MODELING ZERO-INFLATED AND OVERDISPERSED COUNT DATA:

APPLICATION TO IN-HOSPITAL MORTALITY DATA

By

Lisa Nanni

Cuilan Gao
Associate Professor of Statistics
(Committee Chair)

Jin Wang
Professor of Mathematics
(Committee Member)

Sumith Gunasekera
Associate Professor of Statistics
(Committee Member)

Roger Nichols
Associate Professor of Mathematics
(Committee Member)

MODELING ZERO-INFLATED AND OVERDISPERSED COUNT DATA:

APPLICATION TO IN-HOSPITAL MORTALITY DATA

By

Lisa Nanni

A Thesis submitted to the Faculty of the University of
Tennessee at Chattanooga in Partial
Fulfillment of the Requirements of the Degree
of Master of Science: Mathematics

The University of Tennessee at Chattanooga
Chattanooga, TN
August 2019

ABSTRACT

Hyperchloremia (high serum chloride level) is frequently observed in critically ill patients in the intensive care unit (ICU). Clinical evidence shows that hyperchloremia is associated with increased in-hospital mortality. Length of hospital stay (LOS) is often used as an indicator of hospital efficiency, a proxy of resource consumption and is especially important in organizing hospital services. Such data often have a highly right-skewed distribution for non-zero values and possible excess zero counts. Our study aims to examine the association of serum chloride levels at different time points with hospital mortality and to model the length of hospital and ICU stays in conjunction with zero-inflated and overdispersed count data. This research will consider the use of several univariate and multivariate models to evaluate the effects of serum chloride as it pertains to patient mortality. This research resulted from application to more than 1700 critically ill patients from a local hospital.

DEDICATION

This Thesis is dedicated to my husband, Dino Nanni. He has encouraged and believed in me for so many years.

ACKNOWLEDGEMENTS

I would like to acknowledge the University of Tennessee at Chattanooga Mathematics department for the support and guidance. I would particularly like to thank Dr. Lani Gao for being an excellent teacher and mentor. Her knowledge in the field of Statistics both in academia and in private industry have been invaluable in my studies. I would also like to thank all my committee members for their time and efforts toward helping me achieve this degree. Lastly, I would like to thank my friends and family for their constant support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ICU, Intensive Care Unit

LOS, Length of Stay

AIC, Akaike Information Criterion

BIC, Bayesian Information Criterion

NB, Negative Binomial

ZIP, Zero-Inflation Poisson

ZIBN, Zero-Inflation Negative Binomial

GLM, Generalized Linear Model

SCL, Serum Chloride Levels

GFR, Glomerular Filtration Rate

AKI, Acute Kidney Injury

APACHE, Acute Physiologic and Chronic Health Evaluation OR, Odds Ratio

ZABN, Zero-Altered Negative Binomial

CHAPTER 1

INTRODUCTION

Modeling count variables is a common task in the medical and social sciences. The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data sets typically exhibit over-dispersion and/or an excess number of zeros. As Count data frequently occur in many fields, including public health, medicine and epidemiology, it is necessary to find models which will better account for these issues. A few common examples of count data in these fields are the number of deaths, number of cigarettes smoked, and number of disease cases. For such data the Poisson model is a commonly applied statistical model as a starting point. A key feature of the Poisson model is that the mean and the variance are approximately equal.

Departures from a Poisson model can occur in a variety of ways; the main reasons are: (1) some covariates may be omitted and/or may not have a uniform effect on all subjects so that population heterogeneity has not been accounted for, and (2) an excess number of zero events occurred compared to the Poisson distribution (Lindsey 1995; Lindsey and Altham 1998) . For the excessive zeros situation, it could be assumed that a sample is collected from two different sub-populations; one population always produces zero, or no event, while the other behaves like a Poisson distribution.

1.1 Overdispersion in Count Data

In medical research, data are often collected in the form of counts which are related to the number of times that an event of interest occurs. Because of their simplicity, one-parameter distributions for which the variance is directly determined by the mean are often used at least in the first method to model this data. However, the equal mean-variance relationship rarely happens with real-life data (Cox 1983; Dean 1992; Prentice 1986). In most cases, the observed variance is larger than the assumed variance, which is known as overdispersion. If the overdispersion is ignored, statistical inference results in an inaccurate conclusion by underestimating the variability of the data (Cox 1983). If this dispersion is not taken into account, then using these models may lead to biased estimates of the parameters and consequently incorrect inferences about the parameters.

Several statistical methods have been proposed for analysis of count data with overdispersion. Many of them used negative binomial distribution to model the count data (Pounds and Zhang 2012; Auer and Doerge 2012; McCarthy and Smyth 2010). In this research study, we demonstrate the use of various models for overdispersed count data. These are Poisson, negative binomial, Quasi-Poisson, and Zero-inflated models. The real data in this research study deal with the hospital stays and ICU stays of patients in critially ill conditions. The models resulted in differing statistical inferences. The Poisson model, which is widely used in epidemiology research, underestimated the standard errors and overstated the significance of some covariates. The models were compared in terms of covariate estimates along with their statistical inferences. Akaike‘s Information Criterion (AIC) values were used to consider the relative model fitting for the models as a goodness-of-fit statistic.

1.2 Zero-Inflation in Count Data

In psychological, social, and public health related research, it is common that the outcomes of interest are relatively infrequent behaviors and phenomena. Data with abundant zeros are especially frequent in research studies when counting the occurrence of certain behavioral events, such as number of school absences, number of cigarettes smoked, number of day of hospitalizations, or number of ICU days. These types of data are called count data and their values are usually nonnegative with a lower bound of zero and typically exhibit excessive zeros and overdispersion (i.e., greater variability than expected). Except for transforming the outcome to make it normal and using the general linear model, other alternative approaches can be taken in the context of a broader framework. For example, the Poisson distribution becomes increasingly positively skewed as the mean of the response.

Thus, a typical way of analyzing count data includes specification of a Poisson distribution with a log link (the log of the expectation of a response variable is predicted by the linear combination of covariates, i.e., predictors) in a model known as Poisson regression. Equal mean and variance of the response variable is the main condition for using Poisson model. If the condition is not fulfilled, then generalized Poisson and negative binomial models are appropriate (Karlis and Xekalaki 2005; Yau and Wulu 2003; Famoye and McGwin 2002). Generalized Poisson regression and negative binomial regression models are, to some extent, capable of determining dispersion. When the zero data are exceeded, data such as length of stay (LOS) of these models will not be efficient. These methods cannot be used to explain and analyze overdispersion in LOS data, so one of the proper approaches to analyze them is zero-inflated regression models (Hilbe 2011).

Several other more rigorous approaches to analyzing count data include the zero-inflated Poisson (ZIP) models that have been proposed recently to cope with an overabundance of zeros (Greene 1994; King 1989; Lambert 1992; Mullahy 1986). These two types of models both include a binomial process (modeling zeros versus non-zeros) and a count process. The difference between the two models is how they deal with different types of zeros: although the count process of ZIP is a zero-truncated Poisson (i.e. the distribution of the response variable cannot have a value of zero), the count process of ZIP can produce zeros (Zuur and Smith 2009). One of the assumptions of using Poisson regression is that the mean and variance of a response variable are equal. In reality, it is often the case that the variance is much larger than the mean. Variations of negative binomial (NB) models can be used when overdispersion exists even in the non-zero part of the distribution. Although a Poisson distribution contains only a mean parameter, a negative binomial distribution has an additional dispersion parameter (k) to capture the amount of over-dispersion. Thus, the zero-inflated negative binomial (ZINB) model and zero-altered negative binomial (ZANB) model were introduced to deal with both zero-inflation and over-dispersion.

CHAPTER 2

BACKGROUND

A retrospective study of 1724 patients at Erlanger Hospital, Chattanooga, was conducted. The inclusion criteria for this study is critically ill patients at least 18 years of age who were admitted to a medical intensive care unit. This is a cohort study of patients admitted to the ICU on whom data were extracted via electronic medical records.

2.1 Study Significance

Chloride plays a pivotal role in many body functions including acid-base balance, muscular activity, osmosis, and immunomodulation (Berend 2012). Despite its physiological importance, chloride has captured little attention by the scientific community until recently (Yunos and Story 2010) when chloride-rich solutions were associated with hyperchloremic metabolic acidosis (Scheingraber and Sehmisch 1999; Robinson and Smyth 2009) and short-term mortality after non-cardiac surgery (Silva and Santana 2009; McCluskey and Wijeysundera 2013). Hyperchloremia is frequently observed in critically ill patients in the ICU. The primary aim of our study is to determine whether there is independent association of serum chloride (CI) levels at two difference time points of ICU stay with hospital mortality in critically ill patients.

Optimal choice of intravenous fluids in specific clinical scenarios has been a subject of much debate and there are few formal recommendations on the ideal initial fluid choice in the majority of critically ill patients. Normal saline (0.9% NaCl) is the most commonly used

intravenous fluid. However, infusion of the large volumes (> 2 liters) of 0.9% saline induces hyperchloremia, metabolic acidosis, hyperkalemia, and a negative protein balance. This fluid contains 50% more chloride than is typically present, making the moniker 'Normal Saline' a misnomer. The high chloride content is largely responsible for the acidic nature of 0.9% saline as it reduces the strong anion difference resulting in a reduction in pH and the associated systemic effects of this pH shift. Chloride, the most abundant extracellular anion, plays a key role in balancing extracellular cations and extracellular tonicity, and excess chloride is known to contribute to edema.

Increasing evidence suggests that hyperchloremia results in a reduction of the glomerular filtration rate (GFR), likely through afferent arteriolar dilatation caused by increased chloride delivery to the tubules. Clinically, hyperchloremia induced reduction in GFR is independently associated with increased risk for acute kidney injury (AKI).

From perioperative literature, it is evident fluids other than 0.9% saline may provide better choices for most patients. Perioperative infusion of 0.9% saline is associated with increased hospital length of stay and need for blood products in patients following abdominal and cardiovascular surgery. In contrast to perioperative literature, there is only scant evidence for fluid choice in patients in the medical ICU. Initial studies suggest that hyperchloremia is associated with increased mortality in patients with sepsis; however, this relationship requires further validation. Additionally, studies examining the prognostic value of hyperchloremia in critically ill, non-septic patients are lacking.

2.2 Description of the Variables

Of the 1724 patients, variables collected include demographic data (age, race, gender), admission and subsequent levels of serum Cl, length of hospital and ICU stay, mortality, and other laboratory data. Patients for whom all data was not present were excluded from this study. All patient data was used in a de-identified form whereby each patient was assigned a unique identification number.

The data regarding mortality takes the values of "death" or "no death" which represents a binary response variable. This study aims to look at the effect of serum chloride levels at time of admission, 72 hours after admission and the change in these levels in association with outcome of mortality. The associated potential explanatory variables include patient age, gender, race and APACHE (Acute Physiologic Assessment and Chronic Health Evaluation) score.

CHAPTER 3

REVIEW OF CURRENT LITERATURE

This section presents the background of the current models and the theory that has accumulated in regard to logistic regression and count data modeling. The framework for each model is presented along with the relationships between models to examine the potential benefits and problems in model selection.

3.1 Generalized Linear Models (GLM)

The GLM extends ordinary regression models to include non-normal response distributions. Three components specify a generalized linear model:

1. A random component identifies the response variable Y and its probability distribution.
2. A systematic component specifies explanatory variables used in a linear predictor function.
3. A link function specifies the function E(Y) that the model associates with the systematic component.

3.2 Components of Generalized Linear Models

The random component of a GLM can be expresses by a response variable Y which has independent observations from a distribution of the natural exponential type. The probability density function has the form,

$$f(y_i|\theta_i, \phi) = exp\left[\phi^{-1}(y_i\theta_i - b(\theta_i)) + C(y_i, \phi)\right] \quad (1)$$

where ϕ is a known or unknown constant or an estimated parameter. This represents the equation in canonical form with natural or canonical parameter θ which may vary based on the values of the explanatory variables (Lindsey 1995).

The systematic component relates the explanatory variables as a combination of linear predictors.

This linear combination of explanatory variables is then called the linear predictor and can be denoted as,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \quad (2)$$

The link function is the connection between the random and the systematic components. It denotes how the expected value of the response variable relates to the linear predictor of explanatory variables. The model links μ to η = g(μ) where the link is a monotonic differentiable function. The link function that converts the mean to the natural parameter is called the canonical link (Lindsey 1995).

For the normal distribution ϕ = σ, for the Poisson distribution ϕ = 1 and for the binomial distribution ϕ = 1/n, where n is the binomial index. For over-dispersed count data, ϕ can be considered as an over-dispersion parameter to be estimated from the data (Famoye and McGwin 2002).

CHAPTER 4

METHOD

Each statistical analysis includes first a univariate analysis to assess the relationship between each independent variable with the outcome variable followed by a multivariate analysis of the data. An alpha level of significance of 0.1 was used to assess covariates for each model.

4.1 Logistic Regression Model

For this portion of the study, the statistical testing was completed by using a GLM with logistic regression modeling. Because the response variable is binary, or dichotomous, a logit transformation of the response variable will make a correlation between the predictor and response variable linear. This test will be used to test for an association of the mortality with the predictors. The hypothesis will be tested as follows:

H_0 = No association between mortality and predictor

H_a = Association between mortality and predictor

Let π be the probability of an event occurring. This means that the odds ratio, OR, is given by:

$$\frac{\pi}{1-\pi} \tag{3}$$

When the logarithm of the odds ratio is taken, the result is the Log-Odds function,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \tag{4}$$

where β_0 is the intercept and β_1, β_2, β_3… represent predictors. By exponentiating the Log-Odds function, the result is the logistic function,

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots}} \tag{5}$$

Odds ratio were used in these models to measure the association between an exposure and an outcome. Odds ratios look at the odds of a certain outcome with exposure to a predictor versus the odds of the outcome in an absence of the predictor. In these models, the risk factor of serum chloride is tested to determine if mortality is affected by exposure.

The initial testing looked at the initial serum chloride levels, levels after 72 hours and the change in levels as predictors for mortality. Secondary testing looked at individual covariates to determine if any single covariate had a statistically significant effect using the alpha level of significance (type I error) of 0.1. Upon determination of significant covariates, a new model using each of the serum chloride levels along with the covariates was constructed. All models were then exponentiated to show the estimated mortality risk change per 5 unit change of serum chloride levels.

4.2 Count Data Models

For the next portion of this study, the response variables representing length of stay in the hospital and length of ICU stay were examined. As these response variables represent count data, a generalized linear model using a Poisson link function was initially constructed. The histogram in Figure 1 shows the count data distribution for length of hospital stay and ICU stay.

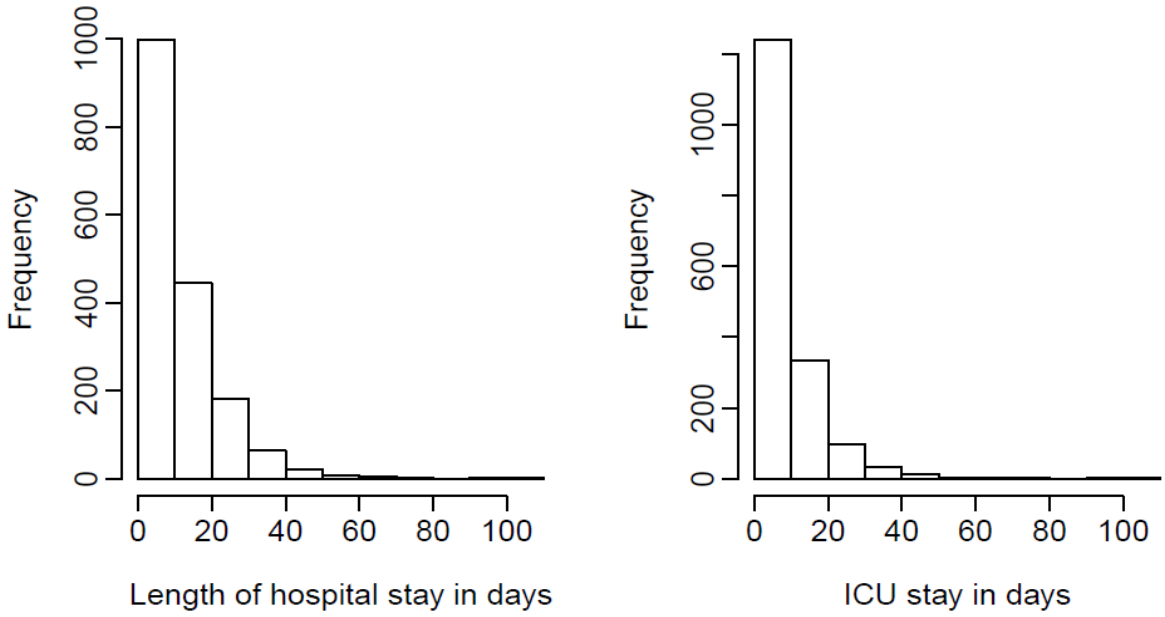

Figure 1 Count Data Length of Hospital Stay and ICU Stay

The testing was completed in a similar manor to the mortality testing in that individual serum chloride levels, then potential covariates, and finally a combined model at each level were tested. The hypothesis will be tested as follows:

H_0 = No association between length of hospital stay/ICU stay and predictor

H_a = Association between length of hospital stay/ICU stay and predictor

4.3 Overdispersion and Zero Inflation Models

For count data, Poisson and negative binomial models have been the basic building blocks. The Poisson density function is described as follows. If y follows a Poisson distribution, then:

$$P(y = y_i|\lambda) = \frac{\lambda_i^y}{e^{\lambda} y_i!} \qquad (6)$$
$$y = 1, 2, 3... \quad \lambda > 0$$

Using the GLM with Poisson link, the assumption is that the mean is approximately equal to the variance. However, a model that displays a greater variability than may be expected for that model indicates overdispersion.

The basic testing was completed to first get an initial check of results. Because there can be under or overdispersion associated with count data, a simulation of the sample data was used to verify if the variance and the mean were equal. A set of 500 simulations using 1000 samples each, taken from the data, were plotted to check the distribution of the mean versus the variance.

There are methods which can be used to correct for overdispersion in count data. Such methods include the quasi-Poisson, negative binomial, and Zero-inflated Poisson/Zero-Inflated negative binomial (ZIP/ZINB) models for correction of overdispersion. Each of these models allows for a relaxing of the restrictive nature of the Poisson model with respect to variance. In these models, the variance is a function of the mean, linear for quasi-Poisson and quadratic for negative binomial, thus large and small counts are weighted differently. Parameters are estimated using the optimization method of iteratively reweighted least-squares. ZIP models are composed of a mixture of two distributions, one of which is Poisson in nature.

Let Y be a random variable where $E(Y) = \mu$ with $Var(Y) = \theta\mu$ and θ representing a dispersion factor. If $\theta = 1$, we have the standard Poisson mean and variance relationship. However, if $\theta > 1$, overdispersion of the model will be present for Poisson. There are cases where the variance is proportional by some weighted value to the mean. This can be noted by observing the bootstrap simulations.

There are alternatives to using a strict Poisson relationship to account for the overdispersion. Both Quasi-Poisson and negative binomial regressions provide options when compensating for overdispersion. The Quasi-Poisson model is only characterized by mean and

variance and do not always have a distributional form. Therefore, theoretical approaches such as AIC (Akaike Information Criteria) or BIC (Bayesian Information Criteria) is not applicable and cannot be effectively directly compared to other models under this criterium.

The density for the negative binomial is defined as follows:

$$P(y = y_i) = \frac{\Gamma(y_i + \delta_i)}{\Gamma(y_i + 1)\Gamma(\delta_i)} \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i}\right)^{y_i} \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

The negative binomial model depends on a parameterization such that E(Y) = μ and Var(Y) = μ + κμ2 with κ > 0. The overdispersion depends on the multiplicative factor 1 + μ. So, for the Quasi-Poisson, the variance is linearly related to the mean and for the negative binomial, the variance is quadratic in the mean. Thus, the weighting in the Quasi-Poisson model is directly proportional to the mean. In the negative binomial model, smaller mean values get smaller weights and larger mean values increase but level off to 1/κ. The Quasi-Poisson model depends only on the overdispersion parameter θ and does not have a direct density function.

The ZIP model is proposed when data shows variations caused by the occurence of extras zeroes either structurally or occurring from sampling. The ZIP may be viewed as a combination of two generating processes, the first of which is a degenerate component centering mass at zero and the second a Poisson governing process which handles count. The ZIP distribution can be given as:

$$\begin{aligned} P(Y = 0) &= p + (1 - p)e^{-\lambda} \\ P(Y = k) &= (1 - p)e^{-\lambda}\lambda^k/k! \quad k = 1, 2, \ldots \end{aligned} \tag{8}$$

If the data shows extra zeroes and unobserved heterogeneity, the ZINB model is recommended. The ZIBN model is given as:

$$P(Y=0) = \pi_i + (1-\pi_i)(1+k\lambda_i)^{-1/k}$$
$$P(Y=y_i) = (1-\pi_i)\frac{\Gamma(y_i+1/k)}{\Gamma(y_i+1)\Gamma(1/k)}\frac{(k\lambda_i)^{y_i}}{(1+k\lambda_i)^{y_i+1/k}} \quad (9)$$

with probability π_i and κ as overall dispersion parameter,

The count data characteristics for all models are given in the following table.

Table 1

Count Data Model Characteristics

Model	$E[y_i]$	$E[y_i \mid y_i > 0]$	$var[y_i]$
P	λ_i	$\lambda_i c_i^{-1}$	λ_i
NB	μ_i	$\mu_i d_i^{-1}$	$\mu_i + \mu_i^2 \delta_i^{-1}$
ZIP	$(1-\pi_i)\lambda_i$	$\lambda_i c_i^{-1}$	$\lambda_i(1-\pi_i)(1+\pi_i\lambda_i)$
ZIBN	$(1-\pi_i)\mu_i$	$\mu_i d_i^{-1}$	$\mu_i(1-\pi_i)(1+\mu_i\delta_i^{-1}+\pi_i\mu_i)$

Note: $\lambda_i, \mu_i, \delta_i > 0, \pi_i \in (0,1)$. Also, $c_i = 1 - exp(-\lambda)$ and $d_i = 1 - (\frac{\delta_i}{\mu_i+\delta_i})^{\delta_i}$

As an alternative to conventional models, assumption adequacy averaging could be applied (Pounds and Zhang 2012; Auer and Doerge 2012; McCarthy and Smyth 2010). By using the law of total probability to proportion the data based on a preset threshold, the data within the threshold could be modeled using a method such as the negative binomial. The data outside the threshold would then be modeled separately. Determining the proper proportion for each model may yield a combined model which shows better fit qualities.

To compare all models, the p-values for each model were considered. For the Poisson and negative binomial models the AIC were also stated and considered. All models were exponentiated and examined based on a 5 unit change of serum chloride levels.

CHAPTER 5

RESULTS

The statistical analysis was performed using R software and includes examination of three response variables: mortality, length of hospital stay and length of ICU stay. The R software package is a language and environment for statistical computing and graphics and is widely used for statistical modeling.

5.1 Mortality Model

For the mortality study, the model was developed first by examining the serum chloride levels at admission, after 72 hours and the difference in levels. From these models, it appears that the initial serum chloride levels alone do not show statistical significance in relation to mortality. However, the levels at 72 hours and the difference in levels from admission to 72 hours show evidence of statistical significance on mortality as shown in Table 2.

Table 2

Univariate Model for Mortality

CIO			**CI72**			**Cidiff**		
OR	p-value	AIC	OR	p-value	AIC	OR	p-value	AIC
0.9052	0.0092	2022.6	1.1734	4.78E-05	2014	1.3088	2.32E-11	1980.1

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels at admission, 72 hours after admission and for the change in levels from initial to 72 hours.

The covariates under investigation included age, gender, race and APACHE score. For the age covariate, the group was segmented into those subjects less than 65 years of age and those 65 years or older.

To determine which covariates should be added to the univariate model, based on a significance level of 0.1, each covariate was assessed for significance on mortality. The age covariate showed statistical significance in terms of mortality, so the covariate was added to the models for each of the serum chloride levels.

$$Mortality \sim \beta_0 + \beta_1 age \tag{10}$$

The multivariate model, using p-values, shows that for serum chloride levels at admission, 72 hours and the difference in levels from admission to 72 hours is statistically significant in reference to mortality. Generally, the odds ratio represents the change in response based on a 1 unit change in predictor. The odds ratio in these models is based on a 5 unit change in serum chloride as the units represented are extremely small. Based on the odds ratio, shown in Table 3, both the change in serum chloride level and the age greater than 65 years demonstrates an increased risk of mortality.

Table 3

Multivariate Model for Mortality

CI0				
CI0		**Age**		AIC
OR	p-value	OR	p-value	1958.8
0.9017	0.0078	32.2322	1.98E-10	
CI72				
CI72		**Age**		AIC
OR	p-value	OR	p-value	1953.6
1.1547	0.0003	3.2106	3.20E-10	
CIdiff				
CIdiff		**Age**		AIC
OR	p-value	OR	p-value	1921.6
1.294	2.72E-10	29.3944	1.28E-09	

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels with age as a covariate.

5.2 Length of Stay Model

The length of hospital stay was next considered as a response variable. Because the response is represented by count data, a GLM using a Poisson link function was considered. The Poisson link assumes that the mean and the variance are close to equal. To test this assumption, a bootstrap simulation using the hospital length of stay data was completed. A sample size of 1000, with replacement, was taken 500 times to check the mean versus the variance. Figure 2 shows the simulation data to compare the mean and variance and strongly shows the relationship is not an equal mean and variance and demonstrates overdispersion.

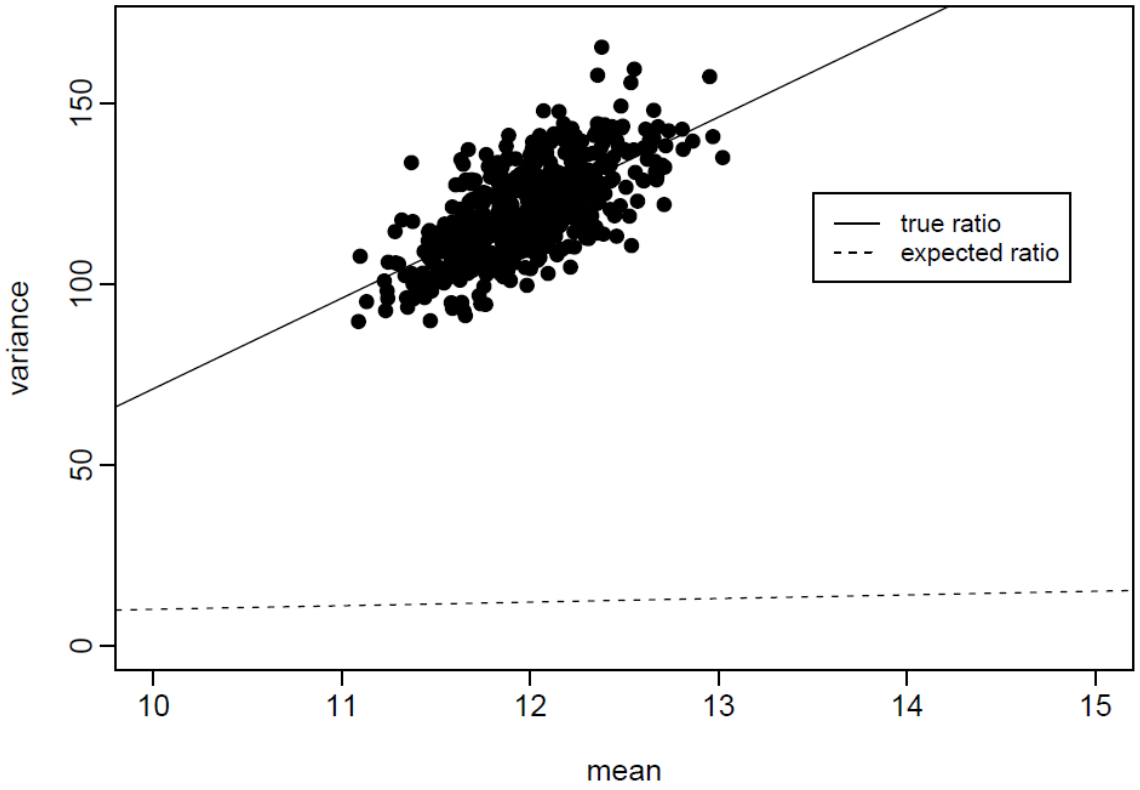

Figure 2 Bootstrap Samples for Hospital Length of Stay

Due to the apparent overdispersion of the model, the model with Poisson link was compared with models using the Quasi-Poisson link and the negative binomial model. For the Poisson and the negative binomial models, the p-values, odds ratio, and AIC were compared. It should be noted that the Quasi-Poisson model does not produce a comparable AIC value and therefore only p-values and odds ratio are shown in Table 3. The univariate model considered the serum chloride levels at admission, 72 hours, and the difference in levels.

All univariate models produced p-values indicating statistical significance associating the length of hospital stay with serum chloride levels. Comparing the Poisson and negative binomial models, the AIC for the negative binomial model was much lower indicating a possible better model quality and selection.

The covariates tested in the mortality model were again used individually with a significance level of 0.1 with length of hospital stay as response. Both APACHE score and age showed statistically significant p-values and were added as covariates for the full model. As shown in Table 4, the Poisson model showed p-values which were significant for all covariates.

$$Length of Stay \sim \beta_0 + \beta_1 age + \beta_2 APACHE score \tag{11}$$

Table 4

Univariate Model Length of Stay

	Poisson			Negative Binomial			Quasi-Poisson	
	OR	p-value	AIC	OR	p-value	AIC	OR	P-value
CI0	0.9732	3.99E-08	20512	0.9729	0.532	11823	0.9732	0.0817
CI72	1.0351	6.75E-12	20495	1.0375	0.0104	11828	1.0351	0.0291
CIdiff	1.0653	2.00E-16	20357	1.0646	1.70E-05	11792	1.0653	5.40E-05

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels using each of the models. There is no AIC for Quasi-Poisson and p-values are used for comparison

This model also shows a higher AIC value than the negative binomial model. The Odds Ratios for the full models of serum chloride at 72 hours and the difference from admission indicate that there is an increased risk for longer hospital stays with a 5 unit increase in serum chloride levels.

Table 5

Multivariate Model Length of Stay

	Poisson			Negative Binomial			Quasi-Poisson	
	OR	p-value	AIC	OR	p-value	AIC	OR	P-value
CI0	0.9746	2.32E-07		0.9752	0.0782		0.9746	0.0978
AP Score	1.0109	2.00E-16	19891	1.0095	0.0272	11539	1.0109	1.03E-05
Age	0.7464	5.22E-05		0.7615	0.1829		0.7464	0.1953
CI72	1.0386	9.78E-14		1.0414	0.0053		1.0386	0.0168
AP Score	1.0111	2.00E-16	19863	1.0098	0.0232	11543	1.0111	5.88E-06
Age	0.7291	1.26E-05		0.7185	0.1057		0.7291	0.1604
CIdiff	1.0667	2.00E-16		1.0653	1.51E-05		1.0667	3.42E-05
AP Score	1.0111	2.00E-16	20357	1.0096	0.026	11507	1.0111	5.78E-06
Age	7.109	2.50E-06		0.7268	0.118		7.109	0.129

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels using each of three models. The covariates of APACHE score and age were used for each model.

The ZIP and ZIBN models were used to test the serum chloride levels as predictors for the length of hospital stay in days. The following table shows the coefficients, p-values and standard errors associated with each model. As these models have two components, the count and zero-inflation portions, it is evident that the zero-inflation in this data is not statistically significant. This was expected since the original data set did not show a large percentage of zero counts. However, this model was reviewed to ensure that all possible outcomes were tested.

Table 6

ZIP and ZIBN Models for Length of Stay

ZIP Poisson						
	Count Coef.	p-value	SE	Zero-Inflation Coef.	p-value	SE
CI0	-0.0055	1.73E-08	0.0009	-0.0344	0.524	0.054
CI72	0.0067	2.62E-11	0.001	-0.0729	0.205	0.0576
CIdiff	0.0125	2.00E-16	0.001	-0.0388	0.543	0.0639
ZIBN Negative Binomial						
	Count Coef.	p-value	SE	Zero-Inflation Coef.	p-value	SE
CI0	-0.0055	5.61E-02	0.0029	-0.0857	0.903	0.7014
CI72	0.0074	1.29E-02	0.0029	-0.1955	0.993	23.355
CIdiff	0.0125	1.56E-05	0.0029	-0.1053	0.977	3.6881

This table gives the coefficients, p-values and standard errors for serum chloride levels using each model. Both the count and the zero inflation portions of each model are shown.

5.3 ICU Stay Model

The final response variable of interest is the ICU stay length. The univariate model tests serum chloride levels at admission, 72 hours after admission and the difference in levels. Similarly to length of hospital stay, ICU stay represents count data and the GLM model with Poisson link was considered as an initial choice.A bootstrap simulation was completed to check the mean versus variance relationship. The results are shown in Figure 3 and demonstrate overdispersion in the model.

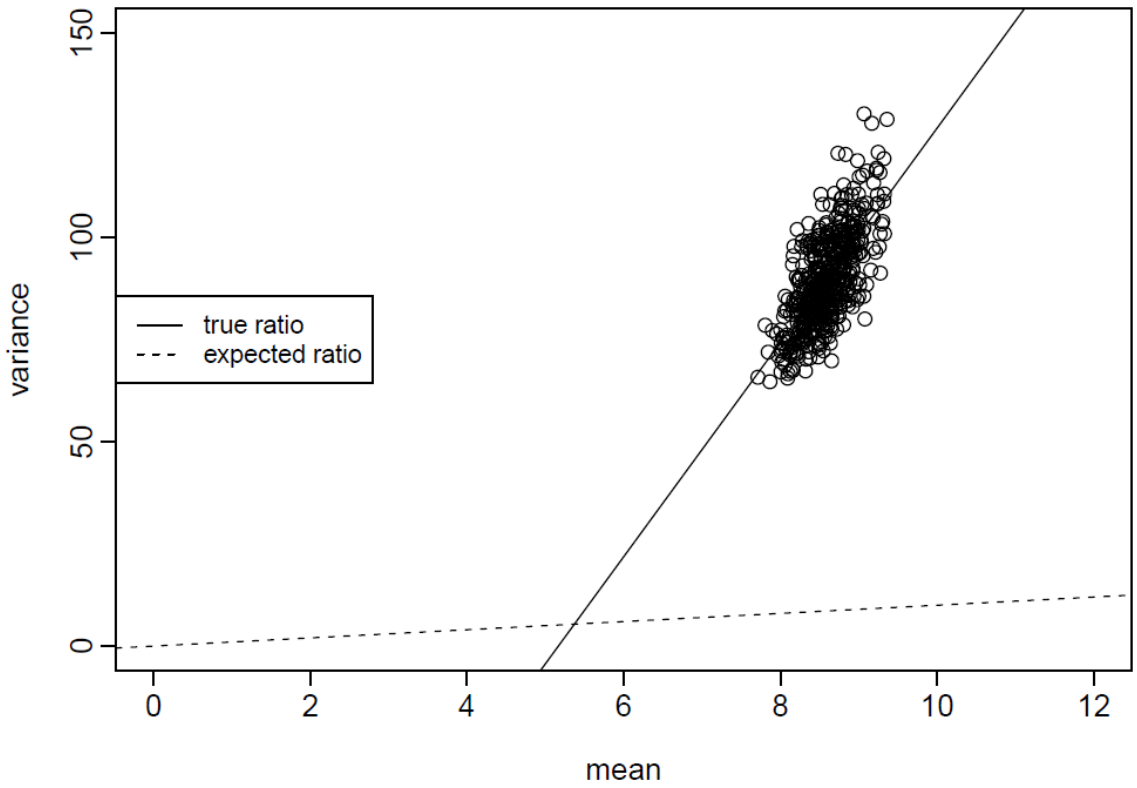

Figure 3 Bootstrap Samples for ICU Length of Stay

Similarly to the length of stay model, the ICU stay simulation shows overdispersion of the data. The covariates were tested using univariate models with the significance level of 0.1 and the results are shown in Table 7.

Table 7

Univariate Model ICU Stay

	Poisson			**Negative Binomial**			**Quasi-Poisson**	
	OR	p-value	AIC	OR	p-value	AIC	OR	P-value
CI0	0.9745	9.38E-06	19187	0.9741	0.104	10884	0.9745	0.171
CI72	1.0848	2.00E-16	19010	1.0896	1.32E-07	10865	1.0848	1.48E-05
CIdiff	1.1141	2.00E-16	18834	1.0646	4.76E-12	10824	1.1141	4.89E-09

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels.

The covariates of APACHE score, age and gender showed statistical significance in relation to length of ICU stay.

$$Lengthof Stay \sim \beta_0 + \beta_1 age + \beta_2 APACHEscore + \beta_3 gender \quad (12)$$

APACHE score shows statistical significance regarding length of ICU stay for Poisson and Quasi-Poisson models. It should be noted that the negative binomial model does not show this effect and the AIC for the negative binomial model is smaller than the Poisson model. Serum chloride levels at 72 hours and the difference between initial and 72 hours is significant for all models. Odds ratios for the Poisson and Quasi-Poisson are also consistent and show an increased risk of ICU stay length as serum chloride increases.

Table 8

Multivariate Model ICU Stay

	Poisson			**Negative Binomial**			**Quasi-Poisson**	
	OR	p-value	AIC	OR	p-value	AIC	OR	P-value
CI0	0.9759	3.28E-05	18716	0.9749	0.1180	10633	0.9759	0.1983
AP Score	1.0107	2.00E-16		1.0078	0.1130		1.0107	4.00E-04
Age	1.1077	0.232		1.1302	0.6000		1.1066	0.711
Gender	1.0243	0.774		1.0206	0.9300		1.0244	0.9292
CI72	1.0872	2.00E-16	18531	1.0912	1.27E-07	10614	1.0872	1.11E-05
AP Score	1.0111	2.00E-16		1.0083	0.0917		1.0111	2.00E-04
Age	1.0556	0.5234		1.0264	0.9106		1.0556	0.8415
Gender	1.0279	0.7433		1.0066	0.9771		1.0279	0.9183
Cidiff	1.1135	2.00E-16	18368	1.1194	9.61E-12	10574	1.1136	6.56E-09
AP Score	1.0111	2.00E-16		1.008	0.103		1.0111	2.00E-04
Age	1.0212	8.05E-01		1.0559	0.815		1.0212	0.9381
Gender	1.0293	7.31E-01		0.9881	0.958		1.0293	0.9139

This table gives the Odds Ratio, p-value, and AIC for serum chloride levels using each of three models. The covariates of APACHE score, age, and gender were used for each model.

Similarly to the data from length of stay modeled by the ZIP and ZIBN, the length of ICU stay models using ZIP and ZIBN exhibit no significance in terms of the zero-inflation portions of the models. This is most likely due to the low proportion of zero counts in this data set. Again, it is important to examine the zero-inflation models to eliminate the possible of zero counts adversely affecting model fit.

Table 9

ZIP and ZIBN Models for ICU Stay

ZIP Poisson						
	Count Coef.	p-value	SE	Zero-Inflation Coef.	p-value	SE
CI0	-0.0050	1.94E-05	0.0012	0.0951	0.2203	0.078
CI72	0.0163	2E-16	0.0012	0.0382	0.593	0.535
CIdiff	0.0215	2.00E-16	0.0012	-0.0587	0.486	0.0844
ZIBN Negative Binomial						
	Count Coef.	p-value	SE	Zero-Inflation Coef.	p-value	SE
CI0	-0.0051	0.115	0.0033	-0.0459	0.999	106.298
CI72	0.0172	2.38E-07	0.0033	-0.0783	0.998	31.476
CIdiff	0.0227	1.316E-11	0.0033	-0.201	0.937	2.529

This table gives the coefficients, p-values and standard errors for serum chloride levels using each model. Both the count and the zero inflation portions of each model are shown.

CHAPTER 6

CONCLUSION AND DISCUSSION

Examining the effect of Hyperchloremia on mortality, it can be shown that increased levels of serum chloride at admission and after 72 hours increase the risk of mortality. As this data was a smaller set than may be available in the future, more tests will be needed to verify these results. Some of the data provided here had some key variables missing and these may be used in future work with larger data sets.

Overall, Hyperchloremia is associated with increased in-hospital mortality in critically ill patients. We found an independent association between higher CI72 and in-hospital mortality in critically ill patients. Most mortality CI72 was associated with hospital mortality in those patients who were already hyperchloremic on ICU admission.

The data also showed increases in length of hospital stay and ICU stay with high admission serum chloride levels and high levels after 72 hours. It was found that the ZIP and ZINB regression models both have some problems in assessing LOS and ICU stay in critically ill patients, especially in the presence of excess zeros and overdispersion in count data. The strengths of our study are the large sample size (>1700), the careful selection of a representative sample of patients with critical ill admitted to the ICU, and the multivariate adjustment for clinical confounders directly linked to hyperchloremia and hospital mortality such as AKI, and comprehensive critical illness severity scores.

None of the studies that have previously revealed the association between serum chloride levels and hospital mortality accounted for confounding. Our study is unique in the multivariate design and patient population.

CHAPTER 7

FUTURE STUDY

To better understand the potential issues involving medical count data and to investigate modeling where multiple factors and covariates are present, the following section presents alternative methods for handling count data.

7.1 Modeling Excess Overdispersion: The Hurdle Poisson Model

The data provided thus far for studying the effects of serum chloride on hospital and ICU length of stay have not accounted for factors such as regional or geographic location. A future study may be needed to identify patients by region to better understand the full effects on patients. To incorporate these regional differences, studies using the Hurdle model may be employed. As length of hospital stay is usually characterized as right-skewed for the non-zero values and possibly zero-inflated or zero-deflated, the Hurdle model was developed to account for these and other issues regarding count data. The Hurdle model accommodates other count data issues such as spatial random effects and other fixed-effects covariates.

The Hurdle model is based on a mixture of zero mass and non-zero count observations following either Poisson or negative binomial distributions. The Hurdle model considers the zeros to be completely separate from the non-zeros.

Let Y_{ij} denote the length of stay in days, $i = 1, 2, ...n$ and j denote patient region, $j = 1, 2, ...J$. The model is given as follows:

$$P(Y_{ij}) = \begin{cases} \pi_{ij} & y_{ij} = 0, \\ (1 - \pi_{ij}) \frac{p(y_{ij};\theta_{ij})}{1-p(0;\theta_{ij})} & y_{ij} > 0 \end{cases} \tag{13}$$

where $\pi_{ij} = P\ (Y_{ij} = 0)$ is the probability of a patient belonging to the zero component and $p(y_{ij}; \theta_{ij})$ is the probability distribution for count data. The parameter vector is θ_{ij} and $p(0; \theta_{ij})$ is the evaluation at zero.

The probability distribution for the Hurdle Poisson is:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & y_{ij} = 0, \\ (1 - \pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{1-e^{-\mu_{ij}}} & y_{ij} > 0 \end{cases} \tag{14}$$

The probability density function for the Hurdle negative binomial model is:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & y_{ij} = 0, \\ \frac{1-\pi_{ij}}{1-(\frac{r}{\mu_{ij}+r})^r} \frac{\Gamma(y_{ij}+r)}{\Gamma(r) y_{ij}!} (\frac{\mu_{ij}}{\mu_{ij}+r})^{y_{ij}} (\frac{r}{\mu_{ij}+r})^r & y_{ij} > 0 \end{cases} \tag{15}$$

where $(1 + \mu_{ij} /r)$ measures overdispersion and as $r \to \infty$, the negative binomial converges to the Poisson distribution.

REFERENCES

A. Zuur, N. W., E. Leno, and Smith, G. (2009), Mixed effects models and extensions in ecology with r, NY: Springer.

Auer, P., and Doerge, R. (2012), "A two-stage poisson model for testing rna-seq," Statistical Applications in Genetics and Molecular Biology, 10, 26.

Cox, D. (1983), "Some remarks on overdispersion," Biometrics, 10, 269–274.

D.J. McCarthy, M. R., and Smyth, G. (2010), "A bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, 26, 139–140.

Dean, C. (1992), "Testing for overdispersion in poisson and binomial regression models," Journal of the American Statistical Association, 87, 451–457.

Famoye, F., and McGwin, G. (2002), "Regression analysis of count data," Indian Socitey of Agricultural Statistics, 55, 220–231.

Greene, W. (1994), "Accounting for excess zeros and sample selection in poisson and negative binomial regression models. working paper ec-94-10," Leonard N. Stern School of Business, New York University, EC-94-10.

Hilbe, J. (2011), Negative binomial regression, Cambridge University Press.

J. Silva, E. N., and Santana, T. (2009), "The importance of intraoperative hyperchloremia," Revista Brasileira de Anestesiologia, 59, 304–313.

K. Berend, R. G., L.H. van Hulsteijn (2012), "Chloride: The queen of electrolytes?" European Journal of Internal Medicine, 23, 203–211.

K. Yau, A. L., and Wulu, J. (2003), "Modelling inpatient length of stay by a hierarchical mixture regression via the em algorithm," Mathematical and Computer Modelling, 37, 365–375.

Karlis, D., and Xekalaki, E. (2005), "Mixed poisson distributions," International Statististical Review, 73, 35–38.

King, G. (1989), "Event count models for international relations: Generalizations and applications," International Studies Quarterly, 33, 123–147.

Lambert, D. (1992), "Zero-inflated poisson regression with an application to defects in manufacturing," Technometrics, 34, 1–14.

Lindsey, J. (1995), Modelling Frequency and Count Data, Oxford University Press.

Lindsey, J., and Altham, P. (1998), "Analysis of the human sex ratio using overdispersion models," Applied Statistics, 47, 147–157.

M. Robinson, D. M., and Smyth, G. (2009), "Impact of normal saline infusion on postoperative metabolic acidosis," Paediatric Anaesthesia, 19, 1070–1077.

Mullahy, J. (1986), "Specifications and testing of some modified count data model," Journal of Econometrics, 33, 341–365.

N. Yunos, R. B., and Story, D. (2010), "Bench-to-bedside review: Chloride in critical illness," Critical Care, 14, 226.

Prentice, R. (1986), "Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors," Journal of the American Statistical Association, 81, 321–327.

S. McCluskey, K. K., and Wijeysundera, D. (2013), "Hyperchloremia after noncardiac surgery is independently associated with increased morbidity and mortality: A propensity-matched cohort study," Anesthesia & Analgesia, 117, 412–421.

S. Pounds, C. G., and Zhang, H. (2012), "Empirical bayesian selection of hypothesis testing procedures for analysis of sequence count expression data," Statistical Applications in Genetics and Molecular Biology, 11, 7.

S. Scheingraber, M. R., and Sehmisch, C. (1999), "Rapid saline infusion produces hyperchloremic acidosis in patients undergoing gynecologic surgery," Anesthesiology, 90, 1265–1270.

VITA

Lisa Nanni was born in Grosse Pointe, MI. She attended the Florida Institute of Technology and earned a Bachelor of Science in Electrical Engineering in 1990. She worked as an engineer for General Dynamics before beginning a teaching career at the Jackson Community College from 1999 to 2011. She served on committees, was a course coordinator and won the instructor of the year award in 2005. In 2010, she earned a Master of Arts in Mathematics at Eastern Michigan University. Lisa relocated to Georgia where she has been teaching Mathematics at Georgia Northwestern Technical College and currently at the University of Tennessee at Chattanooga.