Journal of Data Analysis and Information Processing, 2026, 14(1), 49-62
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Cumulative Link Modeling of Ordinal Outcomes in the National Health Interview Survey Data: Application to Depressive Symptom Severity

Andre Williams, Louisana Louis

Christine E. Lynn College of Nursing, Florida Atlantic University, Boca Raton, FL, USA
Email: wandre719@gmail.com

How to cite this paper: Williams, A. and Louis, L. (2026) Cumulative Link Modeling of Ordinal Outcomes in the National Health Interview Survey Data: Application to Depressive Symptom Severity. *Journal of Data Analysis and Information Processing*, **14**, 49-62.
https://doi.org/10.4236/jdaip.2026.141004

Received: November 15, 2025
Accepted: December 20, 2025
Published: December 23, 2025

Abstract

This study investigates the application of cumulative link models with alternative distributions (hyperbolic secant, Laplace, and Cauchy) to model ordinal outcomes of depressive severity using 2022 National Health Interview Survey data. The primary objective was to assess whether these models provide a better fit to ordinal response data and more accurate predictions than their traditional counterparts with the logit link function. The results indicate that the logit model achieved the highest classification accuracy, correctly classifying 83.54% of the cases. The Cauchy model demonstrated the best model fit, *i.e.*, the lowest AIC and BIC values. This study highlights the importance of considering both classification accuracy and model fit when selecting a statistical model.

Keywords

Cumulative Link Model, Ordinal Outcome, Depressive Symptom Severity, National Health Interview Survey

1. Introduction

Ordinal response data refers to data with a categorical outcome having natural, ordered categories, but with unknown distances between these categories. Cumulative link models (CLMs) are statistical models specifically designed to analyze this data type. These models have found widespread applications in various fields. In the social sciences, for example, CLMs can be used to model attitude responses on a Likert scale [1]. In medicine, they can be used to analyze disease severity with cancer stage. In ecology, CLMs can be used to study the effects of factors on the spatial abundance of a species, which is categorized into distinct levels [2]. CLMs provide a flexible framework for modeling the relationship between an ordinal outcome and a set of predictor variables while preserving the inherent ordering of the response categories [3].

CLMs link the cumulative probabilities of the ordinal response to a set of predictors through a suitable link function. This process assumes that the observed ordinal response is a manifestation of an underlying continuous variable with a corresponding distribution that is not directly observed. Commonly used link functions include the logit, probit, cumulative log-log, and log-log, which correspond to the logistic, normal, Gompertz, and Gumbel distributions [4]. The CLM with a logit link, also known as the proportional odds model, is widely used in research. These link functions impose certain assumptions on the underlying latent variable that generates the observed ordinal response. However, the choice of link function and the associated distributional assumptions can significantly influence the model's performance and the interpretation of its results [5]. Extensions of CLMs include the incorporation of dispersion effects, in which the explanatory variables affect not only the location of the ordinal response but also its spread or variability [6].

While traditional CLMs with logit or probit links are widely used, they may not always be appropriate. Each CLM assumes a distribution for the unobserved latent variable. However, this assumption may not hold when the underlying data exhibit skewness, kurtosis, or boundary inflation, leading to poorer model performance. A limitation is their inability to adequately capture complex relationships in certain situations. For instance, in surveys assessing mental health symptom severity, responses might be heavily skewed towards lower symptomatology levels [7]. In behavioral health studies, many patients may fall into the 'minimal depressive symptomatology' category. In these cases, alternative distributions potentially offering greater flexibility in modeling the shape of the latent variable distribution may be more suitable. Additional CLMs with associated distributions must be evaluated as they may better fit the data when applying the ordinal outcome model.

This manuscript proposes using the hyperbolic secant, Laplace, and Cauchy distributions as a group of candidate distributions for CLMs. Integrating hyperbolic secant and Laplace distributions into CLMs represents a novel consideration within behavioral research. The hyperbolic secant distribution and its generalizations are extensively used in financial modeling; characterized by slightly fatter tails than the normal distribution, it is particularly adept at accommodating datasets with larger-than-average observations [8] [9]. Moreover, this distribution has consistently demonstrated a robust fit across the entire range of data support. The Laplace distribution has also been employed in financial modeling because it captures the leptokurtic and skewed nature of financial data [10]. The Laplace distribution has a sharper peak at the mean than the normal distribution, but heavier tails due to slower decay. The Cauchy distribution is also bell-shaped and symmetric, but with much heavier tails than the normal distribution [11].

Given the limitations of traditional CLMs and the flexibility offered by the distributions under consideration, this research investigates whether CLMs with hyperbolic secant, Laplace, and Cauchy distributions provide a better fit to ordinal response data. Choosing an appropriate latent distribution enables the model to capture the underlying structure of the ordinal data, leading to more accurate and consistent predictions and better model fit. Specifically, we aim to determine whether these models offer improved accuracy, model fit, and variable selection compared to their traditional counterparts with the logit link function. To achieve this, we address the following research question:

- Do CLMs with the hyperbolic secant, Laplace, and Cauchy distributions provide a better fit to ordinal response data and more accurate predictions than traditional CLMs with a logit link?

We conducted a study and analyzed the 2022 National Health Interview Survey (NHIS) dataset with depressive symptom severity (minimal, mild, moderate, and severe) as the ordinal outcome of interest [12] to address the research question. The NHIS is a comprehensive, nationwide survey conducted annually by the Centers for Disease Control and Prevention that collects data on a wide range of health topics, including chronic conditions, health insurance coverage, and access to healthcare services [13]. The models were evaluated based on predictive accuracy, model fit, macro-average F1 score, and mean decrease in variable accuracy. The findings of this research will contribute to the development of additional CLMs that can better capture the complexities of ordinal response data.

2. Materials and Methods

For a given observation *i*, denote $\boldsymbol{x}_i = \left(x_{i1} \;\; x_{i2} \;\; \cdots \;\; x_{ip} \right)$ as a vector of *p* covariates, with $i = 1, 2, \cdots,, n$. For this study, the predictor variables comprised sociodemographic and healthcare access-related variables from the 2022 NHIS dataset. A measure of anxiety was also included as a covariate. In addition, an ordinal outcome of depressive severity was recorded. As such, we denote the outcome vector $\boldsymbol{y}_i = \left(y_{i1}, y_{i2}, \cdots, y_{iJ} \right)$ where $y_{ij} = 1$ if, for observation *i*, the outcome is in the j^{th} category, with all other entries being set to 0. There are J possible outcome levels, with depressive severity measured by the PHQ-8. These levels are categorized as follows:

1. Minimal (PHQ-8 score under five)
2. Mild (PHQ-8 score from five to less than 10)
3. Moderate (PHQ-8 score from 10 to less than 15)
4. Severe (PHQ-8 score of 15 or higher)

For this given study, the goal is to use the covariate vector $\boldsymbol{x}_i$ to predict depressive severity, $\boldsymbol{y}_i$, using the 2022 NHIS data. The aggregation of vectors for all subjects yields the covariate matrix $\boldsymbol{X}$, a *p* by *n* matrix where the i^{th} column is set to $\boldsymbol{x}_i$. We also have a *J* by *n* matrix, denoted $\boldsymbol{Y}$, where the i^{th} column is set to $\boldsymbol{y}_i$. The goal is to develop a function $f : \boldsymbol{X} \to \boldsymbol{Y}$, which aims to predict the $\boldsymbol{Y}$ matrix using the $\boldsymbol{X}$ matrix. To achieve this, an ordinal regression framework was used. First, four CLMs are introduced, two of which are novel applications (based on the hyperbolic secant and Laplace distributions). Next, the models' specifications are outlined. The method is then applied to the 2022 NHIS data, with results on predictive accuracy, macro-averaged F1 score, model fit, and variable importance reported.

2.1. Cumulative Link-Based Outcome Functions

The cost function for the neural network is derived from the log-likelihood of a multinomial distribution:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \times \log\left(\pi_{ij}\right), \tag{1}$$

where $\pi_{ij} = P\left(y_{ij} = 1 | \boldsymbol{x}_i\right)$. Define the cumulative probabilities p_{ij} as $p_{ij} = P\left(y_{ij} = 1 | \boldsymbol{x}_i\right) + P\left(y_{ij-1} = 1 | \boldsymbol{x}_i\right) + \cdots + P\left(y_{i1} = 1 | \boldsymbol{x}_i\right)$. In addition, $p_{iJ} = 1$ and $p_{i0} = 0$. As such, we can write Equation (1) as

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \times \log\left(p_{ij} - p_{ij-1}\right). \tag{2}$$

We are primarily concerned with directly modeling cumulative probabilities and, to this end, employed CLMs.

CLMs are statistical models intended to analyze data that have ordinal outcomes. We are concerned with modeling the cumulative probabilities p_{ij}. In CLMs, p_{ij} is linked to a function of predictor variables through a specified link function. This method enables the use of a covariate set to account for variation in the ordinal outcome while preserving the natural order of the response categories. The four-link functions considered in this study are based on the following distributions:

1. Logistic Distribution (logit)
2. Hyperbolic Secant
3. Cauchy
4. Laplace

These special distributions (hyperbolic secant, Laplace, and Cauchy), rather than the more common alternatives (normal (probit) and Gumbel (complementary log-log)), were chosen due to their tail and kurtosis properties to allow greater flexibility to accommodate the skewed and heavy-tailed characteristics of the behavioral health data that the other distributions may not capture.

The goal of this study is to evaluate the performance of the link functions with respect to predictive accuracy (defined as the model's performance on a validation dataset), model fit, and variable selection (defined as the mean decrease in accuracy).

By employing the logit link, the cumulative probabilities are modeled as:

$$p_{ij}^{[1]} = \frac{1}{1 + \mathrm{e}^{-\left(\mathbf{x}_i \boldsymbol{\beta}_j + b_j\right)}} \tag{3}$$

where b_j is defined as the intercept parameter for the *j*th level. Considering the hyperbolic secant distribution as the latent distribution of interest, the cumulative probabilities are now modeled as:

$$p_{ij}^{[2]} = \frac{2}{\pi}\arctan\left[\exp\left(\frac{\pi}{2}\left\{\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j\right\}\right)\right]. \tag{4}$$

When the underlying latent distribution is assumed to be a Cauchy distribution, we have

$$p_{ij}^{[3]} = \frac{1}{\pi}\arctan\left(\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j\right) + 0.5 \tag{5}$$

Finally, utilizing the Laplace distribution as the underlying latent distribution, we have:

$$\begin{aligned} p_{ij}^{[4]} = {} & I\left(\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j \geq 0\right) \\ & - sign\left(\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j\right)\frac{1}{2}\exp\left(\left\{\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j\right\} \times \left(-sign\left(\boldsymbol{x}_i\boldsymbol{\beta}_j + b_j\right)\right)\right), \end{aligned} \tag{6}$$

These cumulative probabilities can be substituted into equation (2) and solved accordingly. The goal is to find parameter estimates such that:

$$\left(\hat{\boldsymbol{\beta}}_j, \hat{b}_j\right) = \underset{\beta_j, b_j}{\arg\min}\left\{-\log L\right\}. \tag{7}$$

As such, the Adam optimization algorithm was applied [14] [15]. Once the optimal values for $\hat{\boldsymbol{\beta}}$ and $\hat{b}_j$ were computed, they were used to evaluate the models regarding accuracy, model fit, and variable selection.

2.2. Application to NHIS Data

The NHIS is a pivotal, cross-sectional household survey conducted annually by the National Center for Health Statistics (NCHS) under the Centers for Disease Control and Prevention (CDC) [16]. The primary aim of this survey is to monitor the health status of the U.S. population and to track trends in essential health indicators. The survey gathers comprehensive data on a broad spectrum of health-related topics, including chronic and acute conditions, health insurance coverage, utilization of healthcare services, and health-related behaviors. Given that the NHIS provides a nationally representative sample of the civilian noninstitutionalized population, its data are widely used by researchers and policymakers to assess public health needs, evaluate health policies, and inform interventions to enhance Americans' health [16]. In the context of modeling depressive severity, the NHIS is particularly valuable due to its use of the PHQ-8, a widely accepted screening tool for depressive symptoms [17]. The instrument is a truncated version of the PHQ-9, aligns with DSM-IV criteria [18], and assesses symptom frequency over the past two weeks, enabling the determination of depression severity [19]. The survey's rich collection of sociodemographic and health-related variables further enables researchers to explore how a variety of factors, from income and education to access to care, are associated with depression severity across a nationally representative sample of the population.

For this study, the 2022 survey data were used. The outcome variable is an ordinal measure of the PHQ-8, as described earlier in the methods section. Due to the relatively small sample sizes at moderate and severe levels, these two levels were combined into a single moderate/severe level. Selection of specific predictor variables was guided by established literature on sociodemographic and healthcare access-related risk factors for depression [20]. The predictor variables include age, poverty levels (less than 100% Federal Poverty Level (FPL), between 100% and 199% FPL, between 200% and 300% FPL, and greater than 400% FPL), sex (male, female), race/ethnicity (non-Hispanic white, non-Hispanic black, other), education level (some high school, high school graduate, some college, bachelor's, master's, professional or doctoral degree), health insurance (private, Medicare, other, uninsured), any delay in receiving medical care over the past 12 months (yes, no), foregoing medical care due to cost in the past 12 months (yes, no), having a usual place for medical care (yes, no), the number of urgent care visits in the past year (0, 1, >1), the number of emergency room visits in the past year (0, 1, >1), any overnight hospitalizations in the past year (yes, no), living alone (yes, no), and GAD-7. The GAD-7 is represented ordinally as:

1. Minimal (GAD-7 score less than five)
2. Mild (GAD-7 score greater than or equal to five and less than 10)
3. Moderate (GAD-7 score greater than or equal to 10 and less than 15)
4. Severe (GAD-7 score greater than or equal to 15)

To evaluate the models on the 2022 NHIS dataset, the data was split into a training set (80%) and a validation set (20%). The model parameters for the four cumulative link regression models were optimized on the training dataset. Once optimized, the model parameters $\hat{\boldsymbol{\beta}}_j$ and $\hat{b}_i$ were applied and evaluated on the validation data. The metrics reported include the accuracy (percent correctly classified and percent correctly classified in the moderate/severe depressive symptomatology category), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and macro-averaged F1 score [21], and the mean decrease in accuracy for each variable. The macro-average F1 score computes the F1 score for each class in a multi-class classification problem and then averages those scores. It gives equal weight to all classes, regardless of the number of samples in each class, and is useful for comparing performance on imbalanced datasets where some classes may have few examples. The mean decrease in accuracy (also known as Permutation Feature Importance) is a technique for quantifying the importance of a feature by measuring the average drop in a model's prediction accuracy when the feature values are randomly permuted [22]. Permuting the feature values breaks the relationship between the feature values and the true outcome, effectively excluding it from the model. This process was repeated 100 times for each variable, with the average measure being reported. A high score indicates that the model is highly reliant on that feature, as its performance drops significantly when the feature values are randomly permuted. All analyses were performed using the R software program [23].

3. Results

The study applied four cumulative link models to the 2022 NHIS dataset to evaluate their performance in modeling depressive severity. The sample size for this study is 25,208. The models included logit, hyperbolic secant, Cauchy, and Laplace distributions. The primary evaluation metrics were the percentage of correctly classified cases with 95% confidence intervals, AIC, BIC, macro-average F1 score, and variable importance measured by the decrease in accuracy. Table 1 illustrates the distribution of the ordinal outcome within the National Health Interview Survey (NHIS) dataset. A substantial proportion of participants exhibited minimal depressive symptomatology, accounting for 79.07% of the sample. In contrast, a smaller fraction of individuals reported moderate and severe symptomatology, comprising 4.36% and 2.8% of the sample, respectively. Consequently, for the purposes of further modeling analysis, the moderate and severe categories were amalgamated into a single category.

Table 1. Descriptive statistics for the ordinal outcome of PHQ-8 symptom severity.

	PHQ-8 (Categorized)			
	Minimal	Mild	Moderate	Severe
N (%)	19933 (79.07%)	3468 (13.76%)	1100 (4.36%)	707 (2.8%)

Due to the relatively small sample size, moderate and severe were combined into a single group.

Table 2. Ordinal regression network accuracy.

	Model Type			
	Logit	Hyperbolic Secant	Cauchy	Laplace
Correctly Classified (95% Confidence Interval)	83.54% (82.49%, 84.55%)	83.42% (82.36%, 84.44%)	80.78% (79.67%, 81.86%)	83.16% (82.1%, 84.18%)
AIC	7511.46	11830.76	6127.67	8722.97
BIC	7655.4	11974.7	6271.61	8866.91
Macro-averaged F1 Score	0.62	0.61	0.49	0.6

The percentage correctly classified by the four models (logit, hyperbolic secant, Cauchy, and Laplace) applied to the 2022 National Health Interview Survey data.

Table 2 reports the metrics regarding predictive accuracy and model fit as applied to the validation dataset. Regarding model performance in terms of correctly classified percentage, the logit model achieved the highest accuracy, correctly classifying 83.54% of the cases studied. Closely following, the hyperbolic secant and Laplace models achieved accuracies of 83.42% and 83.16%, respectively. Conversely, the Cauchy model exhibited the lowest accuracy, correctly classifying

80.78% of the cases. Lower AIC and BIC scores indicate better model performance. Interestingly, the Cauchy model performed best, presenting the lowest AIC (6127.67) and BIC (6271.61) values, surpassing the other models. In contrast, the logit model yielded AIC and BIC values of 7511.46 and 7655.40, respectively, while the hyperbolic secant and Laplace models recorded even higher AIC and BIC values, indicating poorer performance than the Cauchy model. Regarding the macro-averaged F1 score, the logit model achieved the highest score, followed by the hyperbolic secant and Laplace-based models; the Cauchy model had the lowest score.

To assess the models' predictive capability at the extremes, predictive accuracy was evaluated in the minority class of moderate/severe depressive symptom severity. The percent correctly classified on the validation data is listed as follows:

1. Logistic: 53.37%
2. Hyperbolic Secant: 56.13%
3. Cauchy: 17.79%
4. Laplace: 53.06%

As such, the hyperbolic secant outperformed all other models when predicting the extreme class of moderate/severe depression symptom severity.

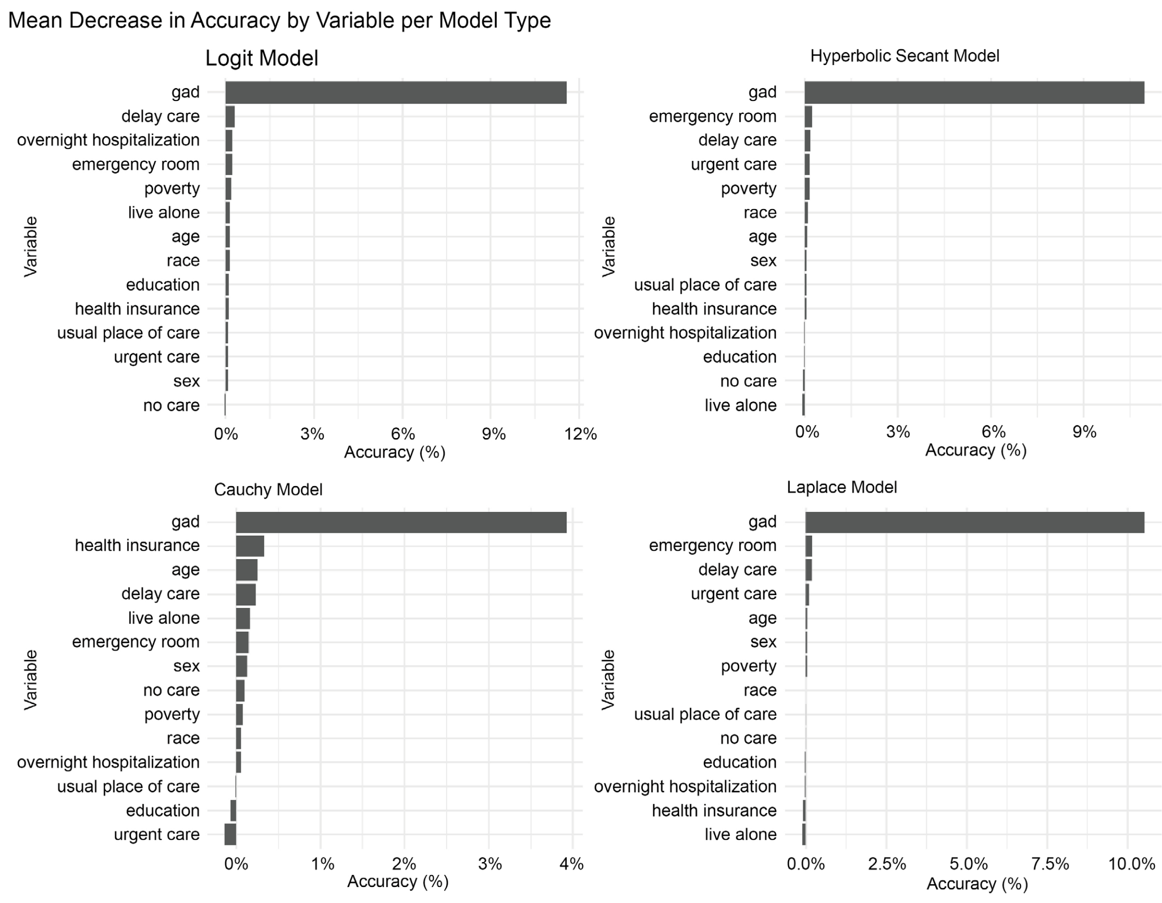

Figure 1. The plot presents the mean decrease in accuracy per variable for the four models presented. The x-axis represents decrease in accuracy, while the y-axis displays the variables used in the model.

Figure 1 shows the average decrease in accuracy across the four models tested on the validation dataset. The mean decrease in accuracy was used to evaluate the importance of predictor variables in each model. When analyzing the mean decrease in accuracy across the four models, the GAD-7 severity variable consistently shows the largest decrease. This indicates that GAD-7 has the greatest effect on model performance. In both the Laplace and hyperbolic secant models, a similar pattern emerges: variables such as past 12-month delay in receiving medical care and past-year number of emergency room visits are associated with small drops in accuracy. Conversely, within the Cauchy model framework, the variables health insurance and age have a more significant influence than in the other models, though still minor. The logit model shows that the variables—any overnight hospitalizations in the past year and any delay in receiving medical care in the past 12 months—are the next most important factors affecting accuracy, after GAD-7.

4. Discussion

This study explored the application of CLMs with alternative distributions—hyperbolic secant, Laplace, and Cauchy—to model ordinal outcomes of depressive severity using 2022 NHIS data. The primary objective was to assess whether these models provide a better fit and more accurate predictions than traditional CLMs with a logit link.

The results indicate that the logit model achieved the highest classification accuracy. The logit also achieved the highest macro-averaged F1 score. The hyperbolic secant-based model achieved the highest predictive accuracy when predicting the minority class of moderate/severe depression symptom severity. A plausible reason is that the logistic cumulative distribution function naturally arises when the log-likelihood of the multinomial distribution is derived. Also, the logit link function performed well in estimating the threshold for the minimal category. However, the Cauchy model demonstrated superior model fit, as evidenced by the lowest AIC and BIC values. A possible hypothesis for the improved performance of the Cauchy's heavy tails is that they do a better job of accommodating the individual whose depression symptom severity scores are in the moderate/severe range. The tails of the Cauchy distribution decay more slowly than those of the logistic distribution, so it might be able to model the extreme observation of moderate/severe depression symptom severity more accurately without being overly influenced by them, leading to a higher likelihood value, and lower AIC and BIC across the dataset. The Cauchy distribution may sacrifice accuracy, but it gains by better representing the true distributional shape, especially at the extremes. These findings suggest that while the logit model is effective for classification tasks, the Cauchy model may offer a more nuanced understanding of the data structure, particularly when model fit is prioritized.

The trade-off between the highest predictive accuracy and better model fit can have significant practical implications for health researchers: choices about which model to use should depend on the study's main goal. For instance, in a healthcare setting, if the objective is to develop a tool that accurately classifies patients into depression severity categories (such as for automated screening or triage), the model that correctly classifies the most cases should be preferred. However, if the goal of a study is to gain a deeper or more nuanced understanding of the relationship between the predictors and the underlying latent variable (like depression severity), model fit (AIC and BIC) should be prioritized.

The variable importance analysis showed that the GAD-7 severity variable consistently had the highest mean decrease in accuracy across all models, indicating its significant impact on predicting depressive severity. The logit and hyperbolic secant models also identified any delay in receiving medical care over the past 12 months and the number of emergency room visits in the past year as key factors, while the Cauchy model emphasized the importance of health insurance and age. This difference in variable importance across models suggests that various distributions may capture different aspects of the data, providing complementary insights.

We acknowledge limitations due to data preparation, namely, the requirement to combine the moderate and severe depressive symptom categories into a single moderate/severe symptom category (due to small sample sizes in the validation dataset), which assured model fitting power and stability, but potentially masks meaningful and clinically relevant differences between moderate and severe depression symptom severity. All future interpretations of the results need to be cautious about the exact clinical meaning of the moderate/severe category, because the model's outputs and variable importance for the moderate/severe outcome reflect the merged group rather than two distinct clinical entities.

The findings underscore the potential of alternative distributions, such as the Cauchy, to capture the complexities of ordinal response data, particularly in behavioral health research. The ability of these models to accommodate skewness and heavy tails makes them suitable for datasets with extreme observations, which are common in mental health surveys. Future research could explore integrating these CLMs with machine learning techniques such as penalized modeling, deep learning, and explainable AI (XAI) to further enhance predictive accuracy, model interpretability, and explainability.

Despite the higher-level performance of the logit and Cauchy functions compared to the hyperbolic secant and Laplace functions, these latter functions retain their value, particularly in the context of sensitivity analyses. These analyses are crucial for validating the assumption that the underlying latent variable conforms to a specified distribution. In the health sciences, the logit link function is predominantly used for ordinal regression because it produces regression coefficients that are readily interpretable as odds ratios [24]. However, the other three models should be considered as viable alternatives, as it is entirely possible that they may more closely align with the underlying latent distribution.

Additionally, there was a high-class imbalance in the ordinal outcome variable of depression symptom severity as measured by the PHQ-8, and some of the input variables were ordinal as well. A possible remedy to this was the application of Traditional Synthetic Minority Oversampling Technique (SMOTE) algorithms to address the class imbalance. However, for this study, SMOTE algorithms could not be directly applied, as none of the current versions accommodate ordinal outcomes and ordinal input variables [25]. Currently, there are no SMOTE-based algorithms in R or Python that can accommodate ordinal input and outcome variables. Additional research in algorithmic development is needed to develop SMOTE algorithms that can accommodate the complexities of high-dimensional biomedical data. A future SMOTE implementation designed to effectively address the methodological gap of class imbalance in ordinal data would require the following specific features and capabilities:

1) Ordinal-Aware Distance Metric: The core capability is a distance function that respects the ordered, non-numeric nature of ordinal variables. For example, the distance between categories 1 and 2 is smaller than that between 1 and 4, but it is not calculated by simple subtraction, as with continuous variables. The algorithm must use metrics that consider cumulative probabilities or employ rank-based distances.

2) Synthetic Sample Generation for Ordinal Outcomes: The algorithm must synthesize new minority-class samples for the ordinal outcome, ensuring they maintain the inherent ordering structure of the outcome variable. New synthetic outcomes should fit logically within existing categories (e.g., a new synthetic severity score should be labeled “Mild” or “Moderate”).

Previous studies have demonstrated the effectiveness of the SMOTE algorithm in improving prediction accuracy for unbalanced data with binary outcomes and no ordinal input variables [26].

This study highlights the importance of considering both classification accuracy and model fit when selecting a statistical model. Researchers should weigh these factors based on the specific objectives of their study, whether they aim to maximize predictive accuracy or to gain deeper insights into model fit. For future project implementation, one can begin by clearly defining the research question, such as whether the focus is on predictive accuracy, accuracy within a specific subgroup, or overall model fit. If pilot, preliminary, or real-world data are available, the candidate models can be evaluated, and the model with the highest performance on the prespecified metric of interest will be selected as the main model for the primary analysis of the main study. Alternative models can be used for secondary and sensitivity analyses. It is also helpful to consider additional CLMs, including probit, cloglog, and loglog. Additionally, the extensive nature of the NHIS data facilitates the examination of health from a caring science perspective, offering insights into the humanistic and relational dimensions of healthcare [27]. This aligns with the core concept of health as the fundamental category of caring, aiming to support and strengthen an individual’s health processes [28].

5. Conclusion

In conclusion, this study highlighted that using CLMs with alternative distributions, such as the hyperbolic secant, Laplace, and Cauchy, offers a promising avenue for improving the analysis of ordinal outcomes in health research. Use of such models may provide a better fit to the latent distribution than traditional models, such as the logit link, can capture effectively due to skewness and heavy tails in real-life data. In the present study, we observed that although the logit link model had the highest classification accuracy and macro-averaged F1 score, the hyperbolic secant model demonstrated the highest accuracy in the extreme cases, and the Cauchy model had the best model fit, *i.e.*, the lowest AIC and BIC values. This shows that model selection depends on the study's objective: whether model fit or predictive accuracy is the primary focus. Further, variation in variable importance across models shows the potential of these alternative distributions to capture different aspects of the data, which could enrich data analysts' options for choosing an appropriate model. By expanding the toolkit of available models, researchers can better address the complexities inherent in real-world data, ultimately leading to more informed decision-making.

Data Availability Statement

The data are accessed via the URL:
https://www.cdc.gov/nchs/nhis/documentation/2022-nhis.html.

Conflicts of Interest

The authors declared no potential conflicts of interest regarding this article's research, authorship, and/or publication.

References

[1] Larasati, A., DeYong, C. and Slevitch, L. (2011) Comparing Neural Network and Ordinal Logistic Regression to Analyze Attitude Responses. *Service Science*, **3**, 304-312. https://doi.org/10.1287/serv.3.4.304

[2] Guisan, A. and Harrell, F.E. (2000) Ordinal Response Regression Models in Ecology. *Journal of Vegetation Science*, **11**, 617-626. https://doi.org/10.2307/3236568

[3] Christensen, R.H.B. (2018) Cumulative Link Models for Ordinal Regression with the R Package Ordinal. *Journal of Statistical Software*, **35**, 1-46.

[4] Tutz, G. (2011) Regression for Categorical Data. Cambridge University Press. https://doi.org/10.1017/cbo9780511842061

[5] Agresti, A. (2010) Analysis of Ordinal Categorical Data. Wiley. https://doi.org/10.1002/9780470594001

[6] Tutz, G. and Berger, M. (2017) Separating Location and Dispersion in Ordinal Regression Models. *Econometrics and Statistics*, **2**, 131-148. https://doi.org/10.1016/j.ecosta.2016.10.002

[7] Arvelo, I. and Plantinga, A. (2023) U.S. Mental Health Dashboard. *The New England Journal of Statistics in Data Science*, **2**, 323-329. https://doi.org/10.51387/23-nejsds52

[8] Fischer, M.J. (2013) Generalized Hyperbolic Secant Distributions: With Applications to Finance. Springer Science & Business Media.

[9] Palmitesta, P. and Provasi, C. (2004) GARCH-Type Models with Generalized Secant Hyperbolic Innovations. *Studies in Nonlinear Dynamics & Econometrics*, **8**, Article

7. https://doi.org/10.2202/1558-3708.1212

[10] Jing, H., Liu, Y. and Zhao, J. (2022) Asymmetric Laplace Distribution Models for Financial Data: Var and Cvar. *Symmetry*, **14**, Article 807. https://doi.org/10.3390/sym14040807

[11] Nolan, J.P. (2013) Financial Modeling with Heavy-Tailed Stable Distributions. *WIREs Computational Statistics*, **6**, 45-55. https://doi.org/10.1002/wics.1286

[12] Zablotsky, B., Weeks, J.D., Terlizzi, E.P., Madans, J.H. and Blumberg, S.J. (2022) Assessing Anxiety and Depression: A Comparison of National Health Interview Survey Measures. https://stacks.cdc.gov/view/cdc/117491

[13] Terlizzi, E. and Zablotsky, B. (2024) Symptoms of Anxiety and Depression among Adults: United States, 2019 and 2022. *National Health Statistics Reports*, No. 213, CS353885.

[14] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv: 1412.6980.

[15] Williams, A.A.A. (2019) Ordinal Outcome Modeling: The Application of the Adaptive Moment Estimation Optimizer to the Elastic Net Penalized Stereotype Logit. *Journal of Data Analysis and Information Processing*, **7**, 14-27. https://doi.org/10.4236/jdaip.2019.71002

[16] Zablotsky, B., Lessem, S.E., Gindi, R.M., Maitland, A.K., Dahlhamer, J.M. and Blumberg, S.J. (2023) Overview of the 2019 National Health Interview Survey Questionnaire Redesign. *American Journal of Public Health*, **113**, 408-415. https://doi.org/10.2105/ajph.2022.307197

[17] Arias de la Torre, J., Vilagut, G., Ronaldson, A., Dregan, A., Ricci-Cabello, I., Hatch, S.L., *et al.* (2021) Prevalence and Age Patterns of Depression in the United Kingdom. A Population-Based Study. *Journal of Affective Disorders*, **279**, 164-172. https://doi.org/10.1016/j.jad.2020.09.129

[18] Ma, F. (2021) Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5). In: u, D. and Dupre, M.E., Eds., *Encyclopedia of Gerontology and Population Aging*, Springer International Publishing, 1414-1425. https://doi.org/10.1007/978-3-030-22009-9_419

[19] Ajele, K.W. and Idemudia, E.S. (2025) Charting the Course of Depression Care: A Meta-Analysis of Reliability Generalization of the Patient Health Questionnaire (PHQ-9) as the Measure. *Discover Mental Health*, **5**, Article No. 50. https://doi.org/10.1007/s44192-025-00181-x

[20] Califf, R.M., Wong, C., Doraiswamy, P.M., Hong, D.S., Miller, D.P. and Mega, J.L. (2021) Importance of Social Determinants in Screening for Depression. *Journal of General Internal Medicine*, **37**, 2736-2743. https://doi.org/10.1007/s11606-021-06957-5

[21] Opitz, J. and Burst, S. (2019) Macro F1 and Macro F1. arXiv: 1911.03347.

[22] Wang, H. (2023) Research on the Application of Random Forest-Based Feature Selection Algorithm in Data Mining Experiments. *International Journal of Advanced Computer Science and Applications*, **14**, 505-518. https://doi.org/10.14569/ijacsa.2023.0141054

[23] R Core Team (2025) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[24] Singh, V., Dwivedi, S.N. and Deo, S.V.S. (2020) Ordinal Logistic Regression Model Describing Factors Associated with Extent of Nodal Involvement in Oral Cancer Patients and Its Prospective Validation. *BMC Medical Research Methodology*, **20**, Arti-

cle No. 95. https://doi.org/10.1186/s12874-020-00985-1

[25] Elreedy, D. and Atiya, A.F. (2019) A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. *Information Sciences*, **505**, 32-64. https://doi.org/10.1016/j.ins.2019.07.070

[26] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S. (2017) Predicting Diabetes Mellitus Using SMOTE and Ensemble Machine Learning Approach: The Henry Ford Exercise Testing (FIT) Project. *PLOS ONE*, **12**, e0179805. https://doi.org/10.1371/journal.pone.0179805

[27] Zajacova, A., Huzurbazar, S. and Todd, M. (2017) Gender and the Structure of Self-Rated Health across the Adult Life Span. *Social Science & Medicine*, **187**, 58-66. https://doi.org/10.1016/j.socscimed.2017.06.019

[28] Bandura, A. (1997) Editorial. *American Journal of Health Promotion*, **12**, 8-10. https://doi.org/10.4278/0890-1171-12.1.8