The AAPS Journal (2019) 21: 102
DOI: 10.1208/s12248-019-0370-6

Check for updates

Commentary

On the Comparison of Methods in Analyzing Bounded Outcome Score Data

Chuanpu Hu[1,2]

Received 1 June 2019; accepted 1 August 2019; published online 26 August 2019

Abstract. Clinical trial endpoints often take the form of bounded outcome scores (BOS) which report a discrete set of values on a finite range. Conceptually such endpoints are ordered categorical in nature, but in practice they are often analyzed as continuous variables, which may result in data range violations and difficulties to handle data skewness. Analysis methods dedicated for BOS data have been proposed; however, much confusion exists among pharmacometricians on how to compare the possible methods. This commentary reviews the main methods used in pharmacometrics applications and discusses their theoretical and practical comparisons. The expected performance of some conceptually appealing methods in different situations is discussed, and a guideline is provided on selecting analysis methods in practice.

KEY WORDS: categorical data; likelihood; model selection; nonlinear mixed-effects modeling; transformation.

Clinical trial endpoints are often bounded outcome scores (BOS) which take restricted values on finite intervals and achieve boundary values (1). Such endpoints are often summations or weighted averages of composite subscores. An example is the Psoriasis Area and Severity Index (PASI) score, ranged 0–72 with 0.1 increments (2). Due to the typically large (>10) number of possible values, they are often analyzed as continuous data, although conceptually they are ordered categorical variables (1). The continuous analysis approach may not handle skewed data well and may predict the data outside of its natural range, and the ordered categorical analysis approach may require too many intercept parameters (1). Several dedicated approaches have been proposed to describe BOS data distributions (1,3–6). How to compare these approaches is a complex issue that represents a significant challenge even for highly experienced, established pharmacometricians and is the focus of this commentary.

ANALYSIS APPROACHES

The main BOS analysis approaches used in pharmacometric applications are briefly summarized below. To streamline the notation and allow comparison ease, it is assumed that, without loss of generality, the original BOS variable Y has been standardized onto the closed interval [0, 1] by a linear transformation. Thus Y takes possible values in the form of k/m, where $k = 0, 1, \ldots, m$. This will result in minor notational differences compared with some of the original method descriptions.

Beta-regression

Beta-regression may be the earliest and most often-used approach in pharmacometrics, originating from psychology data analysis (3). An additional linear transformation $Y^* = Y(1 - \delta) + \delta/2$ with a small correction factor δ, e.g., 0.01, is used to map the data inside the interval [0, 1]. Y^* is then modeled with a beta distribution with density $f(x) = \Gamma(\alpha + \beta)/[\Gamma(\alpha)\Gamma(\beta)]x^{\alpha-1}(1 - x)^{\beta-1}$, where $\alpha, \beta > 0$, and Γ denotes the gamma function.

Censoring

This approach may be motivated from that of analyzing concentration data below the quantification limit (5). It treats the boundary data as outside but censored at the boundary, and the data within the boundary as continuous. The Aranda-Ordaz link function is defined by:

$$\mathrm{x} = \mathrm{h}(\mathrm{y}) = \mathrm{h}(y, \lambda) = \log\left(\frac{(1-y)^{-\lambda}-1}{\lambda}\right) \quad (1)$$

where λ is a parameter to be estimated, was used to accommodate data skewness. A general nonlinear mixed effect model $X = p + g\varepsilon$ was used where p is the model predictor, g is a residual error standard deviation function, and ε is a normally distributed residual error. The conditional likelihood of an observation y on the original scale is given by:

See related article, https://doi.org/10.1208/s12248-019-0343-9

[1] Clinical Pharmacology and Pharmacometrics, Janssen Research & Development, LLC, 1400 McKean Road, PO Box 776, Spring House, Pennsylvania 19477, USA.
[2] To whom correspondence should be addressed. (e–mail: CHu25@its.jnj.com)

$$\left[\phi\left(\frac{h(y)-p}{g}\right)J(y,\lambda)\right]^{I(y\in(0,1))}\Phi\left(\frac{x_L-p}{g}\right)^{I(y=0)}\left[1-\Phi\left(\frac{x_U-p}{g}\right)\right]^{I(y=1)} \quad (2)$$

where ϕ is the normal density, Φ is the cumulative distribution function of ϕ, I is the indicator function, $J(y, \lambda) = \partial h(y)/\partial y$ is the Jacobian, and $x_L = h(0)$ and $x_U = h(1)$ are the transformed boundary values.

Coarsened Grid

The original approach is motivated by the presumption of a latent variable U on the open interval (0, 1), the crossing certain thresholds of which leading to changes in the observed data values (1). Specifically:

$$Y = k/m \text{ if and only if } a_k \le U < a_{k+1}, \text{ for } k = 0, \ldots, m \quad (3)$$

where $a_k = (k - 0.5)/m$, $a_{k+1} = (k + 0.5)/m$, and $a_0 = 0$, $a_{m+1} = 1$. It is further assumed that U follows a logit-normal distribution, i.e., $\text{logit}(U) \sim N(p, \sigma^2)$, where p is the model predictor on the transformed scale, and σ^2 is the variance. The conditional likelihood of an observation $y = k/m$ on the original scale is given by:

$$\Phi\left(\frac{z_k^{(u)}-p}{\sigma}\right)-\Phi\left(\frac{z_k^{(l)}-p}{\sigma}\right) \quad (4)$$

where $z_k^{(l)} = \text{logit}(a_k)$, $z_k^{(u)} = \text{logit}(a_{k+1})$. This may be viewed as an ordered categorical analysis approach with the intercepts fixed up to a scale parameter.

Additional flexible transformations have been used to accommodate skewed data distributions (4), such as the Aranda-Ordaz and the Czado transformation below:

$$h(x,\lambda_1,\lambda_2) = \begin{cases} \dfrac{(x+1)^{\lambda_1}-1}{\lambda_1} & \text{if } x \ge 0 \\ -\dfrac{(-x+1)^{\lambda_2}-1}{\lambda_2} & \text{if } x < 0 \end{cases} \quad (5)$$

where λ_1 and λ_2 are parameters to be estimated. This may be viewed as an ordered categorical analysis approach with a parsimonious estimation of the intercepts (4).

It is well known that replacing the logit link in logistic regressions with the probit link leads to probit regressions. Likewise, the probit link may be used instead of the logit link with CG. That is, assuming a probit-normal distribution for the underlying latent variable leads to replacing the logit function with Φ in the definition of $z_k^{(l)}$ and $z_k^{(u)}$ in Eq. 4. As with the censoring approach, a general standard deviation function g may be used instead of σ. To the author's knowledge, this generalization has not been used in CG model applications. While it has the potential to improve the data likelihood, improving the description of data trends, e.g., assessed with the visual predictive check (VPC) (7), would seem more difficult.

Bounded Integer

This is a categorical analysis approach resembling CG without the flexible transformations. BI uses an equidistant discretization of the cumulative normal distribution with the Z-values, $Z_{1/(m+1)}$ to $Z_{m/(m+1)}$ (6). Assuming a "variance" function g as called in (6), the conditional likelihood of an observation $y = k/m$ on the original scale is modeled as follows: for $k = 1, \ldots, m - 1$,

$$\Phi\left(\frac{Z_{(k+1)/(m+1)}-p}{g}\right)-\Phi\left(\frac{Z_{k/(m+1)}-p}{g}\right) \quad (6)$$

for $k = 0$,

$$\Phi\left(\frac{Z_{1/(m+1)}-p}{g}\right) \quad (7)$$

and for $k = m$,

$$1-\Phi\left(\frac{Z_{m/(m+1)}-p}{g}\right) \quad (8)$$

Comparing Eqs. 4 and 6 shows that, when the same standard deviation function g (e.g., a constant) is used, BI and CG differ only in the intercepts used. Namely, BI uses $Z_{k/(m+1)}$ where CG uses $\text{logit}(a_k)$ or could use $\text{probit}(a_k)$. A latent variable derivation for BI was attempted ((6), supplementary materials) using the same definition of a_k, U, and Eq. 3 as given in the CG description. Note that Eq. 3 states that a_k are the latent variable thresholds, and therefore a_k directly enter the CG likelihood Eq. 4. On the other hand, a_k do not enter the BI likelihood Eqs. 6, 7, and 8, and cannot be both CG and BI latent variable thresholds under U. By construction, a_k correspond to latent variable threshold intervals centered at the observations. This property would be difficult to hold for the actual BI latent variable thresholds if they must differ from a_k. Conceptually, having the observations deviating from the center of the latent variable threshold intervals may lead to biases; however, the practical impact is not yet clear. It is noted that BI, as a newly proposed approach, is not further considered here because its performance characteristics, e.g., under the VPC (7), are not yet fully understood.

There are also a variety of other methods in the statistical literature, such as the CUB family which combines a binomial distribution with a uniform distribution (8). In the author's experience, the CUB approach may suit certain types of survey data, but the uniform distribution may not be sufficiently flexible for many clinical trial endpoints. To limit the scope, this approach will not be further discussed in this commentary, while a detailed investigation is planned in the future.

THEORETICAL COMPARISONS

An important question is how to compare the different BOS approaches, along with the standard continuous and ordered categorical approaches, i.e., the logistic/probit regressions. Likelihood-based methods, such as AIC/BIC, can seem tempting due to their wide use in model selection. However, as it recently has been noted that (9) between the continuous and

categorical approaches, the data likelihoods are not comparable; therefore, neither are the AIC and BIC values. The reason is that the data domains differ between the approaches; therefore, the data are different despite appearing to be the same. Consequently, theoretical model comparisons using AIC/BIC are only possible between the categorical approaches, namely CG, BI and the logistic/probit regressions.

The censoring approach (5) may be viewed as a mixture approach because it treats the boundary as censored but the rest of the data as continuous. Therefore, it cannot be theoretically compared with the other approaches via AIC/BIC.

The beta-regression approach (3) is also not theoretically comparable with the other approaches due to the data transformation. It is noted that, despite its intuitive appeal, the required seemingly innocuous small correction factor δ creates an ill-behavior at the boundary (2), in the sense that the boundary data can be shown to become arbitrarily influential as $\delta \to 0$. Therefore, the approach lacks statistical rigor (9).

PRACTICAL COMPARISONS

Although some approaches cannot be theoretically compared, choices still need to be made in practice. An important factor is the consistency between the observed and model-predicted data, as assessed by VPC. Suitable comparison metrics will need to be chosen depending on the intended use of the model. Many metrics may seem reasonable, e.g., correct classification rates or mean prediction errors on the continuous scale. However, as noted in the above section, the very reason that continuous and categorical approaches are not comparable is that they treat the data on different scales. This implies that there can be no universally best metrics. For example, correct classification rates would in principle favor the categorical approaches, and mean prediction errors would favor the standard continuous approach with a normally distributed residual error.

Although universally best metrics may not exist, the intended use of the model may determine the metrics suitable for the specific application. In clinical trials, certain derived endpoints could be the main objective, e.g., that of modeling PASI scores is often to predict the rates of PASI75/90/100, which are 75%, 90%, and 100% improvement from baseline, respectively (2); the corresponding correct classification rates would then be the natural metrics in this scenario. This may be practically conveniently implemented as VPCs of PASI75/90/100 (2). In other situations, the intended use of the model may be to describe the data distribution on the continuous scale; then, VPC on the continuous scale could be used (4).

In situations where the intended use of the model may yet to be specified, an argument can be made for using the more informative scale, i.e., when comparing continuous and categorical approaches, the data simulated under the categorical model should be treated as continuous. While this may favor the continuous model, the opposite would cause more issues as the mapping from the continuous scale to the categorical scale is not unique in this case. This may affect the claim one wishes to make. For example, in a recent application, a categorical approach has been argued to perform better than a continuous approach because the categorical approach achieved better VPC results even on the continuous scale (9).

The diverse situations make it difficult to systematically investigate the best performing approach. Simulation evaluations would likely favor the approach closer to the data-generating model. Practical complexities also vary; for example, in the case of analyzing PASI scores, satisfactory predictions of PASI75/90/100 could not be achieved (2), despite the extensive efforts with a variety of methods in multiple applications (Matthew M. Hutmacher, personal communication).

Even when formal comparisons are possible, ensuring the consistency between the observed and model-predicted trends may still be important, especially for the categorical approaches. In the author's experience with the standard ordered categorical data analyses, the data usually may support only a between-subject variability (BSV) effect on the intercept (10); with additional BSVs, due to the dependence of the model-predicted response probability on both the predictor mean and variance, the likelihood may often improve significantly but the VPC may still show severe biases.

EXPECTED MODEL PERFORMANCE

As comparing many methods would be practically cumbersome, it is desirable to have an initial understanding of the likely performance of the analysis methods based on their characteristics and the application scenarios. The remainder of this section assumes, somewhat generally, that the main modeling objective is to adequately describe the data distribution on the continuous scale.

Although predicting the data with its specified range is appealing, its practical relevance usually reduces as the number of possible categories increases, which makes the data closer to continuous in nature. Therefore, the standard continuous approach assuming a normally distributed residual error may perform adequately when the number of possible categories is large and the data distributions are approximately symmetric. It is noted that data distributions may be skewed for different reasons, including the presence of BSVs and predictor-dependent variances. These create difficulties for the standard continuous approach, as sizable BSVs may result in model predictions falling outside the boundary. The standard continuous approach also has difficulties to handle predictor-dependent variances because the boundary values may be achieved: for example, additive-plus-proportional error models become ill-behaved at 0 with maximum likelihoods approaching infinity (2).

The censoring approach may better handle skewed data by using transformations. The categorical approaches could also handle predictor-dependent variances to some extent because boundary violations are not possible, and the variances are automatically linked with the predictions. This may be seen from the fact that the variance of a binomial distribution is larger when the parameter p is closer to the center (0.5). Thus, the CG approach without transformation (1) could handle skewness to a minor extent, and more skewness will require additional transformations (4).

Recently, the standard ordered categorical approach has been demonstrated to perform well with sufficient data (11). With the many ($=m$) intercept parameters to describe every category, the approach is more accurate and robust compared with the standard continuous approach (9), and likely with the other approaches as well.

A PRACTICAL GUIDELINE

Considering the expected model performance given above, the proposed guideline below could be used to select the analysis method.

- Standard continuous: if data distributions appear symmetric and the number of categories is large (e.g., > 10);
- CG with no additional transformations: if data distributions appear symmetric and the number of categories is moderate (e.g., 8–10);
- Censoring and/or CG with flexible transformations: if data distributions appear skewed and the sample size is moderate relative to the number of categories;
- Standard ordered categorical: if data distributions appear skewed, and the sample size is large relative to the number of categories.

The term "sample size" here relates to the total number of observations as well as the minimum of the number of observations in each category. The aim of the guideline is to provide a start. VPC is recommended to evaluate model performance, and additional methods should be considered if the results are unsatisfactory.

SUMMARY

BOS analysis methods are complex and evolving. Different methods may suit different scenarios. Intuitions on simultaneously dealing with the continuous and categorical aspects of the data could be tempting but misleading, and the difficulty of theoretically comparing continuous and categorical approaches further complicates the situation. This makes the appropriate selection and evaluation of analysis methods important, especially the consistency between the observed and model-predicted data trends. Further methodological development to better extract information from the data is also necessary.

REFERENCES

1. Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcome scores. Biostatistics. 2007;8(1):72–85.
2. Hu C, Randazzo B, Sharma A, Zhou H. Improvement in latent variable indirect response modeling of multiple categorical clinical endpoints: application to modeling of guselkumab treatment effects in psoriatic patients. J Pharmacokinet Pharmacodyn. 2017;44(5):437–48.
3. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol Methods. 2006;11(1):54–71.
4. Hu C, Yeilding N, Davis HM, Zhou H. Bounded outcome score modeling: application to treating psoriasis with ustekinumab. J Pharmacokinet Pharmacodyn. 2011;38(4):497–517.
5. Hutmacher MM, French JL, Krishnaswami S, Menon S. Estimating transformations for repeated measures modeling of continuous bounded outcome data. Stat Med. 2011;30(9):935–49.
6. Wellhagen GJ, Kjellsson MC, Karlsson MO. A bounded integer model for rating and composite scale data. AAPS J. 2019;21(4):74.
7. Karlsson MO, Holford NHG. A tutorial on visual predictive checks 2008 [updated 2008. Available from: www.page-meeting.org/?abstract=1434. Accessed 9 Aug 2019.
8. Piccolo D, Simone R, Iannario M. Cumulative and CUB models for rating data: a comparative analysis. Int Stat Rev. 2018;0(0):1–30.
9. Hu C, Adedokun OJ, Zhang L, Sharma A, Zhou H. Modeling near-continuous clinical endpoint as categorical: application to longitudinal exposure-response modeling of Mayo scores for golimumab in patients with ulcerative colitis. J Pharmacokinet Pharmacodyn. 2018;45(6):803–16.
10. Hu C. Exposure-response modeling of clinical end points using latent variable indirect response models. CPT Pharmacometrics Syst Pharmacol. 2014;3:e117.
11. Liu Q, Shepherd BE, Li C, Harrell FE Jr. Modeling continuous response variables using ordinal regression. Stat Med. 2017;36(27):4316–35.