MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Assessing the impact of extreme values in clinical studies–a latent variable approach

Abstract

Large power losses were observed in two well-powered placebo-controlled clinical trials as the primary efficacy endpoint excessively declined due to rapid disease progression. Despite the common practice of applying a parametric model on rank transformed data to address the issue of extreme data due to the robustness of ranks less affected by extreme values, there is no clear consensus among practitioners or in published work regarding the conditions under which a rank transformed analysis should be performed. The question remains: in practice, at what point do data become so extreme that parametric tests stop being effective and nonparametric tests become necessary. This paper aims to raise awareness of non-normal data with extreme values in drug development. We evaluate the impact of non-normality and extreme values on statistical performance through the approach of a latent variable model and provide a framework to identify and handle efficacy data with extreme values through the evaluation of the Jarque Bera test and Kurtosis.

Keywords: Rapid disease progression, extreme values, kurtosis, latent variable model, power loss

Volume 8 Issue 3 - 2019

Zhengning Lin

Biometrics, Ascendis Pharma Inc., USA

Correspondence: Zhengning Lin, Ascendis Pharma Inc, 500 Emerson Street, Palo Alto, CA 94301, USA, Tel 4158405017, Email zln@ascandispharma.com

Received: May 08, 2019 | **Published:** May 16, 2019

Introduction

Case studies

Assessment of change in exercise capacity using the 6-minute walk distance (6MWD) tests has been the primary endpoint in many pulmonary arterial hypertension (PAH) and neuromuscular disorder clinical trials. However, large power losses were observed in the primary endpoint of 6MWD analysis in well-powered placebo-controlled studies. One study is a new drug application (NDA) of drisapersen in 2015. The drisapersen NDA included 3 placebo-controlled studies to demonstrate efficacy for Duchenne Muscular Dystrophy, a rare progressive neuromuscular disorder that is ultimately fatal for boys at a young age. Change in 6MWD is the primary endpoint. While the two smaller proof-of-concept pilot studies showed consistent treatment differences, the larger and only well-powered placebo-controlled study failed to detect a treatment difference. The statistical power of the pre-planned primary analysis was reduced from the planned 90% to only 53% as a result of the increased standard deviation from the planned 55 meters to the actual 87 meters, based on the parametric model of mixed model repeated measurement (MMRM) assuming normal data distribution.

Examination of the data revealed the root cause for reduction: The larger study enrolled a broader population with much higher variability due to excessive declines in 6MWD for a portion of patients with rapid disease progression, including those who lost ambulation during the study. In such cases 0 meters were imputed for their missing 6MWD values due to loss of ambulation. Excessive declines were also observed on patients who were still ambulant during the one year treatment period.

It turns out such phenomenon is not unique to 6MWD measures. A second case study is a pirfenidone NDA in 2014 for the treatment of idiopathic pulmonary fibrosis (IPF), a rare and ultimately fatal lung disease. The median survival of patients with IPF is only 2 to 3 years, although some live much longer. Respiratory failure resulting from disease progression is the most frequent cause of death. The primary endpoint is change from baseline in percent predicted forced vital capacity (FVC % pred). During the study, any patients who died before the end of study were recorded as missing FVC % pred, with a 0 imputed during analysis. The clinical course of individual patients varies from slow progression to acute decompensation and death, resulting in highly non-normal efficacy data with a skewed heavy tail. The usage of log-transformations, a technique frequently used in pharmacology studies, is prohibited due to the ties in 0. The sensitivity analysis showed that the statistical power would have been reduced by over 20% from the planned 95% based on a parametric model during the final analysis.

Root cause for power loss

Clinical trials for progressive diseases often observe excessive declines in efficacy endpoints due to rapid disease progression including deaths or loss of capability to perform the efficacy assessment. Such observations can disproportionally influence the mean statistic so that it becomes unstable, causing a large reduction in statistical efficiency. The large impact of extreme values is not well recognized by clinical statisticians working in drug development, partly because they may not appear as outliers that statisticians are trained to detect, and the data may not appear to be so extreme based solely on its values.

While this may seem like an archetypal problem with a well-established solution on the surface, applying a usual parametric model on rank transformed data as a general nonparametric method for non-normal data;[1] it is not clear how extreme is extreme in practice that the parametric model stops being effective, and a nonparametric test becomes necessary. The paper sets out to examine this.

Literature review and current practice

While non-normal data do not necessarily invalidate a parametric analysis, there is no consistent and simple guidance in publications about what kind of non-normal data require distribution-free or nonparametric methods. Text books of nonparametric analysis usually recommend using nonparametric methods when data are not normally distributed.[2,3] Hollander & Wolfe[3] point out that usually

the nonparametric procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal, and they can be mildly or wildly more efficient than these competitors when the underlying populations are not normal. Khan & Rayner[4] evaluated the robustness of analysis of variance (ANOVA, parametric) and Kruskal-Wallis (non-parametric) tests with simulations from g-and-k distributions.[4] One of their conclusions is that the Kruskal-Wallis test clearly performs better than the ANOVA test if the sample sizes are large and the distribution kurtosis is high. They also concluded that skewness has a much smaller effect than kurtosis on statistical power. Simulation studies generally have similar conclusions: when distributions consist of heavier tails or extreme values, the nonparametric methods perform better.[4–6] In contrast, Rasch & Guiard[7] concluded (in psychological research) that in most practical cases the parametric approach for inferences about means is so robust that it can be recommended in nearly all applications. Vickers[8] concluded that the parametric analysis of covariance (ANCOVA) model is the preferred method of analyzing randomized trials with baseline and post-treatment measures compared to the nonparametric Mann-Whitney test for non-normal distributed data.

The apparent differences in conclusions reflect complications of statistical performance evaluation under different scenarios. The comparisons usually rely on simulations of limited scenarios which may not always be generalizable to scenarios practitioners are dealing with. Statisticians in drug development are often left with an impression that a nonparametric analysis may not be worthwhile for their studies, or that there are no clear conditions under which parametric methods are not appropriate and a nonparametric method is necessary. Many statisticians believe that, due to the large sample central limit theorem, a parametric model should perform as well or better than a nonparametric method if the sample size is not small, even if the data are not normal. Parametric methods derived from the normality assumption such as the t-test, ANCOVA, mixed model repeated measurement (MMRM) and others are routinely used without critically evaluating for model appropriateness. The lack of awareness was well-revealed during discussions with statistical experts working in the field of drug development and by on-line searching of information comparing the pros and cons of using parametric method vs a distribution free (nonparametric) method. For instance, posts published on Minitab's blog contained statements that parametric methods are robust with larger statistical power compared to their nonparametric counterparts even when data are not normal.

The purpose of this paper is to raise awareness of non-normal data with extreme values and to evaluate their impact on statistical performance through a latent variable model. It also recommends a method for identifying and handling data with extreme values.

Evaluating method

Without loss of generality, consider the typical situation of hypothesis testing comparing two treatment groups, with the null hypothesis of no treatment difference at the two-sided alpha level of 0.05. Computer simulations were conducted to compare the statistical performance of the parametric analysis assuming normality and its nonparametric counterpart using the same parametric analysis on rank transformed data.[1] Early statistical findings suggest that both the parametric and the nonparametric methods are robust against violation of normality in type I error control.[9] This statistical performance evaluation therefore focuses on the statistical power with various scenarios of non-normal data with extreme values. The parametric analysis is represented by a simple one-way analysis of variance (ANOVA) model with treatment group as the model factor in the simulation. The nonparametric counterpart is the same ANOVA on rank transformed data, virtually equivalent to the Kruskal-Wallis test (or Wilcoxon Rank Sum Test for two groups).[4] Statistical significance is calculated with one-sided $p < 0.025$, which is typical for efficacy claims that the testing drug is superior to the control with an equivalent two-sided $p < 0.05$.

The extreme data issue in the simple case of one-way ANOVA is equally applicable to more complicated design settings. Generalization to other parametric models is discussed in Section 4.

A latent variable approach

Assume that the true disease status can be expressed by a latent variable X following the standard normal distribution, and the clinical endpoint Y is linked with X through a hidden monotonic transformation function $Y = f(X)$. This setting is applicable to clinical studies in general, because for any random variable Y there exists a monotonic function $Y = f(X)$ with X following the standard normal distribution. In fact, for Y with a cumulative density function (CDF) of $G(.), X = \varphi^{-1}(G(Y))$ that follows the standard normal distribution, where $\varphi(.)$ denotes the CDF of the standard normal distribution, and $\varphi^{-1}(.)$ denotes an inverse function of $\varphi(x)$, or quantiles. Therefore, $f(x) = G^{-1}(\varphi(x))$ transforms the variable X with CDF $\varphi(.)$ to Y with CDF $G(.)$, where $G^{-1}(.)$ denotes a generalized inverse function of $G(.)$. Note that in this setting $f(x)$ is not necessarily a one-to-one mapping function, while the examples in this paper are all one-to-one monotonic functions. This model allows simulation of non-normal data with different $f(x)$ to evaluate the impact of any non-normal data in the clinical endpoint Y for the same latent endpoint X. It enables an intuitive comparison of the actual power with a non-normal clinical endpoint Y, and the target power with the underlying treatment effect expressed by a latent endpoint X with normal data distribution.

Let X_t and X_c denote the latent response variables of the testing and control groups, respectively, in a randomized clinical trial with a normal distribution and a common standard deviation. Without loss of generality, assume that the mean X_t and X_c are μ and 0, respectively, with a common standard deviation of 1. In this setting, μ represents the standardized treatment effect of the testing group compared to the control group in the scale of the latent variable X under the condition of equal variance. Table 1 summarizes the required standardized effect size μ for sample sizes of 25, 50, 100, and 1000 per group for statistical power of 80% and 90% with normal distribution. The smaller μ's with larger sample sizes reflect the reality in clinical study design that, with the same statistical power, studies with smaller standardized effect sizes require larger sample sizes. The very large sample size case of 1000 per arm is included to evaluate the effect of large samples on statistical performance. The simulation used the standardized treatment effect μ and the corresponding sample size in Table 1 to generate samples of X_t and X_c from normal distributions with a common standard deviation of 1. A single transformation function $f(x)$ was applied to both X_t and X_c with $f(x)$ representing the hidden relation between the latent variable X and the clinical variable Y. Different $f(x)$'s generate different scenarios of non-normal data. In particular, three $f(x)$'s with one-to-one monotonic mapping were used to generate three scenarios of extreme values: (1) The exponential function $f(x) = e^x$ (resulting in a lognormal distribution); (2) the 3rd power function $f(x) = x^3$; and (3) the 5th power function $f(x) = x^5$.

In addition, some common data distributions for the control group were simulated: the exponential distribution with mean of 1 represents data with extreme values or high kurtosis (excess kurtosis

Citation: Lin Z. Assessing the impact of extreme values in clinical studies—a latent variable approach. *Biom Biostat Int J.* 2019;8(3):85–91.
DOI: 10.15406/bbij.2019.08.00276

= 6); the uniform (0,1) distribution represents data without a long tail or with low kurtosis (excess kurtosis = −1.2); and the standard normal distribution (excess kurtosis = 0) acts as a reference. In this case $f(x) = G^{-1}(\varphi(x))$ with $G(x)$ the CDF of the control group and $\varphi\{\varphi^{-1}[G(x)]-\mu\}$ the CDF for the active group. Statistical powers are compared by simulation between the parametric ANOVA and the non-parametric rank transformed ANOVA.

Table 1 Standardized treatment effect μ for various sample sizes and power under normal data distribution

	25 per group	50 per group	100 per group	1000 per group
80% power	0.809	0.566	0.398	0.125
90% power	0.936	0.655	0.461	0.145

A data dependent model selection approach

As a good statistical practice, analysis model assumptions, including the normality assumption, should be evaluated for model appropriateness once a study is completed for data analysis. Model selection based on the outcomes of model checking using the same data has the potential of inflating type I error. This simulation work therefore evaluates the type I error as well as the statistical power of a data dependent model selection process as described in the following paragraphs.

A data dependent model selection process requires a model residual analysis for appropriateness of the parametric model. Such a process may include pre-specified decision rules to avoid the potential of model selection bias. Relying on a simple test of normality to select the method can be problematic for the following reasons: Because no data are strictly normal in practice, the normality assumption will be rejected with large enough sample size for even minor deviations from normality. Furthermore, non-normal distribution without extreme values or with minor deviation may not have a major impact on parametric models even if the data are clearly not normal. For example, rank transformed data asymptotically follow a uniform distribution with low kurtosis.[10] The normality assumption will be rejected for ranked data with moderately large sample size even though running the usual parametric data on the ranked data is an established non-parametric method. The condition to use a rank-based nonparametric method therefore should include a criterion for extreme values that cause statistical performance issues.

The pre-specified model selection criteria in this simulation include a combination of hypothesis testing of normality and a sample kurtosis evaluation for extreme values. Specifically, the rank-based ANOVA is selected if the following two conditions are both met for the model residuals:

I. The Jaque Bera test for normality is rejected significantly at the significance level of 0.05.

II. The model residual excess kurtosis is greater than 1.

The reason for selecting the Jarque Bera test for normality is to ensure evidence exists that the clinical data are not normal due to high kurtosis. Testing for kurtosis only may be problematic.[11]

Researchers have suggested that holding variance approximately constant, a higher kurtosis implies that there are more extreme observations observed in relation to a given distribution, or that the extreme observations are more deviant. Simulations are shown below under both Normal and Poisson distributions. Variance and the proportion of extreme values are held approximately constant; the X axis measures the extremity of extreme observations. The sample Kurtosis exhibits a consistently upward trend as X increases. Hence the sample kurtosis seems a valid measure for quantifying extreme values, including outliers (Figure 1&2).[10]

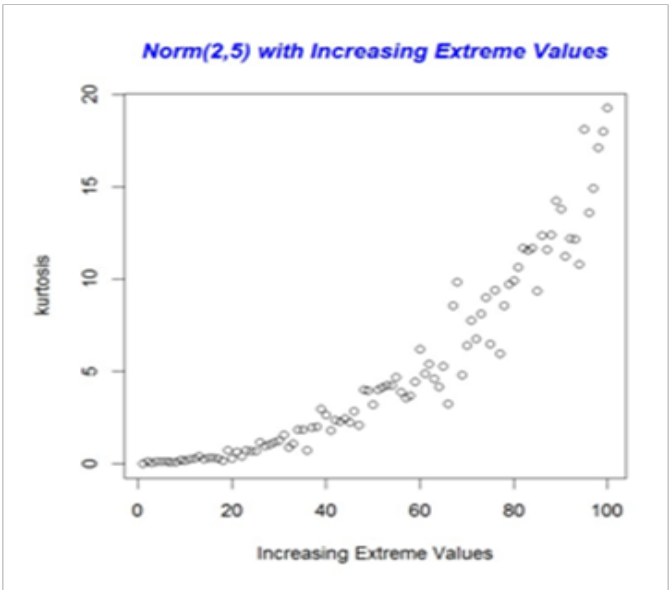

Figure 1 Normal Figure.

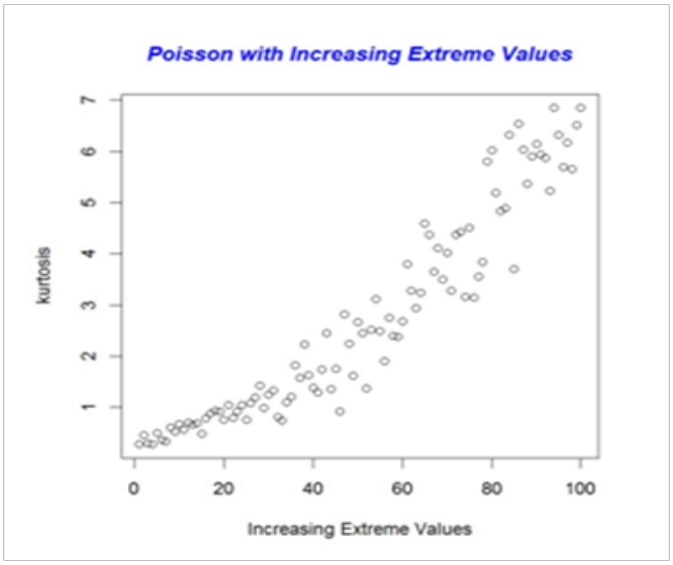

Figure 2 Poisson.

Simulation results

All simulations were performed in R and each of the power and type I error calculations were based on 100,000 replications of

Citation: Lin Z. Assessing the impact of extreme values in clinical studies—a latent variable approach. *Biom Biostat Int J.* 2019;8(3):85–91.
DOI: 10.15406/bbij.2019.08.00276

randomly simulated studies. Power comparisons of the parametric ANOVA and the nonparametric rank ANOVA are included in Table 2 with the underlying treatment effect of the latent variable corresponding to 80% power as summarized in Table 1. Because the transformation functions are one-to-one and monotonic, data rankings are unchanged. Consequently, the power of the rank ANOVA is unaffected between different data scenarios determined by different transformation functions $f(x)$. The power of the rank ANOVA was around 78%, with minor differences across all six scenarios reflecting random errors from each set of the 100,000 study simulations, and with random standard error of 0.13% as the calculation precision. For all three scenarios of extreme values, the parametric approach had a power of 38-65%. This is a decrease of 19-52% compared to the target 80% under a normal distribution. This reduction in power is substantially more than that of the nonparametric approach which had only a 2% decrease. The parametric ANOVA also had substantial power loss of around 10% for the exponential distribution. The two approaches had similar power for the uniform distribution, while for the perfect normal data the parametric model had around 2% power advantage. These results are generally consistent across sample sizes, with a minor tendency of more power loss for the parametric model with the smallest sample size of 25 per arm. It is worth noting that a much larger sample size of 1000 per arm did not help reduce the power loss of the parametric model relative to the target power under normal distribution.

Table 2 Statistical power (%) under different distribution scenarios with μ corresponding to 80% power with normal distribution as in Table 1

Distribution	N = 25+25		N = 50+50		N = 100+100		N = 1000+1000	
	(μ=0.809)		(μ=0.566)		(μ=0.398)		(μ=0.125)	
Scenario	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA
$y = e^x$	60.1	77.9	61	78.5	60.1	77.9	57.5	78.1
$y = x^3$	57.7	77.9	59.3	78	59.5	78.2	58.3	78.1
$y = x^5$	28.1	77.9	32.5	78.1	33.8	77.9	30	77.7
Exponential	69.3	77.9	70.5	78.3	70.5	78	71	77.9
Uniform	77.9	78	77.9	78	78	78	78	78
Normal	80	78	79.9	78.2	79.8	77.9	79.8	77.9

Results corresponding to 90% target power are summarized in Table 3 with similar outcomes. The parametric ANOVA lost power from 17% to 54% compared to the target 90% for all three transformation scenarios of extreme values, while the nonparametric method had only around 1.5% lower power than the target 90%. The power loss for the parametric model was between 7-9% for the exponential distribution. For the uniform distribution the two approaches maintained similar powers with around 1.5% lower power than the 90% target. It is obvious from the above multiple scenarios of extreme values that the parametric model is unacceptably inefficient with high volume of power loss compared to the target power with normal data, while the nonparametric counterpart maintained the power well with only 1-2% power loss compared to the ideal case of normal data. This conclusion is consistent from a wide range of sample sizes from 25 to 1000 per arm which cover most of those for confirmatory studies with continues variables. Therefore, the parametric model is not appropriate under such scenarios of non-normal data and the nonparametric model should be used instead. These results are consistent with those results comparing the Wilcoxon and t-test in their asymptotic relative efficiency or Pitman efficiency.[12,13]

Table 3 Statistical power (%) under different distribution scenarios with μ corresponding to 90% power with normal distribution as in Table 1

Distribution	N = 25+25		N = 50+50		N = 100+100		N = 1000+1000	
	(μ=0.936)		(μ=0.655)		(μ=0.461)		(μ=0.145)	
Scenario	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA	ANOVA	Rank ANOVA
$y = e^x$	73.2	88.6	73.4	88.6	72.6	88.7	70.2	88.6
$y = x^3$	70	88.4	71.8	88.7	71.9	88.5	71.2	88.6
$y = x^5$	36.4	88.4	41.7	88.7	43.2	88.7	38.8	88.7
Exponential	80.7	88.5	82	88.7	82.4	88.7	82.9	88.6
Uniform	88.1	88.5	88.4	88.7	88.6	88.7	88.4	88.4
Normal	90	88.6	90	88.6	90.1	88.7	90	88.6

Citation: Lin Z. Assessing the impact of extreme values in clinical studies−a latent variable approach. *Biom Biostat Int J.* 2019;8(3):85–91. DOI: 10.15406/bbij.2019.08.00276

Results of the data dependent model selection approach are included in Table 4, with both power and type I error outcomes. For all the cases the data dependent approach maintained the power well with only 1-2% lower power compared to the ideal case of normal data, with acceptable type I error close to the target 0.025 level. Furthermore, the statistical power for distributions with extreme values (or high kurtosis) are very close to those based on rank ANOVA as shown in Table 2, while the power for normal distribution was maintained at 80%.

Table 4 Statistical power (%) and type I error (one sided) for the data dependent model selection approach under different distribution scenarios with μ corresponding to 80% power with normal distribution as in Table 1

Distribution	N = 25+25		N = 50+50		N = 100+100		N = 1000+1000	
	(μ= 0.809)		(μ= 0.566)		(μ=0.398)		(μ=0.125)	
Scenario	Power	Type I	Power	Type I	Power	Type I	Power	Type I
	(%)	error	(%)	error	(%)	error	(%)	error
$y = e^x$	77.7	0.0255	78.4	0.025	77.9	0.025	78.1	0.0246
$y = x^3$	77.9	0.0252	78	0.0253	78.2	0.0247	78.2	0.0247
$y = x^5$	77.9	0.0245	78.1	0.0245	77.9	0.0259	77.7	0.0247
Exponential	77.4	0.0269	78.1	0.0249	78	0.0249	77.9	0.0261
Uniform	77.9	0.0265	77.9	0.0247	78	0.0249	78	0.0248
Normal	80.1	0.0257	80	0.0251	79.8	0.0247	79.8	0.0253

Discussion

Regarding the latent variable model

As illustrated, the latent variable model provides a helpful framework to evaluate the impact of non-normal data. It interprets clinical endpoints with non-normal distribution as transformed from a latent variable with normal distribution so that statistical performance of an analysis method can be compared directly between a non-normal distribution and a normal distribution. It makes the statistical efficiency issue intuitive for clinical statisticians and other practitioners. Simulation outcomes with the model clearly demonstrated that non-normal data with extreme values are not appropriate for parametric models focusing on the mean statistic regardless of sample sizes.

Unlike other simulation work, which evaluates power and type I error for a constant treatment effect of a clinical endpoint with non-normal distribution, the latent variable model assumes a constant treatment effect for the latent variable with normal distribution. This setting makes the treatment effect for the clinical endpoint not constant in general. Although one cannot be sure if one model is more realistic than the other in a particular case, assuming a constant effect for a variable with normal distribution is perhaps as reasonable, if not more reasonable, than for a variable with non-normal distribution. Although this model does not assume constant effects for non-normal clinical endpoints, the constant effect is only a commonly used assumption for model simplicity rather than reality. This issue does not affect the validity of hypothesis testing as the constant effect assumption is correct for both the latent and clinical endpoints under the null hypothesis of no treatment difference. The treatment effect interpretation in the mean statistic can be problematic for non-normal data with extreme values, and in this case other statistics such as median, responder rate, quintiles, CDFs etc. can be more appropriate.

In addition, the latent variable model is applicable in a general setting, including datasets with ties or 0s. For example, even if there are ties in an efficacy variable of interest, there could be additional variables to differentiate the ties so that the values are different for the latent variable. In the pirfenidone program, we ranked patients who died earlier worse than those died later, which is an example of breaking the ties. The presence of 0s affects PK log transformation but does not affect the latent variable which is normally distributed. The rank transformation can handle the ties as we applied in the pirfenidone studies. The latent variable model does not require the assumption that the data are transformed into a normal distribution. In this regard, the latent variable has an advantage compared to a specified transformation which may have limitations, like the log transformation.

Method selection

Not all nonparametric methods are suitable to handle extreme values. For example, the permutation test is a common nonparametric method which is distribution-free to handle non-normal data. However, it does not address the performance issue of the mean statistic with extreme values if the mean statistic is the basic statistic for the permutation. Therefore, a nonparametric permutation test with the mean statistic suffers the same performance issue as a parametric model when the data have extreme values. A rank based nonparametric model is recommended to handle the excessive influence of extreme values. The model details should be specified before database lock to avoid the potential of selection bias, as was applied to the pirfenidone NDA. The decision rule of model selection should also be specified before study unblinding to avoid selection bias, as was illustrated by the data dependent approach in the simulation work of this paper.

When a decision rule is not clearly specified while the data are clearly not appropriate for a parametric method, a rank based nonparametric method may be applied and the potential of model selection bias should be addressed. To minimize the potential of method selection bias, applying standardized rank transformation before applying the pre-specified parametric model may be a reasonable default choice based on its high statistical performance and model similarity. In the drisapersen Phase III study, the normality test rejects the hypothesis at $p < 0.0001$ for the model residuals with sample excess kurtosis of 2.2

Citation: Lin Z. Assessing the impact of extreme values in clinical studies—a latent variable approach. *Biom Biostat Int J.* 2019;8(3):85–91.
DOI: 10.15406/bbij.2019.08.00276

and skewness of -0.9. The data are clearly not suitable for parametric models with an unacceptable power loss of close to 40% due to much increased data variability. Applying standardized rank transformation before running the MMRM model is therefore appropriate.

Regarding estimation of treatment effect

This paper focuses on the statistical efficiency evaluation from hypothesis testing, rather than from statistical estimation perspective. However, treatment effect estimation is equally affected by extreme data, with unstable mean statistics and much enlarged standard errors.

Treatment effect estimation on the original variable cannot directly be derived from analysis models based on rank transformed data, since a rank transformation is not a one-to-one transformation which can be back-transformed without loss of information. As pointed out earlier, the mean statistic may not be appropriate for non-normal data with extreme values and alternative statistic may be more representative of the central tendency of treatment effect. Robust statistics such as the median (or the Hodges-Lehmann version for treatment difference[14] can be used to estimate central tendencies when the mean statistic is much affected by extreme values. In practice, a combination of the mean and median statistics works well in assessing the central tendency in most cases, although other robust statistics such as trimmed mean can have better statistical performance[2] although not as commonly used in practice.

Generalization and limitation

This simulation work compares a simple ANOVA and its rank transformed nonparametric counterpart with two treatment groups with equal sample size to illustrate the statistical performance issue of parametric methods for data with extreme values. The conclusions based on these simulations are expected to be generalizable to scenarios with more than two treatment groups and/or with different sample sizes among treatment arms. This is because typically pairwise treatment comparisons are still the focus for studies with multiple treatment arms, and the role rank transformation plays to address the distribution issues remains very similar to that in a study with two treatment arms and with equal sample size.

Because the performance issue is related to the mean statistic of parametric methods, it is reasonable to expect that the conclusions are also generalizable to more complicated parametric models such as ANCOVA and MMRM models since they are also mean statistic based models. The generalization from ANOVA to ANCOVA is intuitive and straightforward if we consider those distribution assumptions of the simulations as those after covariate adjustments in the ANCOVA model. In such applications with subgroup stratification and/or repeated measurements, the rank transformation should be standardized to be on a comparable scale among subgroups and visits. A common rank standardization is to divide the ranks with $(n + 1)$, where n is the sample size for a subgroup at a visit, so that they are in the range of 0 to 1 for all subgroups and visits. The standardized ranking also mitigates the issue of unequal data variations across subgroups, as encountered in the drisapersen Phase III study. The drisapersen study includes a subgroup of patients that are older and more impaired at baseline with much larger functional declines and data variations than those in other subgroups. Stratified nonparametric analysis using standardized ranking addresses both the extreme value issue and the unequal variance issue in the study. It is therefore appropriate for the study. For the case of ANCOVA it often makes sense to rank the covariates so that not only the potential extreme value issues for the covariates are addressed, but also, they are on a relatively comparable distribution with the rank transformed endpoints.

This research work focuses on the issue of extreme values for the application of comparing randomized treatment groups which is typical in drug development. The simulation work is based on the simple case of comparing two independent treatment groups without verifying the conclusions in more general scenarios. While it is reasonable to expect similar conclusions for studies with more than two treatment arms and/or with covariates and strata since the issue of extreme values and the solutions discussed above are very similar, we do not provide such simulation work in this paper partly to focus on the main points without a very lengthy description and discussion for a large variety of scenarios to simulate. On the other hand, the rank transformation strategy discussed above is applicable only to evaluate most typical clinical study designs. A study with more complicated design requires a different rank transformation strategy suitable for the study design, or perhaps a different nonparametric method which is beyond the scope of this paper.

The simulation work in this paper is limited to selected examples of distributions with or without extreme values, and therefore may not be applicable to a study with a very different data distribution or study design. This is because our focus is to raise the awareness of the issues with extreme values, rather than to cover for all scenarios. The latent variable model provides a framework to evaluate different studies with special simulations suitable for their studies. For the same reasons, we do not simulate for sample sizes smaller than 25 per arm where the large sample central limit theorem may not work effectively, and exact methods such as permutation (instead of a parametric model) on rank transformed data may be applied for the calculation of p-values.

Conclusion

The latent variable model provides a useful framework to intuitively evaluate the effects on statistical performance of different non-normal data distributions compared to the normal distribution with an underlying treatment effect on a latent variable. It is demonstrated that the parametric ANOVA model can be unacceptably inefficient and is not appropriate for data with extreme values. The nonparametric counterpart with a (standardized) rank transformation maintains the statistical power well compared to the target power for the latent normal variable. The data dependent model selection process maintains the statistical power well with acceptable type I error. It is recommended that a similar data dependent model selection process be pre-specified to avoid model selection bias.

Acknowledgments

The author would like to thank Dr. Ange Zhou for her valuable contributions of this research work.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for research, authorship, and/or publication of this article.

References

1. Conover WJ, Iman RL. Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician*. 1981;35(3):124129.
2. Conover WJ. *Practical Nonparametric Statistics*. 3rd edn. USA: John Wiley & Sons; 1999.
3. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. 2nd edn. USA: John Wiley & Sons; 1999.
4. Khan A, Rayner GD. Robustness to non–normality of common tests for the many–sample location problem. *Journal of Applied Mathematics and Decision Sciences*. 2003;7(4):187–206.
5. Zimmerman DW, Zumbo BD. Effect of outliers on the relative power of parametric and nonparametric statistical tests. *Perceptual and Motor Skills*. 1990;71:339–349.
6. Sawilowsky SS, Blair RC. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*. 1992;111(2):352–360.
7. Rasch D, Guiard V. The robustness of parametric statistical methods. *Psychology Science*. 2004;46(2):175–208.
8. Vickers AJ. Parametric versus non–parametric statistics in the analysis of randomized trials with non–normally distributed data. *BMC Medical Research Methodology*. 2005;5(35):1–12.
9. Wiedermann W, Eye AV. Robustness and power of the parametric t test and the nonparametric Wilcoxon test under non–independence of observations. *Psychological Test and Assessment Modeling*. 2013;55(1):39–61.
10. DeCarlo LT. On the meaning and use of kurtosis. *Psychological Methods*. 1997;2(3):297–307.
11. Bai J, Ng S. Tests for skewness, kurtosis and normality for time series data. *ASA Journal of Business & Economic Statistics*. 2005;23(1):49–60.
12. Witting H. A Generalized Pitman Efficiency for Nonparametric Tests. *Ann Math Statist*. 1960;31(2):405–414.
13. Nikitin YY. Asymptotic Relative Efficiency in Testing. *Encyclopedia of Mathematics*. 2010.
14. Rosenbaum PR. Hodges–Lehmann Point Estimates of Treatment Effect in Observational Studies. *Journal of the American Statistical Association*. 1993;88(424):1250–1253.

Citation: Lin Z. Assessing the impact of extreme values in clinical studies−a latent variable approach. *Biom Biostat Int J.* 2019;8(3):85–91.
DOI: 10.15406/bbij.2019.08.00276