Analysis of Longitudinal Data in Perinatal Trials when the Length of Follow-up is Informative

Tessa Longstaff

Thesis submitted for the degree of

Master of Philosophy

in

Statistics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

May 14, 2016

Contents

List of Tables

List of Figures

Abbreviations

ARGEE	Autoregressive Generalised Estimating Equation
CWGEE	Cluster Weighted Generalised Estimating Equation
EGEE	Exchangeable Generalised Estimating Equation
GEE	Generalised Estimating Equation
ICS	Informative Cluster Size
IEE	Independence Estimating Equation
MM	Mixed Model
NICS	Non Informative Cluster Size
POPPET	Providing Optimal Protein for Prems via Enteral Tubes
SD	Standard Deviation of the Estimates
SE	Model Based Standard Error
WCR	Within Cluster Resampling

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....

Acknowledgements

I would like to begin by thanking my two supervisors, Associate Professor Gary Glonek and Dr. Lisa Yellend, for their help and guidance throughout my Masters.

I am grateful to the staff at the Women's and Children's Health Research Institute who were involved in conducting the POPPET trial and for sharing the data.

Of course I couldn't do this without the help of my family: my dad, Simon, for always being interesting in what I was doing and my mother, Jacqui, for always encouraging me.

Finally, I would like to thanks my partner, Brock, for his ongoing support and always being there for me.

Abstract

Background: Most commonly used statistical methods assume that the data consist of independent observations. Clustered data occur in many settings, such as longitudinal studies, where outcomes are repeatedly measured over time on each subject. Observations from the same subject are dependent and hence form a cluster. Two commonly used methods of analysis for clustered data are mixed models and generalised estimating equations (GEEs).

Additional complexity arises when analysing clustered data where the cluster size is informative; that is, where the cluster size is related to the outcome. Most methods of analysis for clustered data, including mixed models and GEEs, generally assume non informative cluster size and hence may not be suitable when the cluster size is informative.

Aim: The aim of this thesis is to compare methods for analysing longitudinal data when the cluster size (length of follow up) is informative.

Methods: Both real and simulated data were used to compare methods for analysing clustered data with informative cluster size. A range of methods were considered including: GEEs with independent, autoregressive or exchangeable working correlation structures; cluster weighted GEEs; and mixed models. The real data come from a perinatal trial (the POPPET trial), which investigated the effect of high versus standard protein content human milk fortifier on the growth of 60 preterm infants. This dataset was used to investigate different methods of analysis for estimating the effect of treatment on infant growth when informative cluster size was suspected.

As real data cannot be used to show which methods of analysis are performing best in general, a simulation study was conducted to compare methods when the true parameter values were known. The data were simulated based on the POPPET trial. Different treatment effects, sample sizes, and correlations between the cluster size and the outcomes were considered.

Results: For the POPPET trial, evidence of informative cluster size was found. Different methods of analysis produced quite different parameter estimates but similar conclusions about the effect of the intervention.

The simulation results showed that when cluster size was non informative, all methods performed very well. When cluster size was informative, mixed models and autoregressive GEEs always performed well. However, the independence, exchangeable and cluster weighted GEEs often produced low coverage probabilities and model based standard errors that differed from the standard deviation of the parameter estimates. These methods generally performed better when the trial size was larger and when there was no correlation between individual growth trajectories and cluster size.

Conclusions: It is recommended that mixed models or autoregressive GEEs be used to analyse longitudinal data with informative cluster size in general, including the POPPET trial data. Independence, exchangeable and cluster weighted GEEs should only be used when the sample size is large and there is no correlation between individual growth trajectories and cluster size.