. We then compute the TNSC values for the 124 samples using their DNA methylation data, as described in Section 2.1. The R codes for computing TNSC are accessible online (https://zenodo.org/records/2632938, epiTOC2.R, accessed on 15 January 2024) as indicated by [12]. In this section, we consider the multinomial mixed-link model as described in Section 2.2, and use the computed TNSC as the only covariate to predict the risk level of IM in three categories (Normal, MIM, and IM). For each of 4 × 2 models, the optimal link functions for $j = 1, 2$, respectively, along with their corresponding AIC and BIC values, are listed in Table 1 (see Appendix A for the AIC and BIC values of all link combinations).

Table 1. Best mixed-link models for predicting IM based on TNSC.

Model	Best Link	AIC	BIC
Baseline-category npo	loglog, loglog	146.69	157.97
Cumulative npo	loglog, probit	145.53	156.81
Adjacent-categories npo	loglog, loglog	145.97	157.25
Continuation-ratio npo	loglog, loglog	145.93	157.21

Table 1. *Cont.*

Model	Best Link	AIC	BIC
Baseline-category po	probit, logit	151.12	159.58
Cumulative po	**loglog, logit**	**144.29**	**152.75**
Adjacent-categories po	loglog, logit	146.33	154.79
Continuation-ratio po	loglog, logit	146.96	155.42

Note: The best model overall, along with its links and values, is highlighted in bold.

According to Table 1, the best multinomial mixed-link model with the lowest AIC overall in this case, called Model 1, is a cumulative po model with loglog and logit links for $j = 1$ (Normal) and $j = 2$ (MIM), respectively. Note that by default $j = 3$ (IM) is treated as the baseline category. The fitted Model 1 is provided in (8), where $x_{TNSC,i}$ is the computed TNSC value for the ith tissue sample.

$$\begin{aligned} -\log(-\log(\pi_{i1})) &= \beta_{01} + \beta_1 x_{TNSC,i} = 4.023 - 4.228 \times 10^{-4} x_{TNSC,i} \\ \log\left(\frac{\pi_{i1} + \pi_{i2}}{\pi_{i3}}\right) &= \beta_{02} + \beta_1 x_{TNSC,i} = 4.905 - 4.228 \times 10^{-4} x_{TNSC,i} \end{aligned} \tag{8}$$

In (8), the estimated coefficient of $x_{TNSC,i}$ is -4.228×10^{-4}, which is fairly small. To test whether the effect of TNSC is significant in predicting IM, we obtain its 95% confidence interval $(-4.167 \times 10^{-4}, -4.290 \times 10^{-4})$, which does not contain zero. Actually, the corresponding p-value of its significance test is less than 10^{-6}. As a conclusion, the effect of TNSC is statistically significant in predicting the risk level of IM.

To further check whether Model 1 outperforms the traditional statistical models, as described in Section 2.3, we run a ten-fold cross-validation and compare its cross-entropy loss against other models. For illustration purposes, we choose the baseline-category logit model with npo (also known as the multiclass logistic regression model) and the cumulative logit model with npo (one of the most popular models for ordinal responses) as the alternative models. As for other models, including multinomial logit models and probit models, the conclusions are similar (see Appendix A). To avoid misleading conclusions relying on a particular partition, we randomly generate ten partitions and compute their corresponding CE values. The boxplots of the resulting ten CE values are provided in Figure 1, which shows that the CE values of Model 1 seem to be much lower than those values of the other two models. Although we only run ten random partitions due to computational intensity, our one-sided paired t-tests based on the ten CE values show that the improvements of Model 1 are significant. The p-values of the t-tests for comparing Model 1 against the baseline npo model and the cumulative npo model displayed in Figure 1 are 8.12×10^{-4} and 6.94×10^{-5}, respectively. That is, the recommended cumulative po model with loglog and logit links significantly outperforms the two multinomial logistic models that are commonly used in practice.

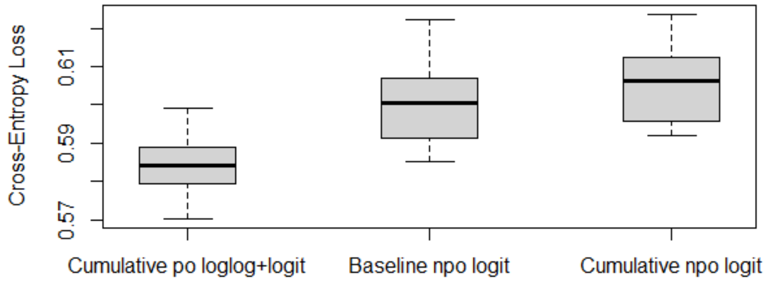

Figure 1. Cross-entropy loss based on ten-fold cross-validations with ten random partitions.

To show how well Model 1 works, we plot in Figure 2 the predictive probabilities $\hat{\pi}_{ij}$ against the true response labels, $j = 1, 2, 3$, respectively.

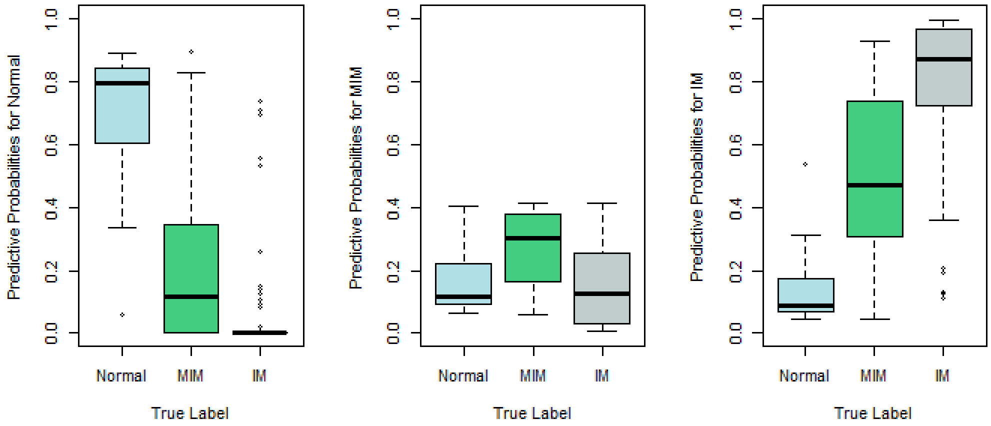

Figure 2. Predictive probabilities $\hat{\pi}_{ij}$ based on Model 1 against true response labels (left panel: $j = 1$; middle panel: $j = 2$; right panel: $j = 3$).

According to Figure 2, the recommended Model 1 works reasonably well. For examples, in the left panel, we plot $\hat{\pi}_{i1}$, which is the predictive probability that the ith tissue sample belongs to Normal, against its true response label. If the true label is Normal, the left boxplot in the left panel of Figure 2, which is apparently higher than the other two boxplots in the same panel, indicates that the corresponding tissue sample tends to be predicted as Normal as well. Similarly, in the right panel, $\hat{\pi}_{i3}$, the predictive probability that the sample belongs to IM, is plotted, and the significantly higher boxplot to the right indicates that the sample with true label IM tends to be predicted as IM as well. Nevertheless, the middle panel, which plots the predictive probabilities for MIM, indicates that the MIM class is not so different from Normal or IM, and thus is more difficult to predict correctly.

3.2. Statistical Model Selection for Predicting IM Based on TNSC and Gastric Atrophy

In this section, we show that when additional information, such as the status of gastric atrophy, is available, the prediction accuracy of the IM risk level can be significantly improved.

In this study, the status of gastric atrophy is a 5-class categorical variable (see Table S3 in [5]), namely, Marked, Moderate, Mild, Negative, and Unknown. In our regression analysis involving the status of gastric atrophy, we replace it with four dummy variables: $x_{mild,i}$, $x_{moderate,i}$, $x_{negative,i}$, and $x_{unknown,i}$. Each dummy variable is binary, taking a value of either 1 or 0, with at most one variable allowed to be 1 for any given sample. For instance, a configuration of $(x_{mild,i}, x_{moderate,i}, x_{negative,i}, x_{unknown,i}) = (1, 0, 0, 0)$ indicates a mild gastric atrophy status for the ith sample, $(0, 1, 0, 0)$ indicates a moderate gastric atrophy status, whereas $(0, 0, 0, 0)$ indicates a marked status, that is, the baseline status. Similarly to Table 1, we list the optimal link functions for $j = 1, 2$, respectively, along with their AIC and BIC values, in Table 2.

Table 2. Best mixed-link models for predicting IM based on TNSC and gastric atrophy.

Model	Best Link	AIC	BIC
Baseline-category npo	logit, probit	109.95	143.79
Cumulative npo	loglog, logit	109.20	143.04
Adjacent-categories npo	logit, logit	109.97	143.81
Continuation-ratio npo	logit, logit	110.97	144.82
Baseline-category po	probit, logit	111.31	131.05
Cumulative po	probit, probit	110.03	129.77
Adjacent-categories po	**logit, logit**	**108.89**	**128.63**
Continuation-ratio po	probit, probit	109.32	129.06

Note: The best model overall, along with its links and values, is highlighted in bold.

With the presence of gastric atrophy, the best multinomial mixed-link model, called Model 2, is an adjacent-categories logit model with po, which is different from the type of Model 1 with TNSC only (see Section 3.1). Since its AIC value, 108.89, is much less than 144.29 in Table 1, Model 2 is expected to outperform Model 1 significantly in terms of prediction accuracy (see [36] for more discussion on AIC differences). The fitted Model 2 is provided in (9).

$$\begin{aligned}
\log\left(\frac{\pi_{i1}}{\pi_{i2}}\right) =& \beta_{01} + \beta_1 x_{TNSC,i} + \beta_2 x_{mild,i} + \beta_3 x_{moderate,i} + \beta_4 x_{negative,i} + \beta_5 x_{unknown,i} \\
=& -1.859 - 4.586 \times 10^{-4} x_{TNSC,i} - 1.144 x_{mild,i} - 2.103 x_{moderate,i} \\
& + 6.469 x_{negative,i} + 3.663 x_{unknown,i} \\
\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) =& \beta_{02} + \beta_1 x_{TNSC,i} + \beta_2 x_{mild,i} + \beta_3 x_{moderate,i} + \beta_4 x_{negative,i} + \beta_5 x_{unknown,i} \\
=& 0.136 - 4.586 \times 10^{-4} x_{TNSC,i} - 1.144 x_{mild,i} - 2.103 x_{moderate,i} \\
& + 6.469 x_{negative,i} + 3.663 x_{unknown,i}
\end{aligned} \tag{9}$$

Similarly to Figure 1, we compare in Figure 3 the cross-entropy loss of two recommended models shown in (8) (Model 1) and (9) (Model 2). It is not surprising that Model 2 with both TNSC and gastric atrophy as predictors has a significantly smaller cross-entropy loss, which implies that the status of gastric atrophy is informative in predicting the risk level of IM.

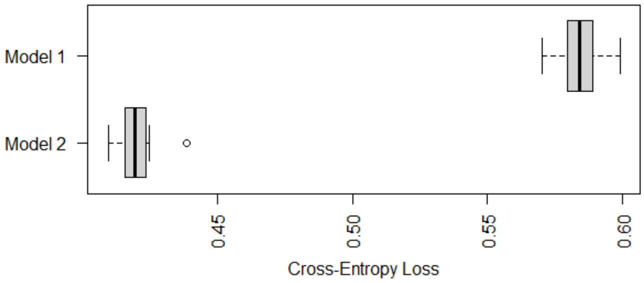

Figure 3. Boxplots of cross-entropy loss of Model 1 and Model 2 based on ten-fold cross-validations with ten random partitions.

Similarly to Figure 2, we plot the predictive probabilities based on Model 1 and Model 2 against the true IM labels in Figures 4–6. When the true IM label matches the predictive label, such as the left panel in Figure 4, the middle panel in Figure 5, and the right panel of Figure 6, Model 2 tends to provide a higher predictive probability than Model 1, which shows that overall Model 2 outperforms Model 1.

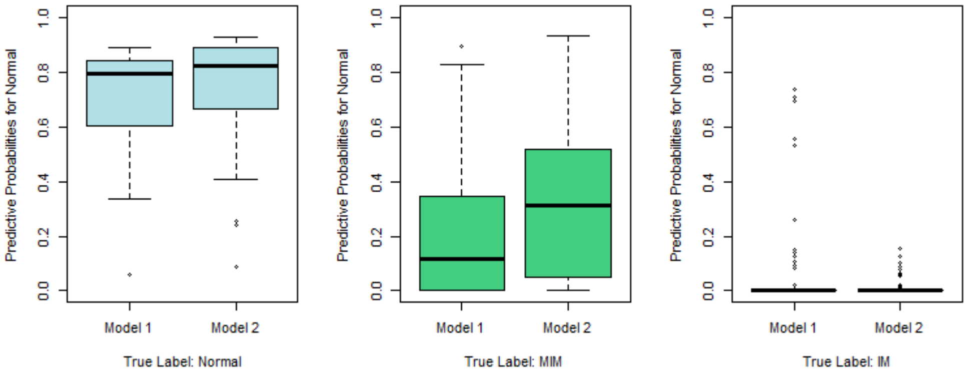

Figure 4. Predictive probabilities for the normal category based on Model 1 and Model 2.

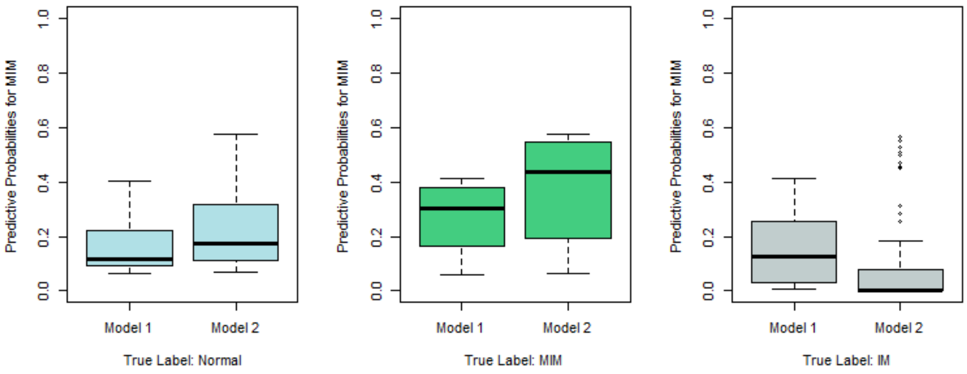

Figure 5. Predictive probabilities for the MIM category based on Model 1 and Model 2.

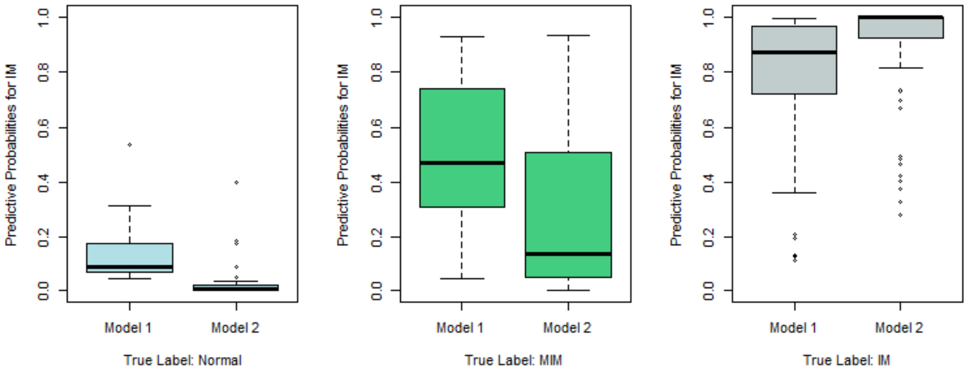

Figure 6. Predictive probabilities for IM category based on Model 1 and Model 2.

3.3. Statistical Model Selection after Removing Unknown and Marked Categories

Among the 124 samples considered in this study, there are only 3 cases with "Marked" status of gastric atrophy, and there are 23 cases with "Unknown" status, which is not informative. In this section, we consider the best multinomial mixed-link model for the 98 cases after removing the samples that belong to Marked or Unknown categories.

In this section, the status of gastric atrophy is a three-class categorical variable restricted to the 98 samples. Similarly to Model 2 in Section 3.2, we replace the status of gastric atrophy with two dummy variables ($x_{mild,i}$, $x_{moderate,i}$). More specifically, ($x_{mild,i}$, $x_{moderate,i}$) = (1,0) stands for mild status, (0,1) for moderate status, and (0,0) for negative status representing the baseline. Similarly to Tables 1 and 2, we provide in Table 3 the optimal choices of link functions for each type of multinomial model. According to Table 3, the best multinomial mixed-link model for this scenario is an adjacent-categories po model with probit links for both $j = 1, 2$. We call it Model 3 and list its fitted model in (10).

Table 3. Best mixed-link models for predicting IM based on TNSC and 3-class gastric atrophy.

Model	Best Link	AIC	BIC
Baseline-category npo	logit, probit	81.43	102.11
Cumulative npo	probit, probit	84.29	104.97
Adjacent-categories npo	logit, probit	83.22	103.90
Continuation-ratio npo	logit, probit	83.56	104.24
Baseline-category po	probit, logit	82.39	95.32
Cumulative po	probit, probit	77.99	90.92
Adjacent-categories po	**probit, probit**	**77.56**	**90.48**
Continuation-ratio po	probit, probit	77.77	90.69

Note: The best model overall, along with its links and values, is highlighted in bold.

$$\begin{aligned}\Phi^{-1}\left(\frac{\pi_{i1}}{\pi_{i1}+\pi_{i2}}\right) &= \beta_{01} + \beta_1 x_{TNSC,i} + \beta_2 x_{mild,i} + \beta_3 x_{moderate,i} \\ &= 3.153 - 3.446 \times 10^{-4} x_{TNSC,i} - 4.260 x_{mild,i} - 5.347 x_{moderate,i} \\ \Phi^{-1}\left(\frac{\pi_{i2}}{\pi_{i2}+\pi_{i3}}\right) &= \beta_{02} + \beta_1 x_{TNSC,i} + \beta_2 x_{mild,i} + \beta_3 x_{moderate,i} \\ &= 5.275 - 3.446 \times 10^{-4} x_{TNSC,i} - 4.260 x_{mild,i} - 5.347 x_{moderate,i}\end{aligned} \tag{10}$$

To compare the performance of Model 3 with Model 1 and Model 2, we use the cross-entropy loss based on ten-fold cross-validations similarly to Sections 3.1 and 3.2. Since Model 3 cannot be applied to cases with marked or unknown status of gastric atrophy, we compare the performance of the three models on samples with mild, moderate, or negative status of gastric atrophy only. Their boxplots of cross-entropy loss based on ten random partitions for ten-fold cross-validations are displayed in Figure 7.

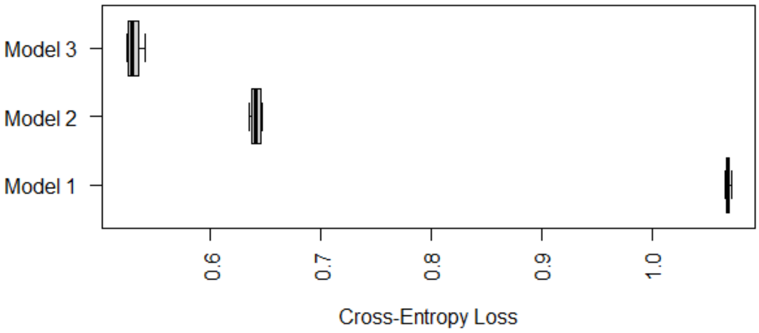

Figure 7. Boxplots of cross-entropy loss (on 98 Samples only) of Models 1, 2, and 3 based on ten-fold cross-validations with ten random partitions.

According to Figure 7, Model 3 has a significantly smaller (average) cross-entropy loss compared with Model 1 and Model 2, in terms of predicting IM for individuals whose gastric atrophy statuses are negative, mild or moderate. Nevertheless, Models 1 and 2 are still useful since they can be applied to cases with marked or unknown status of gastric atrophy as well.

4. Discussion

In Section 3, we presented three models for different scenarios. When only the TNSC (or the DNA methylation profile) is available, we recommend Model 1, a cumulative mixed-link model with po, which works reasonably well with TNSC as the only input. When the status of gastric atrophy is also available, there are two different scenarios. If the status is negative, mild, or moderate, we recommend Model 2, an adjacent-categories logit model with po, which belongs to the traditional multinomial logit models. If the status is marked or unknown, we recommend Model 3 instead, which is an adjacent-categories probit model with po. Each of the three models has its own advantages. For example, although both Model 2 and Model 3 outperform Model 1 in terms of prediction accuracy, Model 1 is still useful when the status of gastric atrophy is not available.

To further compare the performance of Models 1 and 2 on cases with marked or unknown status of gastric atrophy, we display in Figure 8 the (average) cross-entropy loss on predicting those 26 cases with marked or unknown status of gastric atrophy only. According to Figure 8, Model 2 still outperforms Model 1 in predicting the risk level of IM for those 26 cases, which suggests that Model 2 be recommended against Models 1 and 3 for cases with marked or unknown status of gastric atrophy.

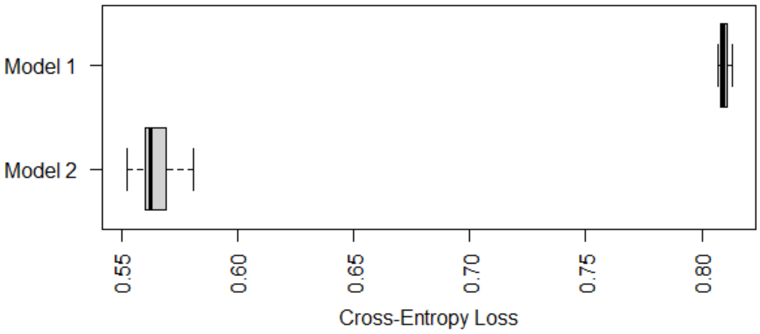

Figure 8. Boxplots of cross-entropy loss (on 26 Samples only) of Models 1 and 2 based on ten-fold cross-validations with ten random partitions.

In practice, more covariates or predictors may be added to the multinomial mixed-link model as well, given their availability. For example, it is known that *Helicobacter pylori* (Hp) infection is an important factor for both IM and gastric cancer development [5,37]. When the Hp status, in terms of Hp serology test result [38], histological examination result [39], or Hp sequence reads [5], is available, one may add it into the model and use AIC, BIC, or cross-validation to determine whether the model with the newly added covariate works significantly better (see Section 2.3).

It should be noted that when using model selection techniques described in Section 2.3, sometimes the differences between the best models are not significant. For example, when selecting Model 3, two other models, a cumulative probit model with po and a continuation-ratio probit model with po, have similar AIC values (see Table 3) that are not significantly smaller than Model 3's [36]. In this case, one may use any of them for prediction purposes. That is saying, with the current data or a finite sample size, those models are comparable or not significantly different from each other.

With an increased sample size, if there is a true statistical model associated with the response and available predictors, then the true model is expected to be among the best models asymptotically [40]. Nevertheless, it does not necessarily mean that the true model is asymptotically identifiable (see [40] for more discussion on asymptotic consistency related to model selections for multinomial models).

In a previous study [5], DNA methylation alteration has been reported as significantly correlated with IM regression at the univariate level. Nevertheless, the significance vanishes when mutation burden and Hp density are incorporated into a multivariate logistic regression analysis [5]. It is worthy of further exploration using the recommended multinomial mixed-link model with the most appropriate link functions selected.

5. Conclusions

In this study, we recommend the newly developed multinomial mixed-link models for predicting Intestinal Metaplasia using DNA methylation profiling. Using model selection techniques, such as AIC, BIC, and cross-validations, we show that the selected multinomial mixed-link model (Model 1) outperforms the traditional multinomial models that assume the same link function for all categories. We also show that when additional information, such as new covariates or predictors, is added to the model, the selection procedure needs to be rerun and the best mixed-link model may change.

When four or more response categories are involved, models other than multinomial mixed-link models have been proposed as well, including two-group models, which can deal with NA or unknown response categories, and po-npo mixture models, which are more flexible than npo, po, or ppo (partial proportional odds) models (see [24] for more examples). Model selection techniques described in Section 2.3 can still be applied, just to a much larger set of candidate models.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/epigenomes8020019/s1, Table S1: IM-epiTOC2-GA-Probabilities.csv, including 124 records with ID, Patient.ID, Response, Site, Gastric.Atrophy, TNSC, predictive probabilities for Normal ($j = 1$), MIM ($j = 2$) and IM ($j = 3$) based on Model 1 (Model1_prob1, Model1_prob2, Model1_prob3), Model 2 (Model2_prob1, Model2_prob2, Model2_prob3), and Model 3 (Model3_prob1, Model3_prob2, Model3_prob3).

Author Contributions: Conceptualization, T.W., Y.H., and J.Y.; methodology, T.W. and J.Y.; software, T.W. and J.Y.; validation, T.W., Y.H., and J.Y.; formal analysis, T.W. and Y.H.; investigation, T.W., Y.H., and J.Y.; resources, T.W., Y.H., and J.Y.; data curation, T.W. and Y.H.; writing—original draft preparation, T.W., Y.H., and J.Y.; writing—review and editing, T.W., Y.H., and J.Y.; visualization, T.W.; supervision, J.Y.; project administration, J.Y.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the U.S. NSF grant DMS-1924859.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: DNA methylation profiling data are publicly available from NCBI (GSE103186) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103186, accessed on 23 January 2024. The clinical data are publicly available online from publication [5] Table S3 at https://www.cell.com/cancer-cell/fulltext/S1535-6108(17)30521-4, accessed on 18 January 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BIC	Bayesian information criterion
CE	cross entropy
CpG	5′—C—phosphate—G—3′ sequence of nucleotides
cloglog	complementary log-log link
DNA	deoxyribonucleic acid
GCEP	Gastric Cancer Epidemiology Program
ID	identifier
IM	intestinal metaplasia
loglog	log-log link
MIM	mild intestinal metaplasia
MLE	maximum likelihood estimate
npo	non-proportional odds assumption
po	proportional odds assumption
PRC2	polycomb repressive complex-2
TNSC	total number of stem cell divisions

Appendix A. AIC and BIC Values of Multinomial Mixed-Link Models Using TNSC for Predicting IM

In this section, we provide a complete list of AIC and BIC values for the multinomial mixed-link models with link functions in {logit, probit, loglog, cloglog}. A "-" in the following tables indicates that the corresponding AIC or BIC value is not available, typically due to numerical issues when fitting the corresponding model.

Table A1. Baseline-category mixed-link models with npo.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	148.49	159.78	147.84	159.12	147.09	158.37	-	-
probit	148.95	160.23	148.27	159.56	147.47	158.75	-	-
loglog	148.83	160.11	148.11	159.39	**146.69**	**157.97**	-	-
cloglog	-	-	-	-	-	-	-	-

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A2. Cumulative mixed-link models with npo.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	148.69	159.97	147.69	158.97	156.46	167.74	-	-
probit	148.35	159.63	147.39	158.67	149.97	161.25	-	-
loglog	146.24	157.52	**145.53**	**156.81**	147.10	158.38	-	-
cloglog	-	-	-	-	-	-	-	-

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A3. Adjacent-categories mixed-link models with npo.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	148.49	159.78	147.87	159.15	146.83	158.11	149.47	160.75
probit	148.54	159.82	147.92	159.20	146.90	158.18	149.51	160.79
loglog	147.65	158.93	147.01	158.29	**145.97**	**157.25**	148.56	159.85
clog log	150.20	161.49	149.62	160.90	148.71	159.99	151.17	162.46

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A4. Continuation-ratio mixed-link models with npo.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	148.95	160.23	148.30	159.58	147.28	158.56	149.81	161.09
probit	148.55	159.84	147.91	159.19	146.88	158.16	149.41	160.70
loglog	147.61	158.89	146.96	158.24	**145.93**	**157.21**	148.47	159.75
clog log	151.95	163.23	151.30	162.58	150.27	161.55	152.81	164.09

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A5. Baseline-category mixed-link models with po.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	173.33	181.79	202.33	210.79	197.84	206.30	188.22	196.68
probit	**151.12**	**159.58**	172.76	181.22	164.81	173.27	162.85	171.31
loglog	159.04	167.50	180.71	189.17	176.63	185.09	170.09	178.55
clog log	156.83	165.29	176.00	184.46	169.15	177.61	166.43	174.89

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A6. Cumulative mixed-link models with po.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	150.23	158.70	194.90	203.36	204.72	213.18	-	-
probit	147.75	156.21	149.25	157.71	155.53	163.99	-	-
loglog	**144.29**	**152.75**	148.81	157.27	147.08	155.54	-	-
cloglog	-	-	-	-	-	-	-	-

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A7. Adjacent-categories mixed-link models with po.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	148.87	157.33	153.58	162.04	153.82	162.28	151.66	160.12
probit	146.57	155.03	148.59	157.05	148.14	156.60	148.17	156.63
loglog	**146.33**	**154.79**	149.87	158.34	149.56	158.02	148.63	157.09
clog log	148.21	156.67	150.03	158.49	149.74	158.20	149.68	158.14

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

Table A8. Continuation-ratio mixed-link models with po.

	logit		probit		loglog		cloglog	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
logit	154.43	162.89	167.30	175.76	165.20	173.66	162.05	170.51
probit	147.39	155.85	153.58	162.04	152.13	160.59	150.99	159.45
loglog	**146.96**	**155.42**	153.40	161.86	151.96	160.43	150.83	159.29
cloglog	152.17	160.63	161.47	169.94	159.67	168.13	157.42	165.88

Note: The AIC/BIC values, associated with the best pair of links, are highlighted in bold.

References

1. Jencks, D.S.; Adam, J.D.; Borum, M.L.; Koh, J.M.; Stephen, S.; Doman, D.B. Overview of current concepts in gastric intestinal metaplasia and gastric cancer. *Gastroenterol. Hepatol.* **2018**, *14*, 92.
2. Filipe, M.I.; Muñoz, N.; Matko, I.; Kato, I.; Pompe-Kirn, V.; Jutersek, A.; Teuchmann, S.; Benz, M.; Prijon, T. Intestinal metaplasia types and the risk of gastric cancer: A cohort study in Slovenia. *Int. J. Cancer* **1994**, *57*, 324–329. [CrossRef] [PubMed]
3. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef] [PubMed]
4. Correa, P. The biological model of gastric carcinogenesis. *IARC Sci. Publ.* **2004**, *157*, 301–310.
5. Huang, K.K.; Ramnarayanan, K.; Zhu, F.; Srivastava, S.; Xu, C.; Tan, A.L.K.; Lee, M.; Tay, S.; Das, K.; Xing, M.; et al. Genomic and epigenomic profiling of high-risk intestinal metaplasia reveals molecular determinants of progression to gastric cancer. *Cancer Cell* **2018**, *33*, 137–150. [CrossRef]
6. Ushijima, T. Epigenetic field for cancerization. *BMB Rep.* **2007**, *40*, 142–150. [CrossRef] [PubMed]
7. Teschendorff, A.E.; Jones, A.; Fiegl, H.; Sargent, A.; Zhuang, J.J.; Kitchener, H.C.; Widschwendter, M. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* **2012**, *4*, 24. [CrossRef] [PubMed]
8. Wang, T.; Tsui, B.; Kreisberg, J.F.; Robertson, N.A.; Gross, A.M.; Yu, M.K.; Carter, H.; Brown-Borg, H.M.; Adams, P.D.; Ideker, T. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol.* **2017**, *18*, 57. [CrossRef] [PubMed]
9. Yamashita, S.; Kishino, T.; Takahashi, T.; Shimazu, T.; Charvat, H.; Kakugawa, Y.; Nakajima, T.; Lee, Y.C.; Iida, N.; Maeda, M.; et al. Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1328–1333. [CrossRef]

10. Tao, Y.; Kang, B.; Petkovich, D.A.; Bhandari, Y.R.; In, J.; Stein-O'Brien, G.; Kong, X.; Xie, W.; Zachos, N.; Maegawa, S.; et al. Aging-like spontaneous epigenetic silencing facilitates Wnt activation, stemness, and BrafV600E-induced tumorigenesis. *Cancer Cell* **2019**, *35*, 315–328. [CrossRef]
11. Cole, J.J.; Robertson, N.A.; Rather, M.I.; Thomson, J.P.; McBryan, T.; Sproul, D.; Wang, T.; Brock, C.; Clark, W.; Ideker, T.; et al. Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. *Genome Biol.* **2017**, *18*, 58. [CrossRef] [PubMed]
12. Teschendorff, A.E. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Med.* **2020**, *12*, 1–17. [CrossRef] [PubMed]
13. Suzuki, K.; Suzuki, I.; Leodolter, A.; Alonso, S.; Horiuchi, S.; Yamashita, K.; Perucho, M. Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. *Cancer Cell* **2006**, *9*, 199–207. [CrossRef] [PubMed]
14. Glonek, G.; McCullagh, P. Multivariate logistic models. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 533–546. [CrossRef]
15. Zocchi, S.; Atkinson, A. Optimum experimental designs for multinomial logistic models. *Biometrics* **1999**, *55*, 437–444. [CrossRef] [PubMed]
16. Bu, X.; Majumdar, D.; Yang, J. D-optimal designs for multinomial logistic models. *Ann. Stat.* **2020**, *48*, 983–1000. [CrossRef]
17. Dousti Mousavi, N.; Aldirawi, H.; Yang, J. Categorical data analysis for high-dimensional sparse gene expression data. *BioTech* **2023**, *12*, 52. [CrossRef] [PubMed]
18. Aitchison, J.; Bennett, J. Polychotomous quantal response by maximum indicant. *Biometrika* **1970**, *57*, 253–262. [CrossRef]
19. Agresti, A. *Categorical Data Analysis*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2013.
20. Greene, W. *Econometric Analysis*; Pearson Education: Hoboken, NJ, USA, 2018.
21. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc. Ser. B* **1980**, *42*, 109–142. [CrossRef]
22. Yang, J.; Tong, L.; Mandal, A. D-optimal designs with ordered categorical data. *Stat. Sin.* **2017**, *27*, 1879–1902. [CrossRef]
23. O'Connell, A. *Logistic Regression Models for Ordinal Response Variables*; Sage: London, UK, 2006.
24. Wang, T.; Tong, L.; Yang, J. Multinomial link models. *arXiv* **2023**, arXiv:2312.16260.
25. Tomasetti, C.; Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **2015**, *347*, 78–81. [CrossRef] [PubMed]
26. Klutstein, M.; Moss, J.; Kaplan, T.; Cedar, H. Contribution of epigenetic mechanisms to variation in cancer risk among tissues. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2230–2234. [CrossRef] [PubMed]
27. Johnstone, S.E.; Gladyshev, V.N.; Aryee, M.J.; Bernstein, B.E. Epigenetic clocks, aging, and cancer. *Science* **2022**, *378*, 1276–1277. [CrossRef] [PubMed]
28. Zheng, S.C.; Widschwendter, M.; Teschendorff, A.E. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics* **2016**, *8*, 705–719. [CrossRef] [PubMed]
29. Zhou, W.; Dinh, H.Q.; Ramjan, Z.; Weisenberger, D.J.; Nicolet, C.M.; Shen, H.; Laird, P.W.; Berman, B.P. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **2018**, *50*, 591–602. [CrossRef] [PubMed]
30. Hannum, G.; Guinney, J.; Zhao, L.; Zhang, L.; Hughes, G.; Sadda, S.; Klotzle, B.; Bibikova, M.; Fan, J.B.; Gao, Y.; et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **2013**, *49*, 359–367. [CrossRef] [PubMed]
31. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, 2–8 September 1971; Akademiai Kiado: Budapest, Hungary, 1973 .
32. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
33. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
35. McCullagh, P.; Yang, J. Stochastic classification models. In Proceedings of the International Congress of Mathematicians, Madrid, Spain, 22–30 August 2006; Volume III, pp. 669–686.
36. Burnham, K.P.; Anderson, D.R. Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
37. Correa, P.; Piazuelo, B.M.; Wilson, K.T. Pathology of gastric intestinal metaplasia: Clinical implications. *Am. J. Gastroenterol.* **2010**, *105*, 493–498. [CrossRef] [PubMed]
38. Veijola, L.; Oksanen, A.; Sipponen, P.; Rautelin, H. Evaluation of a commercial immunoblot, Helicoblot 2.1, for diagnosis of Helicobacter pylori infection. *Clin. Vaccine Immunol.* **2008**, *15*, 1705–1710. [CrossRef] [PubMed]
39. Calvet, X.; Sánchez-Delgado, J.; Montserrat, A.; Lario, S.; Ramírez-Lázaro, M.J.; Quesada, M.; Casalots, A.; Suárez, D.; Campo, R.; Brullet, E.; et al. Accuracy of diagnostic tests for Helicobacter pylori: A reappraisal. *Clin. Infect. Dis.* **2009**, *48*, 1385–1391. [CrossRef] [PubMed]
40. Wang, T.; Yang, J. Identifying the most appropriate order for categorical responses. *arXiv* **2024**, arXiv:2206.08235.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.