Transnational University Limburg is a unique collaboration of two universities in two countries:
the University of Hasselt and Maastricht University.

KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Persistence of resistance selection by common antibiotic substances in streptococci: a case-control study as surrogate for a randomized placebo controlled trial

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof.dr. SAMUEL COENEN

Abdalla Mtumwa

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

2012•2013

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Persistence of resistance selection by common antibiotic substances in streptococci: a case-control study as surrogate for a randomized placebo controlled trial

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof.dr. SAMUEL COENEN

Abdalla Mtumwa

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Acknowledgement

First and foremost I would like to give my heartfelt gratitude to the Almighty God for His Grace, help, protection and blessing which he has given me through all the time. He is the only and everything to my success. I would also like to sincerely thank my internal supervisor Professor Niel Hens for his support, encouragement, critical comments and constructive ideas which resulted to this work. Next, my gratitude also goes to my external supervisors Prof. Samuel Coenen, Mr. Boudewijn Catry and Katrien Latour. I express my deep feelings and gratitude to them not only for supervising my work but also for many more helpful suggestions they extended me without which it would have not been possible for me to complete my thesis.

I'm deeply grateful to the Flemish Interuniversity Council (VLIR) for the scholarship which has enabled me to pursue this valuable Masters program in Universiteit Hasselt. Without forgetting Martine Machiels who was not only a program manager but also a mother of the international students. I would like to thank all of my respected teachers and staffs of the Censtat, UHasselt. Not forget to express my especial thanks to Alex Sila, Kelvin Mchau, Pendael Machafuko, David Amonya, Gilbert Rukundo, Josephine Shabani and Mohammad Romel.

Finally, I would like to express my heartiest honour to my parents, and my wife Saumu Hilal Moh’d , their encouragement enabled me to finish this thesis.

Abdalla Hussein Mtumwa

Contents

LIST OF TABLES

Abstract

Background: Disease-causing microbes that have become resistant to antibiotic drug therapy are an increasing public health problem. Antibiotic use has been identified as the major driver of resistance selection both at ecological and at individual patient level. To assess the effect of previous antibiotic use (*macrolides/tetracyclines* and *penicillin/cephalosporins*) on the risk of resistant pathogenic bacteria respiratory samples (*streptococci*) and to compare with results obtained from earlier randomized controlled trials, we undertook a case-control study with prospective measurement of outcomes in 15 voluntary diagnostic laboratories in Belgium during the year 2005.

Methods: Respiratory samples were taken from the patients, those with streptococci their medical records were examined. Case patients were those with macrolides or penicillin resistant infections (including intermediate) and control patients had infections that were susceptible to antibiotics.

Results: After carefully having taken into account inclusion (e.g. prescription data obtained for the respective patients) and exclusion criteria, a total of 640 observations from 595 patients were retained in the final dataset. Herein, 37.35 % of 83 observation tested for macrolides (prior use of macrolides/tetracyclines) developed resistance while 15.62 % of 557 tested for penicillin (prior use of penicillin/cephalosporins) were resistant. Since of all patients included 94.29 % had only one observation, the analysis was done in two parts. Firstly, a multiple logistic regression model was fitted for those patients having only one observation. The result of this model showed that controlling for *timecat* effect in the model, patients with resistant streptococci, were significantly more likely to have been prescribed *macrolides/tetracyclines* compared to *penicillin/cephalosporins* (4.024, 95% confidence interval (CI): 2.142- 7.559). For the second part of the data, Generalized Estimating Equation (GEE) models with exchangeable working correlation and Generalized Linear Mixed Model (GLMM) with random intercepts were used to model the binary response (resistance versus not resistance).

Conclusions: Within the population setting, exposure to macrolides/tetracyclines increased the risk for resistant streptococci as *compared to* penicillin/cephalosporins while in cluster (patient) level no covariate found to be associated with susceptibility results of streptococci.

Keywords: Streptococci, Antibiotics, Resistance, EB GEE, CWGEE, GLMM

1 Introduction

1.1 Background

In the past 60 years, antibiotics have been critical in the fight against infectious diseases caused by bacteria and other microbes. Antimicrobial chemotherapy has been a leading cause for the dramatic rise of average life expectancy in the Twentieth Century. However, disease-causing microbes that have become resistant to antibiotic drug therapy are an increasing public health problem. One part of the problem is that bacteria and other microbes that cause infections are remarkably resilient and have developed several ways to resist antibiotics and other antimicrobial drugs. Another part of the problem is due to increasing use, and in appropriate prescriptions can elicit antibiotic-resistant bacteria in human medicine (Mellon et al., 2000).

In 1998, in the United States, 80 million prescriptions of antibiotics for human use were added. This equals 12,500 tons in one year animal, aqua and horticultural uses of antibiotics are added to human use. These agricultural practices account for over 70% of antibiotic usage in the U.S, so this adds an additional 18,000 tons per year to the antibiotic burden in the environment

(Mellon et al., 2000). An alarming increase in resistance of bacteria that cause community acquired infections has also been documented, especially in the staphylococci and pneumococci (*Streptococcus pneumoniae*), which are prevalent causes of disease and mortality. Microbial development of resistance, as well as economic incentives, has resulted in research and development in the search for new antibiotics in order to maintain a pool of effective drugs at all times. While the development of resistant strains is inevitable, the slack ways that we administer and use antibiotics has greatly exacerbated the process. Unless antibiotic resistance problems are detected as they emerge, and actions are taken immediately to contain them, society could be faced with previously treatable diseases that have become again untreatable, as in the days before antibiotics were developed (Todar, K., 2009).

The first antibiotic, penicillin, was discovered in 1929 by Sir Alexander Fleming, who observed inhibition of staphylococci on an agar plate contaminated by a Penicillium mold. By accident, Fleming was discovering an extremely useful and safe antibacterial compound. He noticed that a patch of the mold *Penicillium notatum* had grown on a plate containing the bacterium Staphylococcus and that around the mold there was a zone where no *Staphylococcus* could grow. After more research, he was able to show that culture broth of the mold prevented growth of the *Staphylococcus* even when diluted up to 800 times. He named the active substance penicillin but was unable to isolate it.

Several years later, in 1939, Ernst Chain and Howard Florey developed a way to isolate penicillin and used it to treat bacterial infections during the Second World War. The new drug came into clinical usage in 1946 and made a huge impact on public health. In 1946, penicillin became generally available for treatment of bacterial infections, especially those caused by staphylococci and streptococci. Initially, the antibiotic was effective against all sorts of infections caused by these two Gram-positive bacteria. Penicillin had unbelievable ability to kill these bacterial pathogens without harming the host that harbored them. It is important to note that a significant fraction of all human infections are caused by these two bacteria (i.e., strep throat, pneumonia, scarlet fever, septicemia, skin infections, wound infections, etc.).

There has probably been a gene pool in nature for resistance to antibiotics long before antibiotic production, for most microbes are antibiotic producers and intrinsic resistant to their own antibiotic. In retrospect, it is not surprising that resistance to penicillin in some strains of staphylococci was recognized almost immediately after introduction of the drug in 1946. Likewise, very soon after their introduction in the late 1940s, resistance to streptomycin, chloramphenicol and tetracycline was noted. By 1953, during a Shigella outbreak in Japan, a strain of the dysentery bacillus (*Shigella dysenteriae*) was isolated which was multiple drug resistant, exhibiting resistances to chloramphenicol, tetracycline, streptomycin and the sulfonamides. Over the years, and continuing into the present almost every known bacterial pathogen has developed resistance to one or more antibiotics in clinical use.

The impact of antibiotic use on antibiotic resistance in oropharyngeal streptococcal flora of individuals has been assessed in randomized placebo controlled trials(RCTs), one on macrolides (*clarithromycin and azithromycin*) in healthy volunteers and one on amoxicillin in adult patients presenting in primary care with acute cough (Malhotra-Kumar et al., 2007). It was shown that persistence of resistance selection after exposure to macrolides lasts more than 6 months, while it is estimated to be much shorter following amoxicillin use. The objective of this study is to assess persistence of resistance selection of common antibiotic substances (*macrolides/tetracyclines* and *penicillin/cephalosporins)* among *streptococci* isolated from respiratory samples based on a large dataset from a case-control study, and to compare the case-control study results with the RCT results for *macrolides* and *amoxicillin (penicillin).*

2 Data

The significance of any research depends on using a reliable source of data. This section provides a brief description of the data, analysis plan and other related issues of the study. Briefly, laboratory results of were obtained from 15 voluntary diagnostic laboratories in Belgium during the year 2005 in collaboration with the Intermutualistic Agency (IMA) coordinating the national health insurance funds. Susceptibility profiles were predominantly based on the Kirby Bauer disk diffusion test according to CLSI guidelines and interpretation was often recorded by semi-automated systems. All participating labs were certified by an external quality control organization. Data were coupled with antimicrobial prescription details and patient charact-eristics as outlined below. Further details on data collection have been outlined underneath and additional information can be found elsewhere (Catry et al., 2008).

2.1 Variables Description

In this study, covariates listed in Table 1 were available to explain the response variables.

Table 1. *Description of the variables in the study*

Variable	Label	Description
Age	Age	Age (in years) of the patient
agecat	1.[0-14]years 2.15-54]years ≥ 55 years	Categorized version of age years
Patient- Sex	1.Male 2.Female	Sex of a patient
n_prescription	n_prescription	the number of times exposed to antibiotics (prescriptions) in days before the sample date
n_cat	1. ≤ 2 days 2. >2 days	Dichotomized version of the n_prescription
Treat	1. Macrolides 0. Penicellins	macrolides and tetracyclines were combined penicillins and cephalosporins were combined
Timecat	1. (0-60] days 2.>60 days	Dichotomized version of time in days between the prescription date and sample date
Result	1.R=Resistant 0.S=Susceptible	The response variable which is susceptibility test result (intermediate was set as resistant)

2.2 Data Description

In this study, data was collected from multicenter for studying the association between previous antibiotic use and resistance in the individual patient (see above). Starting from this dataset, two new databases were created*:'streptopatient'and'strepto-prescription data'*. The former contained 4738 patients who delivered at least one sample with a *Streptococcus* in it and the latter contained 4227 patients. Susceptibility testing for an antibiotics was performed by the Kirby-Bauer disk diffusion method which relies on the inhibition of bacterial growth measured under standard conditions and the boundaries of susceptibility classification (resistant(R), intermediate (I) or susceptible (S)) are defined based on the Clinical and Laboratory Standards Institute (CLSI, 2006).The microbiological results retrieved from 15 voluntary participating Belgian clinical microbiological laboratories between periods 1st July and 31st December 2005 were merged with the individual antibiotic consumption patterns. In this study, *macrolides-/tetracyclines*, *quinolone and penicillin/cephalosporins* use before the sample date, and the susceptibility of *streptococci* found in respiratory tract samples (including upper respiratory tract, upper respiratory tract (ear), lower respiratory tract and lower respiratory tract (sputum) samples) for these antibiotics were considered as they were our primary interest. The initial *streptopatient* database, obtained from the Scientific Institute of Public Health (WIV-ISP, Brussels), contained 1671 respiratory track samples of interest from 1534 patients, implying that one patient could have had more than one sample during the study period. Each sample was tested for the presence of bacteria and the susceptibility of each found bacterium was tested for different antimicrobials, resulting in 18292 records in the database. Susceptibility tests results in these samples with streptococci were 10477(57.28%) susceptible, 600(3.28%) intermediate, 1386(7.58%) resistant, and 5829(31.87%) were not reported. For this study, only a part of the general database was used. From the *Strepto prescription* database only patients from pharmacy with oral admini-stration route and prior use of *penicillins*, *cephalosporins* , *macrolides* or *tetracyclines* were considered while from the *streptopatient* , for those samples of interest taken from respiratory tract, patients who had samples with *Streptococcus* B, *Streptococcus* A, *Streptococcus* C, *Streptococcus* D, *Streptococcus* F, *Streptococcus* G, *Streptococcus pneumonia* or *Streptococcus- pyogenes* , and susceptibility of these bacteria tested for *macrolides* and *penicillin* were considered. Since there were few patients contributing for the information of quinolone use and quinolone resistance test results, they were dropped in this study. Also susceptibility results not reported for the other compounds of interest (*penicillins, cephalosporings, macrolids,* and *tetracyclines*) were dropped whereas an intermediate susceptibility result was considered as resistant. Patients with resistant infections (including intermediate) to macrolides or penicillin were considered as cases and control patients had infections that were susceptible to antibiotics. Out of 1027 remained respiratory tract samples, 169 were tested for *macrolides* (prior use of *macrolides/tetracyclines*) and 955 for *penicillin* (prior use of *penicillin-/cephalosporins*). For 169 samples tested for macrolides resulted to 191 observatio-ns, in which 66(34.55%) were resistant. While the remaining 955 samples tested for penicillin resulted to 2600 cases with 398(15.31%) of them were resistant. Moreover, 548 were male and 409 female patients.

3 Methodology

3.1 Flow of Analyses

This section provides the analysis plan and procedure to address the specific objectives. Firstly, in section 4.1, data exploration techniques are presented to review the data structure. Secondly, in section 4.2, multiple logistic regression analysis was carried out to identify the primary important risk factors for the first part of the study for those patients each having only one observation. Afterwards, section 4.2.1, logistic regression analysis was carried out for the second part of the study with more than one sample per patient ignoring the clustering effect in the data. Finally, the sections 4.2.2, 4.2.3,4.2.3and 4.2.5 discuss two basic approaches which take the clustering effect into account: one using a population-averaged method, more specifically the generalized estimating equation (GEE) model; another is the generalized linear mixed model (GLMM), a random effect model.

3.2 Exploratory Data Analysis (EDA)

This fundamental step has been carried out in order to gain better insight into the data set. Simple descriptive statistics (cross-tabulation) and unadjusted odds ratio were mainly used to study the association between the response variables and the set of explanatory variables.

3.2 Statistical Analysis

3.2.1 Multiple Logistic Regression Model

Regression analysis can be used to assess the effect and relationship between independent (explanatory) variables and response variable. Many types of regression analysis exist and logistic regression is one of them. This type of model is used in studying the relationship between a binary response variable with a set of predictors which can either be quantitative or qualitative. Deciding which covariates to be kept in the statistical model is not an easy task for data analysts. Examination of each covariate with the response variable can provide a preliminary idea how important the variable is. Consequently, a univariate logistic regression model was fitted and variables with p-value < 0.25 were considered as candidates for the multiple logistic regression models (Hosmer and Lemeshow, 2000). Agresti (2002) stressed that "Statistical significance" is not the only reason to keep a covariate in a model. Other variables known to be important, but not significant could be included in model. Since the response variable (susceptibility result) was dichotomized (R or S), multiple logistic regressions was used to identify risk factors and predict the probability of success. This model belongs to a family of generalized linear models under the assumption of binomial distribution of the response. Various link functions such as logit, probit e.t.c are permissible. Letting y_i be the binary response and $x_i = (x_1, x_2, ..., x_p)$ as the explanatory variables $\pi(x) = P(y_i = 1 | X = x_i)$ is the probability of success in this case a resistant result (R) was success. The model is thus given by $f(y_i | x_i)$ is given by

$g\{\pi(x)\} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ .Where g(.) is the link function and β_p's being the maximum likelihood parameter estimates. The backward selection procedure was used to build the model to identify the primary important risk factors. Eventually, variables with p-value < 0.05 were retained for further statistical analysis. For the second part of the study each patient had more than one sample. Hence, the traditional standard error estimates for logistic regression models based on maximum likelihood from independent observations is no longer appropriate since observations in the same clusters (patients) tend to have similar characteristics and are more likely correlated with each other. The variance of y_i in the binary case can then be inflated (Agresti, 2002). Hence ignoring clustering in analyses may exaggerate the precision, so risk factors are reported as significant even when this may not be correct(Bennett et al., 1991; Faes et al., 2006).The fact that, multiple logistic regression assumes multiple observations coming from the same patient are independent, ignores the intraclass correlation (correlation within repeated measurements). The appropriate model was fitted using Generalized Estimating Equations (GEEs).

3.2.2 Marginal Models (GEE)

The term marginal in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or other responses. The marginal model is used when the researcher investigates the overall”population–average“ trend as a function of the covariates while accounting for the correlations in the data. The association structure is then typically captured using a set of association parameters, such as correlations, odds ratios, etc (Molenberghs and Verbeke, 2005). Within the marginal model family there exist several methods that may be either quasi-likelihood or full likelihood. Full likelihood methods include the Bahadur, dale and probit models. These models are beneficial in term of efficiency, maximum likelihood can be unattractive due to excessive computational requirements, especially when high dimensional vectors of correlated data arise and of course, increase the risk of model misspecification (Molenberghs and Verbeke, 2005). As a consequence, alternative methods have been in demand. When we are mainly interested in the first-order marginal mean parameters and pairwise interactions, a full likelihood procedure can be replaced by quasi-likelihood based methods (McCullagh and Nelder,1989).These include Generalized Estimating Equations (GEE) proposed by Liang and Zeger(1986), Prentice (GEE) and Alternating Logistic Regression(ALR). Prentice (1988) extended GEE to allow joint estimation of probabilities and pair wise correlations. Later, modification was performed to allow modeling of the association through odds ratios (OR) rather than marginal correlation through ALR. GEE proposed by Liang and Zeger (1986), require only the correct specification of the univariate marginal distributions provided one is willing to adopt “working” assumptions about the association structure. The existence of clustering is recognized but considered a nuisance characteristic. The essential idea behind the GEE approach is to generalize and extend the usual likelihood equations for a generalized linear model for a univariate response by incorporating the covariance matrix of the vector of responses. Generalized estimating equations do not use the information on the association structure in estimating regression coefficients and hence it gives consistent and asymptotically normal main effect estimators even when the association structure is misspecified. Various working correlations can be assumed, which include independence, exchangeability, autoregressive and unstructured. Autoregressive can be used only for equally spaced and exchangeable can be used for unequally spaced and unbalanced data. However unstructured covariance structure appears suitable when the number of repeated measurements is small and is balanced (equal) across individuals (Liang and Zeger, 1986). Nonetheless in this particular case, due to the unequally samples per patients auto regressive as well as unstructured structures are not suitable options, therefore exchangeability was preferred for analysis. In this thesis, GEE was applied to assess the average trend of previous antibiotic used and susceptibility result of streptococci. Suppose that Y_{ij} is a binary response, taking the value of 0(denoting ‘failure’) for our case susceptible or 1 denoting success (resistant), and it is of interest to relate change in $E(Y_{ij}) = \Pr(Y_{ij} = 1)$ to the covariate. With binary response, the distribution of each Y_{ij} is Bernoulli and the probability of success is often modelled using a logit or probit link function. The marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on the covariates X_{ij} , through a known link function

$$g(\mu_{ij}) = X'_{ij}\beta.$$

The GEE estimator of β for marginal models can be thought of as arising from minimizing the following objective function:

$$\sum_{i=1}^{N}\{y_i - \mu_i(\beta)\}'V^{-1}\{y_i - \mu_i(\beta)\},........(1)$$

with respect to β, where $V_i = A_i^{\frac{1}{2}} Corr(Y_i) A_i^{\frac{1}{2}}$ is the marginal covariance matrix of Y_i which is treated as known (by ignoring its dependence on β through μ_i), $\mu_i = \mu_i(\beta) = X_i\beta$ is the vector of mean responses and $Corr(Y_I)$ is the marginal correlation matrix. Using calculus it can be shown that if a minimum of the function given by (1) exists, then the regression parameters β are estimated by solving the estimating equations $\sum_{i=1}^{N}\frac{\partial\mu_i}{\partial\beta}V^{-1}(y_i - \mu_i(\beta)) = 0$

3.2.3 Cluster Weighted Generalized Estimating Equations (CWGEE)

In the GEE approach, the correlation in the working covariance matrix can be seen as weights that are assigned to the data from each cluster. If the outcome measured among cluster members is independent of cluster size (i.e., if cluster size is uninformative), clustering only enters the analysis to obtain a valid variance estimates (using the sandwich variance estimator). The inference is valid even if the working correlation is misspecified. However, when cluster size is informative (cluster size is related to the outcome of interest), then weighting the data in different ways, as is done by choosing different working correlations, can result in different marginal models. In a GEE model with independent working correlations, each observation is given the same weight. In that case, the choice of a working correlation matrix becomes an important issue and inappropriate choice resulting in misleading and biased parameter estimates (Williamson et al., 2003; Aerts et al., 2010). Williamson et al. (2003) present a modification of the generalized estimating equations (GEE) for handling binary response data with informative cluster size. They propose the use of weighted generalized estimating equations, where the contribution to the estimating equation from a cluster is weighted by the inverse of the cluster size, with an independence working correlation matrix. As a result, all clusters are given equal weight and individuals in large clusters are no longer over weighted. The marginal parameter in the cluster-weighted GEE (CWGEE) will have a cluster-based interpretation. This is in direct contrast to GEE, where large clusters are weighted more than small clusters. The CWGEE seems to be a nice alternative to the GEE in case of informative cluster sizes. However, the performance of the

CWGEE with a working correlation matrix different from the independence working correlation is not yet understood.

3.2.4 Random Effect Model (GLMM)

In the previous sections, we discussed how marginal models can be considered as an extension of generalized linear models that directly incorporate the within-cluster association among the repeated measurements. In a certain sense, marginal models account for the consequences of the correlation among the repeated measures but do not provide any explanation for its potential source (Fitzmaurice et al., 2004). An alternative approach for accounting for the within-cluster association, and one that provides a source for the within-cluster association is via the introduction of the random effects in the model for the mean response. In this section, a generalized linear mixed model (GLMM) is used to model the risk of bacteria to be resistant to an antibiotic. In GLMMs, the model for the mean response is conditional upon both measured covariates and unobserved random effects; it is the inclusion of the latter that induce correlation among the repeated response marginally, when averaged over the random effects. However, with non-linear link functions, the introduction of random effects has important ramifications for the interpretation of the "fixed-effect" regression parameters. In a random effects model, it is assumed that there is natural heterogeneity across the clusters. This heterogeneity can be modeled by a probability distribution which implies that the regression coefficients are varying from one cluster to another. Conditionally on random effects for each cluster, it is assumed that the cluster-level outcomes are independently distributed as:

$$Y_{ij} / b_i \sim Benoulli(\pi_{ij})$$

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = X_{ij}'\beta + Z_{ij}'b_i$$

Where Y_{ij} is the j-th outcome observed for cluster *i, i =1... N, j = 1...* n_i . b_i is a random vector which is assumed to be normally distributed with mean vector 0 and covariance matrix D. X_{ij} and Z_{ij} are $(n_i \times p)$ and $(n_i \times q)$ dimensional vectors of known covariates. Similarly, β is a p-dimensional vector of unknown fixed effect regression parameters.

3.3 Model Building

Statistical model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables. The major issue in building a statistical model is selecting which independent variables to be included in the model. The objective of model building is to find an optimal model characterized by principles of parsimony and goodness-of-fit based on model selection criteria. Model selection criteria are statistical tools help to find the best fitting model for the data on hand among the set of candidates models. Several criteria for selecting subset of covariates that describe the data optimally differ from likelihood to non-likelihood estimation methods. The Akaikes Information Criterion (AIC) is widely used model selection criterion when the likelihood function is fully specified. When the likelihood function is not fully specified, e.g., as in the GEE, the AIC can no longer be used. An alternative model selection technique QIC based on the quasi-likelihood function (McCullagh and Nelder 1989) can be used instead. The smaller the QIC is, the better the model. In order to obtain the parsimonious working correlation, the discrepancies in standard errors (empirical and model based standard errors) were checked and the correlation that resulted to the least discrepancy was considered.

3.4 Software

The statistical packages SAS (version 9.3) were used to analyze the data. A 5% level of significance was used throughout the study

4 RESULTS

4.1 Exploratory Data Analysis

The refined data set combined the information obtained from the antibiotic result test data base (*streptopatient*) and antibiotic consumption data base (*strepto-prescription).*This data base contains patients with prior use of *macrolides*, *penicillin*, *tetracyclines* or *cephalosporins* and considered all respiratory tract samples and their respective prescription date was found to be from July 2, 2004 to Dec 16, 2005 which resulted in 640 observations obtained from 595 patients. Table 2 below shows Susceptibility results of streptococci and number of patients with different Characteristics of the refined database. In this database 412 patients used penicillin,118 (cephalosporins),14(tetracyclines) and 63 patients used macrolides before the sample date. Looking for macrolide resistance we are mainly interested in prior use of macrolides and tetracyclines respectively. On the other hand, for *penicillin* resistance we are mainly interested in prior use of penicillin and cephalosporins. Therefore the final analysis combined the information of (i)macrolides and tetracyclines use in one group called '*macrolides/tetracyclines'* to look at macrolides resistance and(ii) penicillin and cephalosporins in another group called '*penicillin/ cephalosporins'* to look at penicillin resistance. Of all patients, 345(57.98%) were males. In addition, there were 118 observations with *streptococci* resistance, where by 66 were resulted from patients used *penicillin,* 21 from *Cephalosporin*s user, 22 from *Macrolides* user and 9 from *Tetracyclines* user. Among all observations with resistant *streptococci* 67 were from males and 51 from females. Of 595 patients, 561(94.29 %) have only one observation. The maximum number of observations per patient was 6 and this was found only to one patient. Due to large proportion of patients with only one observation the analysis of the study will be done in two separate parts: Firstly, for those patients with only one observation and the second part of the analysis will consider patients with more than one observation.

Table 2: *Susceptibility results of streptococci and number of patients with different characteristics of the refined database.*

	Number of patients with streptococci by patient sex(n=595)		
	Frequency	Percent(%)	
Males	345	57.98	
Females	250	42.02	
	Number of patients with streptococci by antibiotic use		
Penicillin	412	69.24	
Cephalosporins	118	19.83	
Macrolides	63	10.59	
Tetracyclines	14	2.35	
	Susceptibility result of streptococci by antibiotics use (n=640)		
	Resistant(R)	Susceptible(S)	Total
Penicillin	66(15.03%)	373(84.97%)	439
Cephalosporins	21(17.8%)	97(82.2%)	118
Macrolides	22(34.38%)	42(65.62%)	64
Tetracyclines	9(47.37%)	10(52.63%)	19
	Susceptibility result of streptococci by patient sex(n=640)		
	Resistant(R)	Susceptible(S)	Total
Males	67(17.68%)	312(82.32%)	379
Females	51(19.54%)	210(80.46%)	261

Table 3 below, displays the summary statistics of the variable time (days) between the prescription and sample date .On average ,those patients tested for *macrolides* resistance had larger time difference between the prescription and sample date as compared to *penicillin*. For example, the average time for patients tested for macrolides and having susceptible streptococci result was 194 days while for those under penicillin was 109 days. Also, either the streptococci was tested for *macrolides* or *penicillin*, those with resistant streptococci had less mean average time as compared to those with susceptible streptococci result. For example, for those tested with penicillin the average time for those with resistant and susceptible streptococci was 63 and 109 respectively. This means that those samples taken within small time from the prescription date are more likely to have streptococci resistance as compared to susceptible one.

Table3: *Summary statistics of time (days) between the prescription and sample date by Susceptibility result and antibiotic tested from 595 patients in refined database.*

Result	Antibiotic tested	Min	Max	Mean	N
Resistant(R)	macrolides	5	414	179	31
	penicillin	1	326	63	87
Susceptible(S)	macrolides	6	426	194	52
	penicillin	1	521	109	470

Of the 561 patients each having only one observation, 95(16.9%) had streptococci resistance to either macrolides or penicillin and were defined the cases for this part of the study and 466(83.1%) were controls. 493(87.9%) had been prescribed *penicillin/cephalosporines*. Two hundred and thirty two of the patients were females. Patients with streptococci resistance were significantly more likely to have been prescribed *macrolides/tetracyclines* in previous time before the sample date as compared to *penicillin/cephalosporines* (OR =2.530, 95% confidence interval (CI):1.430-4.477). For patients with length of time between the most recent prescription date and the sample date being more than 60 days were less likely to have resistance as compared to those prescribed with 60 days (OR=0.475, 95% confidence interval (CI): 0.302-0.747). The OR of male to have streptococci resistance as compared to female was 0.745(CL: 0.478-1.160). Also 121 patients were prescribed antibiotics for more than 2 days before the sample date and their OR of being resistance as compared to those patients that were prescribed not more than 2 days was 1.676 (CL: 1.021-2.753). Moreover age might not associate with the risk of streptococci resistance.

Table 4. *The risk of streptococci resistance by different variables for 561 patients with one observation.*

variable	category	Resistant	susceptible	OR	95% CL	
SEX	Females	45	187	ref		
	Males	50	279	0.745	0.478	1.160
Prescription (treat)	Pen/ceph	74	419	ref		
	Mac/tetra	21	47	2.530	**1.430**	**4.477**
Timecategory	(0-60]days	51	204	ref		
(timecat)	>60 days	36	262	0.475	**0.302**	**0.747**
	<-2 days	67	373	ref		
(n_pcat)	>2 days	28	93	1.676	**1.021**	**2.753**
	[0-14]years	43	249	ref		
Agecategory	[15-54]years	12	70	0.993	0.497	1.984
(agecat)	≥ 55 years	40	149	1.576	0.979	2.538

Pen/ceph: *penicillin/cephalosporines, Mac/***tetra:** *macrolides/tetracyclines*

The top panel of table5 below, displays the summary statistics of age (years) and number of prescription (days) of antibiotics for the 561 patients with only one observation. Since the distributions of these variables were skewed it is better to use median and interquartile range for description. The median age of the participant patients was 11 years and the interquartile range of age was found to be 61 years. The median number of prescription of antibiotic a patient had, was 1 days with interquartile range of 1. The minimum age and number of prescription was o and 1 respectively, while their respective maximum was 97 and 23. For the second part of the study for those 34 patients with more than one observation, the median age was 35.5 years with interquartile range of 64. Also the median number of prescription of antibiotic a patient had, was 2 days with interquartile range of 1.The summary for the second part of the study are shown in bottom panel of table5.

Table 5. *Summary statistics for age, and number of prior prescriptions of antibiotics*

Variable	Median	Interquatile Range	Minimum	Maximum
561 patients with only one observation				
Age	11	61	0	97
n_prescription	1	1	1	23
34 patients with more than one observation				
Age	35.5	64	1	81
n_prescription	2	2	1	9

4.2 Statistical Analysis

The objective of this study is to assess persistence of resistance selection of common antibiotic substances (*macrolides/tetracyclines* and *penicillin/cephalosporines*) among pathogenic bacteria respiratory samples (*streptococci*) based on a large dataset from a retrospective cohort study, and to compare with results obtained from earlier randomized controlled trials. These objectives were reformulated using the following hypotheses:

The primary hypothesis can be stated as the odds of having *streptococci* resistant result, compared with an antibiotic-susceptible, would be modified by type prior use of the antibiotic. A secondary hypothesis was that the odds would depend on other covariates. Depending on the nature of the data these hypotheses were assessed by calculating adjusted ORs using multiple logistic regression models for those patients with only one observation and by performing clustered data analysis for those with more than one observation.

4.2.1 Multiple Logistic Regression Models

Building logistic regression models when there are many possible covariates can confuse. It is often useful to work hierarchically, looking at increasingly more complex structure of nested models, using test statistics like likelihood ratio or wald in deciding which covariates are important or not in predicting the response. A univariate logistic regression was used to select first candidate variables among many unidentified possible explanatory variables; in this part of the study it was fitted for 561 patients given antibiotics orally from pharmacy having only one observation so as to satisfy independent assumption. Hosmer and Lemeshow recommend including any covariate in multivariate analysis which had p-value less than 0.25 in univariate analysis. Patient *sex*(p=0.193), and *age*(p=0.087) were not associated with the susceptibility results but were kept in multiple logistic model based on Hosmer and Lemeshow (2000) cutoff, *treat* which is previous antibiotic used (*macrolides/tetracyclines* and *penicillin/cephalosporines*) , *timecat* (ref=[1-60])days and n_pcat (ref=[1-2])days were found to be associated with the risk of streptococci resistance, so were also included in building of the multiple logistic regression model. For further analysis a 5% level of significance was used to retain variables in the model. None of the two way meaningful interactions between any of the two covariates were associated to the susceptibility result. Keeping the other three covariates in the model the patient *sex* (p=0.296), and *age* (p= 0.106) were found not to be significant. Hence they were deleted from the model. Although the interaction effect of treat and timecat (*treat*timecat*) was not significant (p=0.685), we retained it in the model since the previous study showed that the effect of *treat* varied with *timecat* (Malhotra-Kumar et al., 2007). Different link functions existed for logistic regression but the most popular are logit (AIC value, 490.170) and probit (AIC value, 490.313). Based on the smaller AIC value a logit model was a candidate. One advantage of the logistic regression model over the probit model is that the logistic regression effects can also be interpreted using odds ratios. The logistic regression model with treat*timecat was:

$$\log it(\pi) = \log it(p(result = R)) = \beta_0 + \beta_1 * n\_pcat + \beta_2 * timecat + \beta_3 * treat + \beta_4 * treat * timecat$$

The top panel of table 6 displays parameter estimates together with the standard errors (s.e) of the model without the interaction effect (treat*timecat) indicated a highly significant *treat* effect implying that the type of the previous antibiotic used had a significant effect on susceptibility result. Dichotomized time (*timecat)* and n_prescription (n_pcat) effects were also significant. Once the model has been applied the model, one needs to assess how well it fits the data, or how close the model-predicted values are to the corresponding observed values. Test statistics that assess fit in this manner are known as goodness-of-fit statistics. Several test statistics exist to assess the fit of the model such as Deviance, Pearson chi-square, and Hosmer and Lemeshow's statistic. However some of the groups had less than 10 subjects and more than 25% of the predicted cell counts were less than 5. Hence chi-square approximations of Deviance, Pearson chi-square may not satisfy .Therefore the goodness-of-fit was assessed using the Hosmer and Lemeshow test (p= 0.916) indicating that the model fit the data well. Controlling for other covariates in the model, patients with streptococci resistant were significantly more likely to have been prescribed *macrolides/tetracyclines* in the previous time compared to *penicillin-/cephalosporines* (OR: 4.024, CL: 2.142-7.559). Controlling for prescription type (*treat*), those patients whose samples were taken after

60days from the date they had been prescribed are less likely to have resistant result as compared to those patients prescribed within 60 days (OR: 0.415, CL: 0.254-0.677.In the group of patients where the most recent number of prescription was >2 days, there was a statistical significant increased risk of having streptococci resistance as compared to those with ≤ 2 days (OR: 1.833, CL: 1.074-3.127). Including the treat and timecat interaction (*treat*timecat*) did not change the effect of the other covariates. The fit of this model was assessed using the Hosmer and Lemeshow test (p=0.947) indicating that this model also fit the data well. The results of this model are presented in bottom panel of table 6.The conclusion obtained from this model is the same as that of the model without interaction effect between treat and timecat.

Table 6. *Parameter estimates and standard errors for multiple logistic regression model for patients with streptococci (n=561)*

Parameter	Estimate	Standard error	P-alue
	WITHOUT TIMECAT*TREAT INTERACTION		
Intercept	-0.989	0.173	<0.001
n_pcat	0.303	0.136	0.026
TimeCat	-0.440	0.125	0.001
treat	0.696	0.161	<0.001
	WITH TIMECAT*TREAT INTERACTION		
Intercept	-0.960	0.188	<0.001
n_pcat	0.306	0.136	0.025
TimeCat	-0.483	0.164	0.003
treat	0.717	0.169	<0.001
treat*TimeCat	-0.066	0.164	0.685

4.2.2 Generalized Estimating Equations (GEE)

For the second part of the study, 34 patients having more than one observation which resulted to 79 total observations were considered. First, we ignore the clustering by treating the observations from the same patient as if were all independent. In this case an ordinary logistic regression model was fitted and the results are found in table 9 (Appendix). Although this model seems to fit the data quite well, we have overlooked certain aspects in the data. First, the assumption that observations within patients are independent will in general be too strong. While this typically leaves the consistency of point estimation intact, but the same is not true for measures of precision (Hens et al.2007). In case of a “positive” clustering effect (i.e., observations within a patient are more alike than between patients), then ignoring this aspect of the data overestimates precision and hence underestimates standard errors and lengths of confidence intervals. This might result in significant results which are not true. Secondly, the outcome of interest may be related to the cluster size. This is termed ‘informative cluster size’ (Hoffman et al. 2001). Observations selected from a large cluster might have different probability to be resistant compared to observations selected from a small cluster; the logistic regression model weighs each observation equally. As a result, large clusters have more impact on the analysis in comparison with small clusters. Thus, the risk of an observation to be resistant will be different, compared to a method that weighs each cluster equally.GEE was therefore chosen as an alternative method that account for the correlation in the data. For simplicity, we assume an independence working correlation matrix. This choice is justified since the GEE method is robust against misspecification of the working correlation structure, at the cost of efficiency of the parameter estimates. The same model as logistic regression was fitted. The model-based and empirically corrected standard error estimates are given in top panel of Table 10 (Appendix). The empirically corrected standard errors are quite a bit larger than the model-based ones. This implies that ignoring the correlation in these data could lead to invalid conclusions. Note that the working correlation structure does not need to hit the true correlation structure to obtain valid inferences. To increase the efficiency of the parameter estimates it is better to choose a working correlation matrix that is close to the true one (Fitzmaurice et al., 2004). As another typical choice, an exchangeable working correlation matrix was considered, hypothesizing that the correlation between any two observations within a patient is constant. The corresponding model parameter estimates, the model-based and empirically corrected variance estimator are displayed in the bottom panel of table 10 (Appendix). The model-based and empirical (robust) standard errors of the parameter estimates for exchangeable working correlations’ were found to be closer compared to the independent working assumption. Further interpretation was made on the exchangeable working correlation. The results are as shown in top panel of table 7 below. The previous antibiotic used before the sample date (t*reat*) was statistically significant on the risk of streptococci resistant. The estimated odds ratio was obtained by taking exponent of the regression parameter estimate (log odd ratio). Keeping the age and n_pcat fixed odds of having resistant susceptibility result for those patients who had prescribed macrolides-/tetracyclines is 7.021 times that of patients who had prescribed penicillin/cephalosporines. This means that for patients who had prescribed *macrolides/tetracyclines* before the sample date had more chance to have streptococci resistance as compared to *penicillin/cephalosporines*. The odds of resistant versus susceptible result for unit increase in *age* are equal to1.033. However, taking correlation into account the *n_pcat* effect was no longer significant (p=0.062) as compared to the logistic regression which ignores the correlation between the measurements within cluster (patient). Although, the effect of n_pcat was not significant, we decided to keep it the model since previous studies together with the first part of this study showed the significant of it.

4.2.3 Informative Cluster Size

The cluster size is informative when the cluster size is related with the outcome of interest. In this part of the study, cluster sizes varied from 2 to 6 observations per patient. When dealing with informative cluster size, interest can either extend to probability of resistant of a randomly sampled unit(sample or observation) from the set of all units for which no adjustment to the analysis has to be made, or ,to the probability of resistant a random sampled unit from a randomly selected patient(Hens et al.2007). For the letter situation, Williamson et al.(2003) proposed to weigh each unit from from cluster(patient) with the inverse of its cluster size to obtain equal weight for all clusters. The generalized estimating equations implicitly presume that the size of the cluster is unrelated to the parameters under study. When the risk for a streptococci resistance is positive related with number of repeated measurement per patient (cluster size), then the GEE method will estimate the risk of having resistant streptococci as relative high ,as compared with a method that weight each patient equally (Faes et al., 2006; Williamson et al., 2003). It is clear that when cluster size is informative, sometimes the cluster-based marginal model may be more relevant than the observation-based model. The informativeness of the cluster size was checked by including the cluster size as a covariate in the model and it was not significant and found to be ignorable (p=0.301). The results are as shown in middle panel of table 7 below. Controlling for cluster size, the treat effect was not significant. This is due to the fact that treat associate with cluster size (p= 0.001).

4.2.4 Cluster-Weighted Generalized Estimating Equations (CWGEE)

Although the cluster size was independent of susceptibility result in the previous analysis, we fitted cluster-weighted GEE to compare the parameter estimates. The bottom panel of table 7 shows the result for cluster weighted GEE with independent working correlation. It can be seen that all the parameter estimates are similar in their magnitude to that of unweighted GEE under exchangeable working correlation matrix. The cluster weighted GEE and inclusions of cluster size as covariate are two methods to account for non ignorable cluster size. Both approaches yielded results that seem to be consistent in interpretation. The parameter estimates of Weighting the GEE are smaller than that of unweighted GEE under independent working correlation matrix. The results of unweighted GEE under independent working correlation matrix are presented in top panel of table 10 in appendix. The same thing happened when including cluster size as a covariate in the model .The interpretation of parameter estimates in CWGEE is different from GEE with “cluster size” as a covariate. Whereas CWGEE gives an estimate of the probability of a randomly selected sample from a randomly selected patient while GEE with “cluster size” as a covariate gives an estimate of the probability of a sample have resistant result from a patient with a specific number of repeated measurements.

Table 7. *Parameter estimates and standard errors for unweighted GEE, GEE model corrected for cluster size as covariate and using Independent correlation matrix ,and CWGE for patients with streptococci (n=34)*

Parameter	Estimate	Standard Error	P-value
UNWEIGHTED GEE			
Intercept	-3.442	0.906	<0.001
age	0.032	0.014	0.020
n_pcat	1.754	0.941	0.062
treat	1.949	0.807	0.016
GEE WITH CLUSTER SIZE AS COVARIATE			
Intercept	-4.313	1.292	0.001
age	0.0318	0.014	0.019
n_pcat	1.712	0.899	0.057
treat1	1.657	0.865	0.055
clustersize	0.375	0.362	0.301
WCGEE			
Intercept	-3.133	0.799	<0.001
age	0.030	0.013	0.022
n_pcat	1.492	0.870	0.086
treat	1.387	0.865	0.109

4.2.5 Random Effect Model (GLMM)

So far, we have dealt with marginal models to identify the risk factors associated with streptococci resistance. In contrast to the marginal model, it is our interest to investigate which risk factors are related to the risk of streptococci resistance within the specific cluster (patient) and to explain the differences between clusters. The random effects or cluster-specific model, specifically, the random intercept model was considered to account for heterogeneity among patients as well as allow for patient-specific inference. In this study, a generalized linear mixed model was used to model the risk of streptococci resistance as a function of covariates and parameters specific to a patient. Model fitting began by adoption of the marginal model fitted in GEE which was extended by allowing random intercept. The analysis has been performed using the SAS procedures NLMIXED. Different estimates and standard errors were observed with different quadrature values, and convergence was achieved at quadrature point 50 (table11 Appendix). The results for our analyses are summarized in table 8.The result showed that no covariate was significantly related to the risk of streptococci resistance at cluster-specific level. The estimated variance of random intercepts is relatively large, 72.406.This implies that there is substantial variability in the propensity to experience streptococci resistance among the patients.

Table 8. *Parameter estimates and standard errors for the regression coefficients obtained from analysis based on adaptive Gaussian quadrature with 50 quadrature points from NLMIXED for patients with streptococci (n=34)*

Effect	Parameter	Estimate	Standard Error	P-value
intercept	beta0	-17.63	10.309	0.097
Age	beta1	0.175	0.115	0.137
n_pcat	beta2	8.719	6.756	0.206
Treat	beta3	9.637	6.559	0.151
Variance of intercept	d	72.406	86.090	

4.2.6 Empirical Bayes Prediction

The Empirical Bayes (EB) estimates were obtained for our analysis. The estimates reflect between-clusters (patients) variability and are as such useful in detecting special trends. EB estimates are important whenever interest in prediction of cluster-specific trends. The scatter plot of random intercepts plotted in figure 2 shows that there is variability among the patients, and indicates the presence of outlying observations. The disadvantage of empirical Bayes estimates is that, they are underestimated when the parameters in the models are replaced by their estimators.

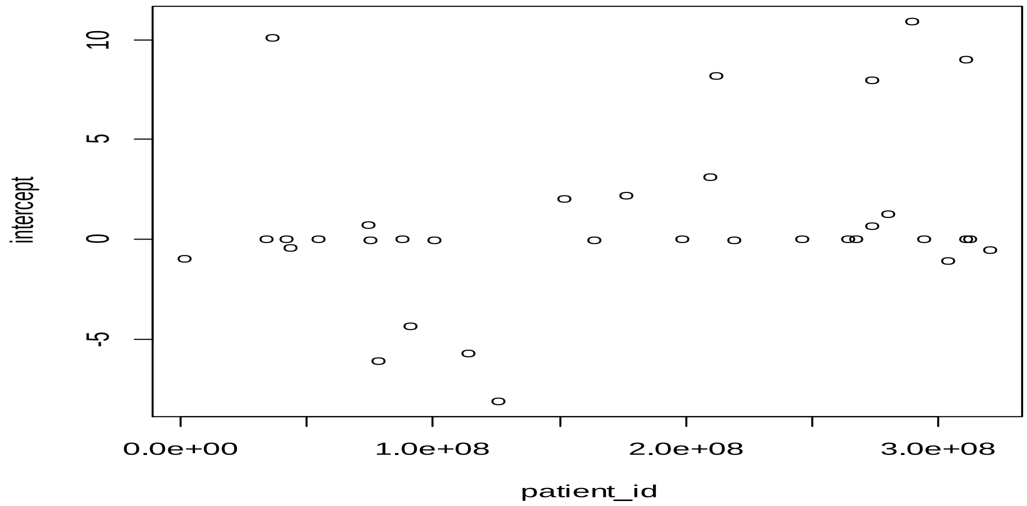

Figure 1. *Scatter plot of Empirical Bayes(EB) Estimates for patients with streptococci (n=34)*

5 Discussions and Conclusions

The objective of this study was to assess persistence of resistance selection of common antibiotic substances (macrolides/tetracyclines and penicillin/cephalosporins) among pathogenic bacteria respiratory samples (streptococci) based on a large dataset from a retrospective cohort study, and to compare with results obtained from earlier randomized controlled trials. To be able to answer the question of a researcher, understanding the nature of the data was an important step. In this study a substantial part of the data had a repeated nature in the sense that some of the patients (5.91%) had more than one observation imposing correlation among these measurements with binary response. Therefore, a model which takes into account the correlated nature had to be fitted. Because most of the patients (94.29%) had only one observation imposing independent observations, we decided to split the data into two parts. In this study logistic regression model was fitted for the first part of the data for 94.29% of the patients each having one observation. This model led to the conclusion that: Controlling for time category (*timecat*) covariate in the model patients with streptococci resistant were significantly more likely to have been prescribed macrolides/tetracyclines as compared to penicillin/ cephalosporins (OR: 4.024, CL: 2.142-7.559).This finding coincides with the results obtained from earlier randomized controlled trials where macrolides (clarithromycin and azithromycine) was found to be more resistance as compared to penicillin (*ampicillin*).

To account for correlation for second part of the data, first we considered the marginal model (GEE) proposed by Zeger and Liang (1986) to identify the risk factors for outcomes. The main finding was that the type of previous antibiotic used (*treat*) did have significant effect on susceptibility result, and in this case having prescribed macrolides/tetracyclines increased the probability of streptococci found in samples to be resistant compared topenicillin/cephalosporins. Since our data was clustered by nature, it was necessary to check for informative cluster sizes. Cluster size is informative when the cluster size is related with the outcome of interest. To check the informativeness of the cluster size, the Generalized Estimating Equation (GEE) was fitted with potential risk factors. The results showed that, the cluster size was uninformative. Despite that the cluster size was found to be uninformative, both unweighted and cluster weighted Generalized Estimating Equation (CWGEE) were fitted to compare the parameter estimates. The parameter estimates of weighting the GEE were smaller than of unweighted GEE under independent working correlation matrix. To identify the patient-specific risk factors associated with streptococci susceptibility result, a generalized linear mixed model (GLMM) was fitted. In this case no covariate was associated with response. The between patient variance was 72.406 indicating that there were large differences between patient.

Within the population setting, exposed to *macrolides/tetracyclines* increased the risk for resistant streptococci as *compared to penicillin/cephalosporins*. The probability of having a resistant result also influenced by another factors such as *age*, *timecat* and *n_pcat*. However, in cluster (patient) specific model, no covariate was found to be associated with susceptibility results of streptococci.

Reference

Aerts, M., Faes, C., Hens, N., and Molenberghs, G. (2010) Handling Missingness and Informative Cluster Sizes When Modeling Prevalence and Force of Infection. ENAR-JSM, 3491-3501.

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002) Topics in Modeling of Clustered Data. London: Chapman and Hall.

Agrest, A. (2002). Categorical Data Analysis (2nd ed.). New York: John Wiley and Sons.

Bennett, S., Woods, T., Liyanage, W.M., and Smith, D.L. (1991).A simplified general method for cluster-sampling surveys of health in developing countries. World Health Stat. Q., 44 (3), 98–106.

Catry, B., Hendrickx, E., Preal, R., and Mertens, R.(2008). Verband Tussen Antibiotica-consumptie en Microbiele Resistentie bij de Individuele Patient.

Faes, C. (2004) .Flexible Modelling of Correlated Multivariate Data with Applications in Animal Studies. Center for Statistics, Limburgs Universitair Centrum.

Faes, C., Hens, N., Aerts, M., Shkedy, Z., Geys, H., Mintiens, K., Laevens, H., and Boelaert, F. (2006) Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. Appl. Statist. 55(5), 595–613.

Fitzmaurice, G., Laird, N.,and Ware,J.(2004).Applied longitudinal analysis

Hillier, S., Roberts, Z., Dunstan, F., Butler, C., Howard,A., and Palmer,S.(2007). Prior antibiotics and risk of antibiotic-resistant community-acquired urinary tract infection: a case–control study.

Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. Biometrika, 88, 1121–1134.

Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression. New York: John Wiley and Sons, 2nd Edition.

Liang, K.Y., and Zeger,S.L(1986).Longitudinal data analysis using generalized linear Models.Biometrika,73,13-22.

Malhotra-Kumar, S., Lammens, C., Coenen, S., Van Herck, K.,and Goossens H.(2007). Impact of azithromycin and clarithromycin therapy on pharyngeal carriage of macrolide-resistant streptococci among healthy volunteers: a randomised, double-blind, placebo-controlled study.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models(2nd edition). London: Chapman and Hall.

Mellon, M., Benbrook, C. and Benbrook,K. (2000).Hogging It!: Estimates of Antimicrobial Abuse in Livestock. Cambridge, MA

Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data.

New York: Springer.

Williamson, J.M., Datta, S., and Satten, G.A. (2003) Marginal analyses of clustered data when cluster size is informative. Biometrics, 59, 36–42.

Todar,k.(2009). Bacterial Resistance to Antibiotics. University of Wisconsin-

Madison Department of Bacteriology

Appendix

Table 9: *Parameter estimates and standard error for logistic regression model ignoring Cluster effect for patients with streptococci (n=34)*

Parameter	Estimate	Standard error	P-value
Intercept	-1.691	0.599	0.005
age	0.035	0.011	0.001
treat	1.025	0.357	0.004
n_pcat	0.803	0.333	0.016

Table 10: *Parameter estimates (model-based standard error Empirical corrected standard error) for GEE under independent and exchangeable working assumption (n=34)*

Parameter	Estimate	Standard Error		P-value
		Empirical	Model -Based	
		INDIPENDENT		
Intercept	-3.518	0.898	0.777	<0.001
age	0.035	0.014	0.011	0.009
n_pcat	1.606	0.905	0.667	0.076
treat	2.049	0.911	0.714	0.024
EXCHANGEABLE				
Intercept	-3.442	0.906	0.973	<0.001
age	0.032	0.014	0.014	0.020
n_pcat	1.754	0.941	0.842	0.062
treat	1.949	0.807	0.731	0.016

Table 11: *Summary of parameter estimates and associate standard error from fitting random effect model, for varying numbers Q of quadrature points and for Adaptive Gaussian quadrature.The obtained maximized approximate log-likelihood is denoted by* ℓ

	Q=5	Q=25	Q=50	Q=100
β_0	-30.499(19.269)	-18.376 (8.966)	-17.632(10.309)	-17.326 (11.463)
β_1	0.312 (0.214)	0.181 (0.103)	0.175 (0.115)	0.1734(0.127)
β_2	13.364 (10.612)	9.340(5.871)	8.719 (6.756)	8.415 (6.765)
β_3	17.177(10.067)	10.089 (6.016)	9.637 (6.559)	9.377(6.819)
d	290.810 (395.95)	77.287(71.351)	72.406 (86.090)	68.7131(90.326)
-2ℓ	50.0	48.7	48.9	48.9

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Persistence of resistance selection by common antibiotic substances in streptococci: a case-control study as surrogate for a randomized placebo controlled trial

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Mtumwa, Abdalla

Datum: **12/09/2013**