Reference Interval Determination by Bhattacharya Analysis on Skewed Distributions – Problems and Pitfalls.

Graham RD Jones
Department of Chemical Pathology, St Vincent's Hospital, Sydney. gjones@stvincents.com.au

Introduction and Aim

Bhattacharya analysis is a useful method for isolating Gaussian distributions in the presence of other data (1) and has been used for reference interval determination using pathology databases (2).

Biological analytes are often non-Gaussian in their distributions. Examples include serum ALT, CK, proBNP and amlyase. These skewed distributions are sometimes commonly log-transformed to allow the use of parametric statistics or Bhattacharya analysis.

The distributions of some analytes have been shown to be skewed to a greater or lesser extent than can accurately be described by log transformation (3).

Assumptions about the underlying analyte distribution in healthy populations may influence the process of estimating reference intervals using the Bhattacharya was evaluated.

Aim

The aim of the study is the examine the effect of assumptions concerning the underlying distribution of serum ALT on reference intervals derived using Bhattacharya analysis using transformed data.

Methods

Data sets for serum ALT (males only) were obtained for healthy individuals from the NORIP and NHANES III studies as well as from outpatients from our pathology service.

Varying assumptions on the underlying distributions were evaluated using the Box-Cox transformation methods.

Bhattacharya analysis was performed with an Excel spreadsheet application. The best fit was determined by optimisation of the linearity of the Bhattacharya curve (determined by the correlation coefficient) and the percent of the data included in the analysis.

ALT Data sets:

- NHANES III
  - National Health and Nutrition Examination Survey III (1998 - 1994, USA) Data from reference 2
  - N=2593, highest health category only
  - ALT: IFCC method, Roche Hitachi Analyser
- NORIP (Nordic Reference Interval Project)
  - Data Source: www.furst.no/norip
  - ALT: DGKC 1997 standardisation
  - n = 1080, healthy by pre-test questionnaire (2002)
  - NORIP conclusion: 10 - 70 U/L
- SydPath data
  - Results from patients attending General Practitioners.
  - ALT: IFCC method, Olympus analyser and reagent, Roche calibrator (2006 data)
  - Current SydPath Interval: 0 - 30 U/L

Box - Cox transformations

Box-Cox is a family of transformations which includes most others. The value of λ is selected to transform the data (T(Y)) into the best fit for a Gaussian distribution.

Formula: $T(Y) = (Y^{\lambda} - 1)/ \lambda$

- λ = 1: No change is distribution
- λ = 0.5: Square root transformation
- λ = 0.2: Skewed right (less than Log)
- λ ~ 0: Log transformation
- λ = -0.2: Heavily skewed right (more than log)
- λ = -0.5: More heavily skewed right

λ values for a range of analytes from various data sets are available in Horn and Pesce (1)

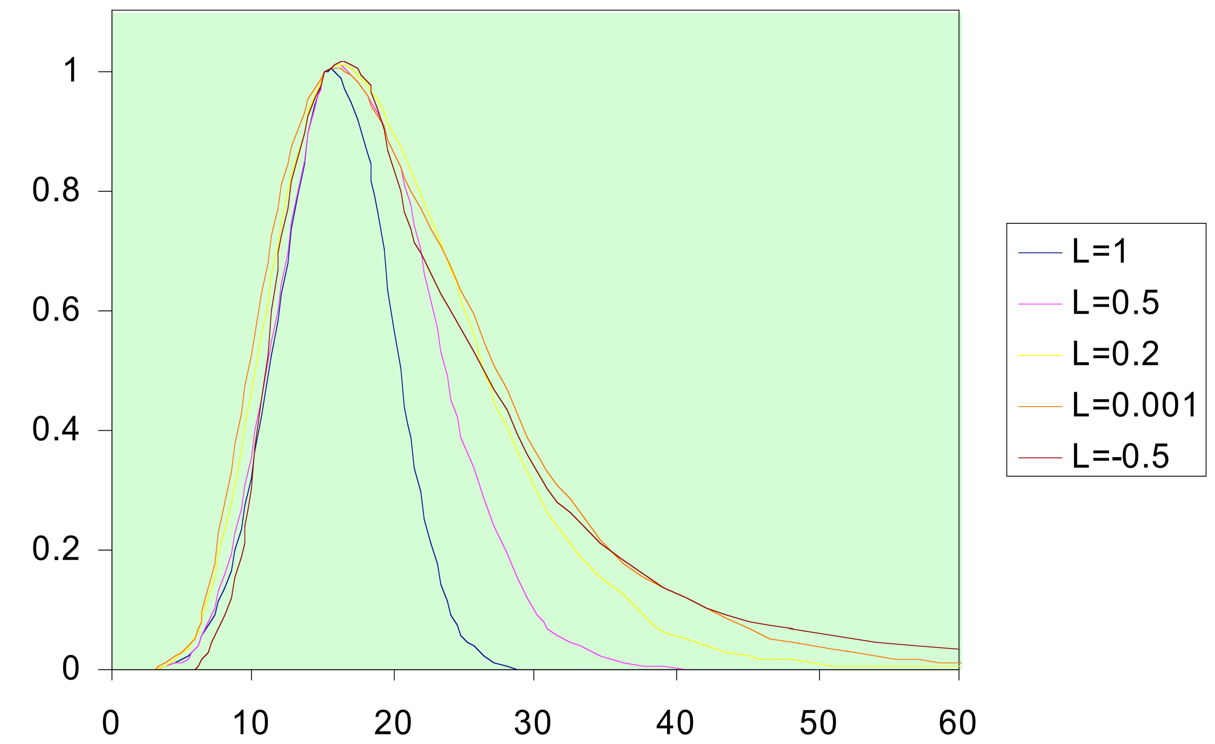

Figure 1. Examples of distributions with different Lambda values

Results

Inspection of all three ALT data sets revealed heavily right-skewed distributions (figure 2).

For all three data sets, the value of λ giving the best fit using Bhattacharya analysis, indicated by the percent of data included in the transformed Gaussian distribution and the correlation coefficient, was -0.5. (figure 3, table 1). This indicates a native distribution which is more right skewed than a logarithmic distribution.

The same transformation was optimal for the SydPath GP samples, which comprises patients presenting with unknown health issues, as for the two "healthy" data sets.

The reference intervals determined were highly dependent on the assumed distribution, with the assumption of a more skewed distribution giving markedly higher upper reference limit (figure 3, table 1).

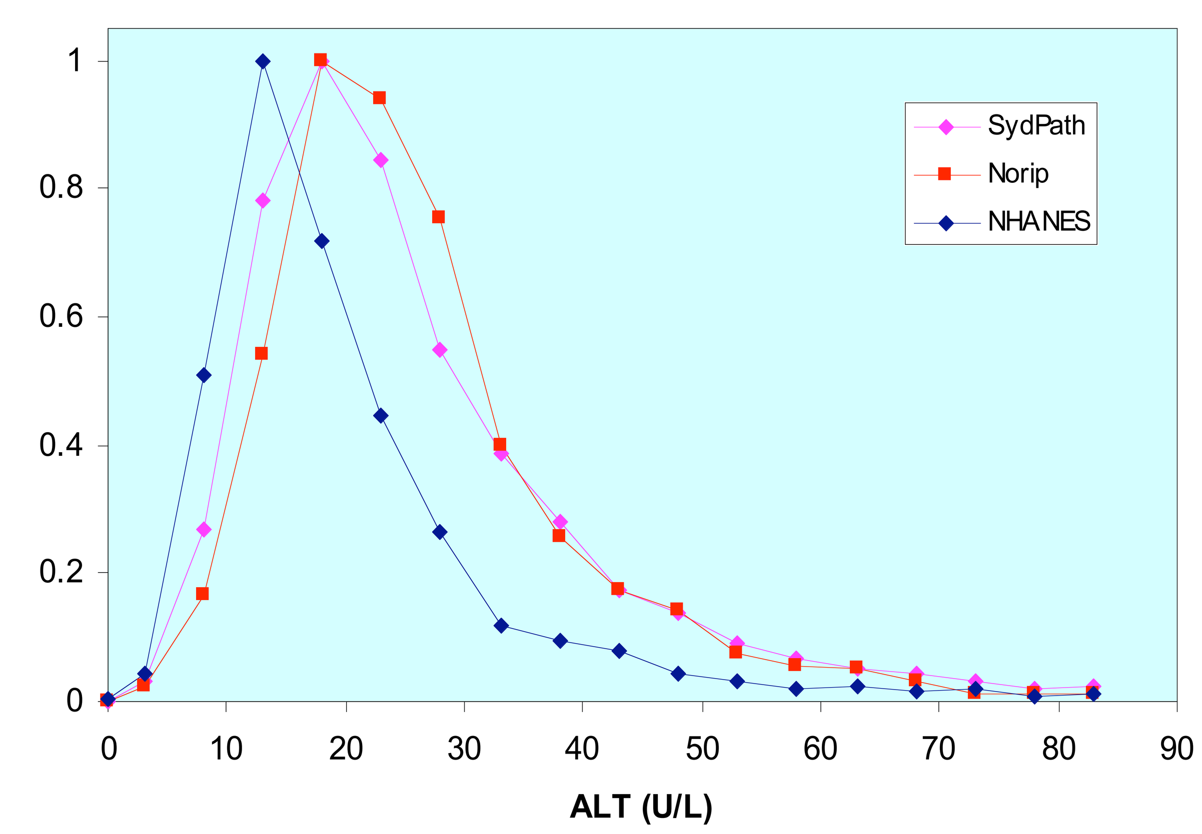

Figure 2. Population distributions for male ALT data sets.

Discussion

The study has demonstrated that assumptions made about the distribution of ALT has a profound effect on the reference intervals produced using Bhattacharya analysis.

ALT can be considered an example of other skewed distributions where data transformation may be used prior to reference interval determination.

The healthy populations show no difference to the outpatient population in the nature of the distribution however the different results may indicate analytical or population differences.

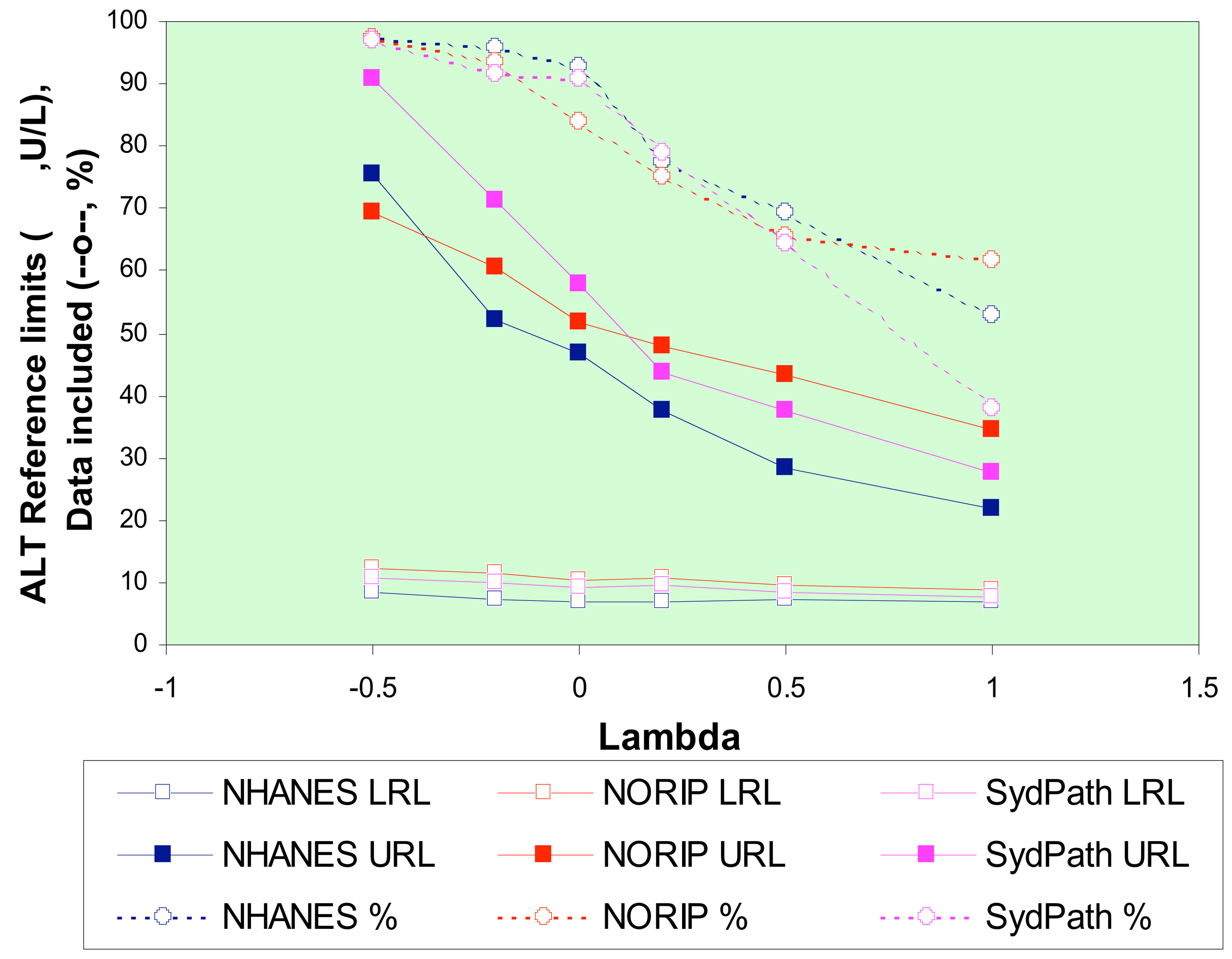

Figure 3. Derived upper and lower reference intervals and % data included in the analysis for varying vaues of lambda.

	Lambda	%	r	URL
NHANES III	0	93	-0.998	59
	-0.2	96	-0.998	63
	-0.5	97	-0.999	91
NORIP	0	84	-0.997	52
	-0.2	93	-0.996	61
	-0.5	98	-0.999	76
SydPath	0	90	-0.997	52
	-0.2	92	-0.997	61
	-0.5	97	-0.999	100

Table 1. Upper reference limits and Bhattacharya parameters for various values of lambda.

Conclusions

Bhattacharya analysis has previously been shown to be a robust tool for reference interval determination for analytes with an underlying Gaussian distribution, however the derived reference intervals for analytes with skewed distributions are highly dependent on the assumptions made about the underlying distribution.

Factors other than pure statistical analysis are likely to be required for reference interval determination of analytes with skewed distributions.

References

1. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. Journal of the Biometric Society. 1967;23:115-135.
2. Taylor N et al. Patient data defined reference intervals. Clin Biochem Revs, 2001;23:89.
3. Horne and Pesce A. Reference Intervals, a user guide. AACC Press 2005

AACB Annual Scientific Meeting, Hobart, 2006