

MULTIPLE IMPUTATION OF ORDINAL VARIABLES USING ADAPTIVE THRESHOLDS

Lynne Moore PhD, James Hanley PhD,
André Lavoie PhD, Alexis Turgeon MD MSc, FRCPC

Affiliations:

Department of Biostatistics and Epidemiology, McGill University

Unité de Traumatologie-Urgence-Soins Intensifs du CHA, Université Laval

Funding:

Canadian Institutes of Health Research

Fonds de la Recherche en Santé du Québec



Context

- Ordinal variable (non-Gaussian distribution) with missing data
- Multiple Imputation under a multivariate normal model



Possible strategies

- Single linear term
- $K-1$ dummy variables:
 - Impute $k-1$ dummy variables
 - Transform back using threshold 0.5



Problem

Conventional threshold of 0.5 may not work well for proportions close to 0 or 1



Potential solution

Adaptive thresholds:

$$C(p) = p - \Phi^{-1}(p) \times \sqrt{p(1-p)}$$



OBJECTIVE

Evaluate the performance of adaptive thresholds for imputing data measured on an ordinal scale



METHODS



Study data

- Trauma registry of the level I trauma center in Quebec City 1999-2006
- Glasgow Coma Score (GCS: 3-15)
- Use observations with observed GCS (60%)
- Impose 40% missing data completely at random



Multiple imputation

- Series of 12 dummy variables
- MCMC (Multivariate normal model)
- EM starting values
- Non-informative prior
- Single chain
- 5 imputes



Specification of the GCS

- Single linear term
- 12 dummy variables back-transformed using:
 - Conventional threshold of 0.5
 - Adaptive thresholds



Auxiliary variables

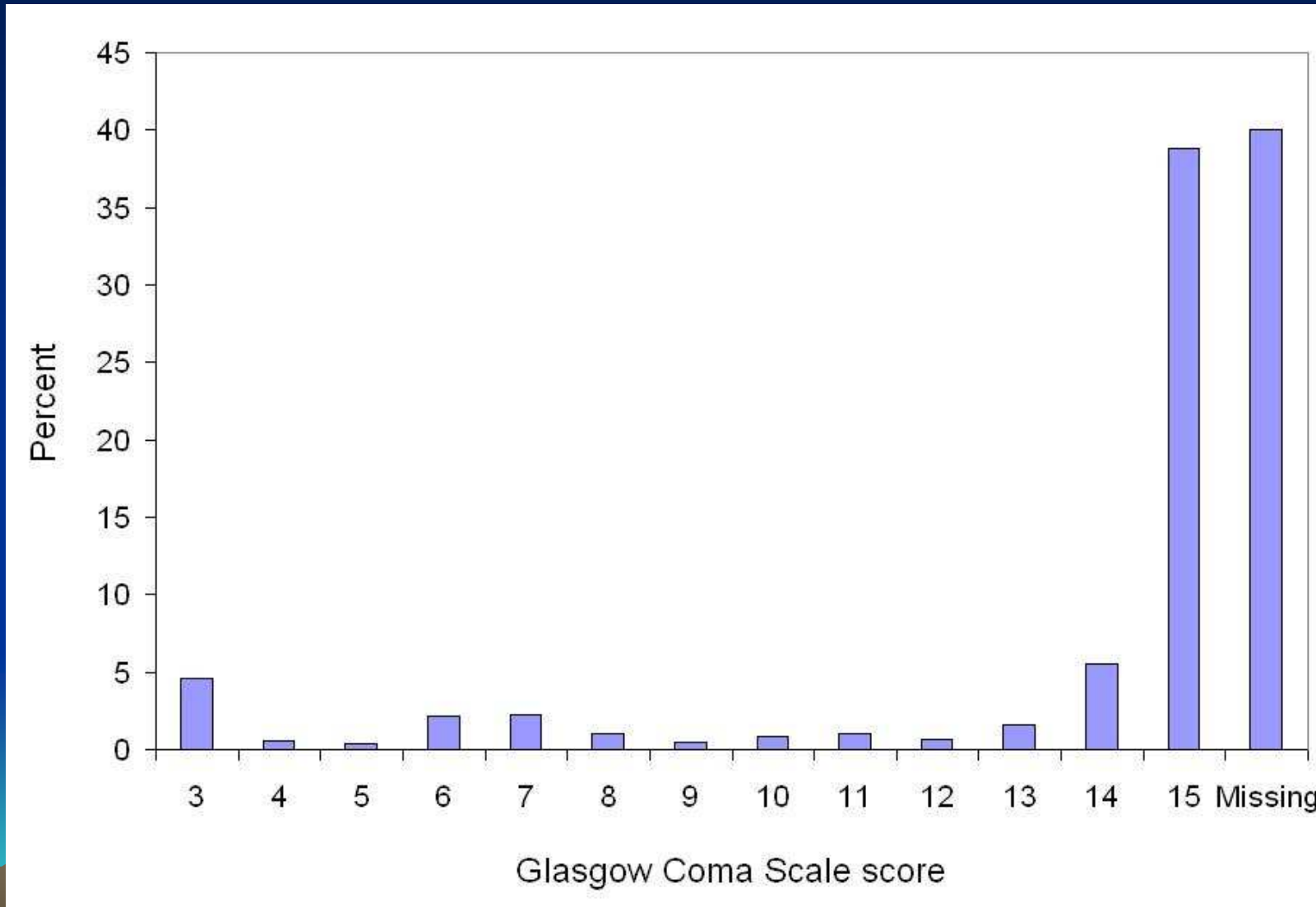
- Age
- New Injury Severity Score
- Systolic blood pressure
- Respiratory rate
- Pupil reaction
- Transfer status
- Injury mechanism
- Destination on discharge



RESULTS



Glasgow Coma Score

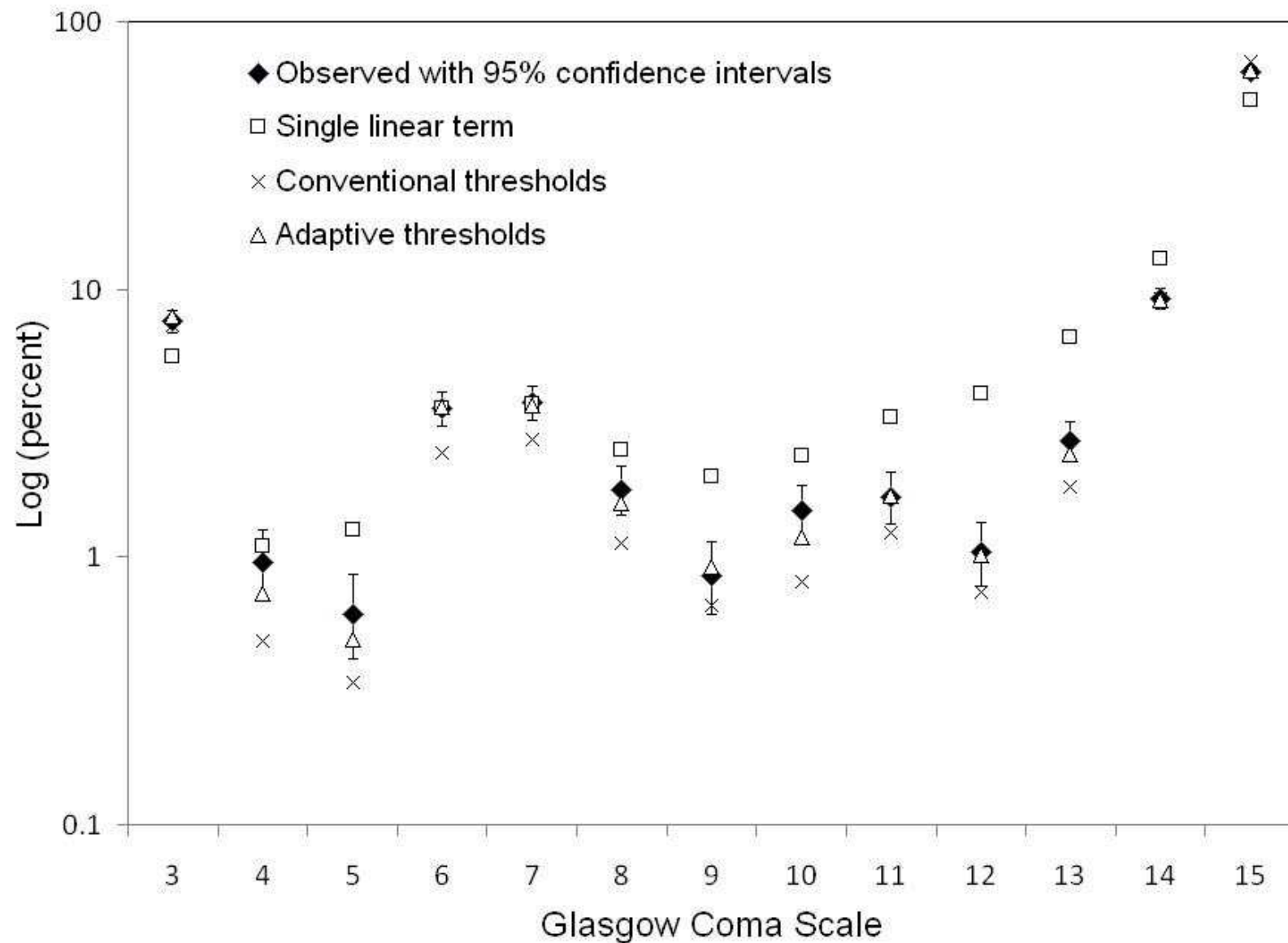


Results

- 4,720 / 7,867 (60%) with non missing GCS
- 1,888 (40%) missing GCS imposed
- Autocorrelation and time series plots OK
- relative efficiency > 95%



Observed and simulated frequency distribution of the Glasgow Coma Score



Summary

- MI Software often limited to multivariate normal model for arbitrary missing data patterns
- Model specification for ordinal data?
- Study results suggest:
 - Single linear terms inappropriate
 - Dummy variables back-transformed using $p=0.5$ can lead to biased frequency distributions
 - Dummy variables back-transformed using adaptive thresholds lead to valid frequency distributions



Limitations and future directions

- Only one variable and one sample used
- Data Missing Completely At Random
- Impact on regression coefficients (e.g. association with mortality)?



CONCLUSION

Results suggest that ordinal data can be imputed in a multivariate normal model using dummy variables providing adaptive thresholds are used



Thank You!



Maximum imputed value

