2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Modelling the Persistence of Trimethoprim and Sulfonamide Resistance in Escherichia coli Bacteria

Promotor :
Prof. dr. Niel HENS

Promotor :
Mr. BOUDEWIJN CATRY

Melkie Chernet Leykun

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

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Maastricht University

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

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Modelling the Persistence of Trimethoprim and Sulfonamide Resistance in Escherichia coli Bacteria

Promotor :
Prof. dr. Niel HENS

Promotor :
Mr. BOUDEWIJN CATRY

Melkie Chernet Leykun

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

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Contents

List of Tables

List of Figures

Abstract

Urinary tract infections (UTIs) are common infections often caused by *Escherichia coli* and trimethoprim/sulfonamide is a first line antimicrobial therapy in many countries. Research demonstrates that antibiotic resistance leads to therapy failure and higher medical costs because alternative agents have to be administered.

The goal of this study was to explore determinants for trimethoprim/sulfonamides resistance of *E. coli* strains and the time frame at which resistance persists after therapy with this very antimicrobial agent. This information can help clinicians to avoid making unnecessary changes in therapy.

For this thesis, susceptibility results of *E. coli* isolates originating from older adults (65+ years) who provided a urine sample in 2005 to 17 voluntarily participating laboratories were considered. Among the 228 019 strains isolated in 9 967 urine samples from 7 621 patients , 30.5% were resistant and 69.5% were susceptible to trimethoprim or/and trimethoprim/sulfonamide. A model was fitted which takes into account the correlated nature of the response, Generalized Estimating Equations (GEE) with Exchangeable working correlation. Patient sex, time gap between the prescription date and urine sample date (categorized version) and the number of trimethoprim or/and trimethoprim/sulfonamide administrations during the previous 6 months were found significantly associated to the susceptibility result. Strains retrieved from men had a higher chance to be resistant as compared to women. The closer the prescription date to the sample date the higher chance bacteria were resistant.

Keywords: Urine sample, *E.coli*, trimethoprim and sulfonamide, antimicrobial resistance, GEE, CWGEE.

Acknowledgement

Firstly, I would like thank almighty God! Next my special thanks goes to my internal supervisor Professor Niel Hens for his continuous support for my project, from initial contacts and creating a conducive atmosphere, and through ongoing advice and encouragement to end of the project. My next special thank goes to the Vlaamse Interuniversitare Raad (VLIR) university cooperation development program for sponsoring me financially, so that I can focus not only on my thesis research but also on my whole study without any extra difficulties in life. Also my special thank goes to the external supervisors Dr. Boudewijn Catry and Katien Latour for their generous support for all the time and making me to feel free in any aspect as well as providing reference journals. I am also grateful to thank Adriaan and Sylvanus for sharing their generous ideas and precious time during our meeting(s) in the Scientific Institute of Public Health (WIV-SIP), Brussels Belgium. In addition, I would like to thank staffs imparting their knowledge and colleague students for supporting and encouraging in one or another ways in the department of Interuniversity Institute for Biostatistics and statistical Bioinformatics at the university of Hasselt.

Melkie

September 2012

Genk, Belgium

1 Introduction

1.1 Background

Physicians prescribe antibiotics to treat bacterial infections depending/basing on their past experiences, therapeutic guidelines and symptoms of the infection (severity of illness). However, with the increase in bacterial resistance to antibiotics taken by patients, it has become more difficult for doctors to prescribe an effective, first line therapy, drug (antibiotics). Antimicrobial resistance is a major public health problem leading to increased morbidity and mortality, longer stay in hospital, require multiple antimicrobial therapies resulting in increased healthcare costs (Gyssens, 2011).

A urinary tract infection (UTI) may be primary or recurrent, complicated or uncomplicated and mostly bacteria are involved in the infection of the urinary system. Antibiotics are used to cure bacterial infections, as with UTIs. This type of infection is among the most commonly diagnosed and treated infectious disease in hospital, clinics, chronic care facilities and ambulatory practice (Florian *et al.*, 2011), and exhibits a high recurrence rate (greater than 25%) within 6 months (Mysorekar and Hultgren, 2006). Symptoms of urinary tract infection may vary according to age, sex, and location of the infection throughout the tract, though most common symptoms include frequent urge to urinate, painful or burning urinating, and blood in the urine (Bishop, 2004).

The risk of UTI depends on a variety of factors, including age, gender, lifestyle, anatomy, and disease process. Women are highly affected by urinary tract infections, and possible reasons include using birth controls like diaphragms, undergoing menopause, and the short length of the female urethra. In general, diseases or underlying conditions that lead to urinary obstruction include genetic abnormalities, prostatitis, kidney stones, and inability to maintain good hygiene.

Escherichia coli (*E. coli*) is one of the primary bacteria involved in for UTI in the western worlds (White et al 2005). *E. coli* is often characterized by co-resistance, that is resistance to one antimicrobial simultaneously associated with resistance to another antimicro-

bial. Trimethoprim and trimethoprim with sulfonamide (cotrimoxazole) are common antibiotics to treat urinary tract infection caused by *E. coli*. These antibiotics can be administered through the oral or parenteral route.

An appropriate management of UTI is prevention and avoiding antimicrobial therapy for asymptomatic bacteria (i.e. symptoms treated as a suspected infection) in the urine. In addition to the latter principle, knowledge of local antimicrobial resistance profiles can guide effective treatment while limiting unnecessary hospitalization (expensive), broad-spectrum antimicrobial use, and diagnostic testing.

Though antibiotics are used to treat/cure bacterial infections, bacteria have the potential to develop resistance at any time. This means that antibiotics once used to kill or inhibit their growth, may no longer be effective. This resistance is continuously increasing in UTI (Naber *et al.*, 2011). Within a given bacterial species, antimicrobial resistance can be intrinsic (natural or inherent) or acquired. Acquired antimicrobial resistance is an increasingly serious problem which limits options for effective treatment of bacterial infections with antimicrobials (Vellinga *et al.*, 2012).

Antimicrobial therapy is mostly guided by laboratory susceptibility testing. The results from this test will help doctor's to determine which antibiotic will be most effective in treating bacterial infection (Jorgensen and Ferraro, 2009). A number of antimicrobial susceptibility testing (AST) methods are available to determine bacterial susceptibility to antimicrobials. The selection of a method is based on many factors such as practicality, flexibility, automation, cost, reproducibility, accuracy, and individual preference. The ones consistently providing reproducible and repeatable results when appropriate guidelines are followed, (Clinical and Laboratory Standards Institute (CLSI, 2006) are: disk diffusion, broth dilution, and agar dilution. The turn around time of these test often takes 3 to 5 days after sampling. In the near future, alternatives based upon DNA detection (PCR, micro-array) or spectrometric identification systems likely will become standard in the future given the rapid availability of test results.

1.2 Objectives

The main goal of this study was to highlight the effect of antimicrobial use on emergent resistance for individuals and, ultimately, to reduce antimicrobial prescribing and to explore the (quantitative) relationship between the prescription of antimicrobials and subsequent antimicrobial resistance. For this thesis, susceptibility results of *E. coli* bacteria isolated in urine samples from older adults (65+ years) were selected.

Specific aims:

(1) to explore characteristics of the included urine samples and patients for*E. coli*
(2) to explore, after adjustment, variables to be associated with resistance to the antibiotics trimethoprim and trimethoprim combined with sulfonamide
(3) to assess time (e.g. in days) between last exposure to an antibiotic and sample result

2 Data

To have a complete knowledge on solving problems one needs a firm understanding of variables meaning and their use. Some of the important variables, potential explanatory variables, have been described in table 1.

Table 1: Variables and their descriptions.

Variable	Label	Description
Patient_Sex	1. Male (Men) 2. Female (Women)	Sex of a patient who provided a urine sample
Age	Age	Age (in years) of the patient from birth to 2005
Agecat	1. 65 to 84 2. 85 to 103	Categorized version of age in years
Prior_Trimeuse	Prior_Trimeuse	Number of times a patient is prescribed trimethoprim or trimethoprim/sulfonamide six months prior to urine sample
Sample_date	Sample_date	Date of urine sample taken
Prescription_date	Prescription_date	Date of antibiotics prescribed
Timecat	1. $\leq$ 30 days 2. (30 -60] days 3. (60 -90] days 4. (90-180] days 5. $>$ 180 days	Categorized version of time difference in days between sample date and prescription date
Prior_CDDD	Consumed DDD	Amount of antibiotics consumed in DDD unit six months prior to the urine sample
Result	1. R=Resistantt 2. S=Susceptible	The response variable which is susceptibility test result (intermediate was set as resistant)

Next to defining the variables in the data, exploring their distribution is also important. For this thesis, data collected in light of a multicenter study on the association between antimicrobial consumption and resistance in the individual patient was used. There were two databases: antimicrobial prescription and laboratory data on positive urine samples. The former contained 12 249 patients and the latter contained 13 083 patients. Determination of bacterial resistance to antimicrobials is an important part of the management of

infections in patients (Kirby *et al.*, 1959). Susceptibility of an antibiotics was performed by 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 17 voluntary participating Belgian clinical microbiological laboratories between periods 1st January and 31st December 2005 were merged with the individual antimicrobial consumption patterns (which were pooled from seven Belgian health insurance funds intermutualistic agency, IMA), extended with a half year prior to the first laboratory observations (July 2004–December 2005). In this study, trimethoprim and combination of trimethoprim with sulfonamide use and the susceptibility of *E. coli* bacteria found in urine samples from elderly (65 years or older) for these antibiotics were considered as they were our primary interest.

The original database, obtained from the Scientific Institute of Public Health (WIV-ISP, Brussels), contained 19 038 urine samples from 13 083 elderly patients (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 481 091 records in the database. Of the 19 038 urine samples *E. coli* was found in 9 754 urine samples (51.2%) from 7 625 patients and tested for different antimicrobials resulting in 228 019 records.

Susceptibility tests results in urine samples with *E. coli* were 59.3% susceptible, 3.1% intermediate, 8.3% resistant, and 29.3% were not reported. In this study, only a part of the general database was used. From the antimicrobial consumption patterns only trimethoprim and sulfonamide with their analogues were considered (ATC J01E class) while from the microbiological results only data originating from patients aged 65 and above, urine samples, *E. coli* bacteria, and susceptibility of this bacteria tested for trimethoprim, trimethoprim and sulfonamide including their derivatives were considered. Also susceptibility results not reported were dropped and an intermediate susceptibility result was considered as resistant, a bacteria which has started producing resistant genes will quickly develop full resistant

genes to an antibiotic. Of 9 967 urine samples positive for *E. coli*, 1 278 were tested for trimethoprim and 8 689 for trimethoprim/sulfonamide. Moreover, 1 911 were men and 7 843 women patients.

Table 2: Exploring the original prescription and microbiological databases with different characteristics.

Item	Frequency	Percentage (%)
Patients older than 65 years with at least one urine sample (n=13 083)		
Men	3 978	30.4
Women	9 105	69.6
Patients older than 65 years with at least one urine sample (n=13 083)		
At least one antibiotic prescription during study	12 249	93.6
No antibiotics prescription	834	6.4
Number of urine samples (n=19 038)		
Men	5 799	30.5
Women	13 239	69.5
Bacteria found in urine samples (n=21 063)		
E. coli	9 754	51.2
K. pneumonia	966	5.1
others	10 343	43.7
Number of urine samples with *E. coli* (n=9 754)		
Men	1 911	19.6
Women	7 843	80.4
Number of susceptibility tests in urine samples with *E. coli* (n=228 019)		
Susceptible	135 220	59.3
Intermediate	7 078	3.1
Resistant	18 822	8.3
Not reported	66 899	29.3
Susceptibility of *E. coli* for J01E antibiotics (n=9 967)		
Trimethoprim	1 278	12.8
Trimethoprim + sulfonamide	8 689	87.2
Susceptibility of *E. coli* for J01E antibiotics (n=9 967)		
Susceptible	6 932	69.6
Intermediate	18	0.2
Resistant	3 017	30.3

3 Methodology

To address the objectives of this study, a procedure was followed depending on the characteristics/nature of the data. As a patient could have been prescribed different antibiotics in same day, or could have one or more urine samples, data were clustered or repeated in a patient. Hence data was first explored to get insight and simple (univariate) logistic regression was fitted to find possible risk factors associated with resistance of *E. coli* ignoring the repeated nature of the data. Then generalized estimating equations were fitted which accounts the clustered or repeated nature of the data and it is population averaged, marginal model. The generalized estimating equation gives equal weight for the unbalanced number of repeated measurements per patient and it was extended and weighted generalized estimating equations was fitted.

3.1 Exploratory Data Analysis

In order to get insight into the data structure and study distribution of the response variable with different demographic characteristics of a patient, possibly potential explanatory variable, cross tabulation were done.

3.2 Statistical Analysis

The analysis of research data is often confronted with the issue of building up models. It is obvious that a variety of models can be fitted for a given dataset. Thus, the model selection process based on sound and scientific principles is imperative (Burnham and Anderson, 2002).

3.2.1 Logistic Regression Analysis

Deciding which covariates to be kept in the statistical model has always been a difficult task for data analysts. Regression analysis can be used to assess the effect and relationship between explanatory variables and response variable. Many types of regression analysis exist

and logistic regression is one of them. In logistic regression, we are interested in studying how risk factors were associated with presence or absence of an event. Here we used simple logistic regression to select candidate risk factors (explanatory variables) for further analysis with a p-value cutoff point 0.25 (Hosmer and Lemeshow, 2000). That is explanatory variables in a simple logistic regression with p-values less than 0.25 would be included in a model. 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 fitting.

Here also the multiple logistic regression was used to see the usual purpose, i.e estimation of risk factors and prediction. However, it ignores the repeated nature of the data and the estimates of the parameters might be unbiased but the standard errors would be underestimated and leads to a significant results which might not be true.

In logistic regression the mean and response are related through different link functions. Thus, the model can be written as (Agresti, 2002),

$$g(\pi(x_i)) = \sum_{i=0}^{k} \beta_k x_k, \tag{1}$$

where $g(.)$ is the link function, $x_0 = 1$, Y_i is the binary response $x_i = (x_1, x_2, ..., x_k)$, is the explanatory variables, $\pi(x_i) = P(Y_i = 1|X = x_i)$ is the probability of success (it can be good or bad, in this case a resistant result (R) was "success") and β_k are maximum likelihood estimates which are obtained as solutions to the full likelihood equations.

Several samples were considered for the same patient and therefore indicated of correlated data, the variance of Y_i in the binary case can then be inflated (Agresti, 2002).

Since, multiple logistic regression assumes multiple observations coming from the same patient are independent, it ignores the intraclass correlation (correlation within repeated measurements). The appropriate model was fitted using Generalized Estimating Equations (GEEs).

3.2.2 Generalized Estimating Equations

The generalized estimating equations (GEE), introduced by Liang and Zeger (1986), is a method of analyzing correlated data in which measurements are taken on subjects who share a common characteristic such as belonging to the same patient. A number of working correlations exist from simple (independent, no correlation) to complex (unstructured), even if the correlation structure is misspecified, parameter estimates remain consistent (Molenberghs and Verbeke, 2005). Though an appropriate working covariance should be used; 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). Fitting GEE requires the correct definition of clusters of observations. Observations within a cluster are assumed to be correlated while observations from different clusters are assumed to be independent. One urine sample could be combined with one or more prescription and this might arose for different urine samples for a patient. Given the complexity of the data characteristics here, it was difficult to define the cluster (repeated measurements). It was investigated for two clusters and different working correlations. When the definition of clusters is ambiguous, there results instability on the parameter estimates for different working correlation assumptions (Bishop, Die, and Wang, 2000). Another issue that leads to instability of the parameter estimates in a model was multicollinearity.

GEE is used to characterize the marginal expectation of a set of outcomes as a function of a set of explanatory variables (Agresti, 2000):

$$g(\mu_{ij}) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{x}_{ij}$ is a p times 1 vector of covariates, $\boldsymbol{\beta}$ consists of the p regression parameters of interest, $g(.)$ is the link function, and denotes the j^{th} outcome (for j=1,...,J) for the i^{th} subject (for $i = 1, \ldots, N$)

Liang and Zeger (1986) proposed generalized estimating equations of the form:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_{ij}}{\partial \boldsymbol{\beta}_k^{'}} \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}, \tag{3}$$

where $\boldsymbol{\mu}_i$ is the vector of probabilities associated with $\mathbf{Y}_i$ and the parameters are evaluated at their current estimates Agresti (2002)

$$\boldsymbol{V_i} = \phi \boldsymbol{A_i}^{1/2} R(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$$

Where ϕ is a dispersion parameter, $\mathbf{A}$ is a diagonal matrix of variance functions, and $R(\boldsymbol{\alpha})$ is the working correlation matrix of $\mathbf{Y}$.

3.2.3 Cluster-Weighted Generalized Estimating Equations

When the number of repeated measures called "cluster size" is informative, generalized estimating equations (GEEs) may lead to invalid inferences. An approach proposed by Williamson *et al.* (2003) to fitting marginal models to clustered data when cluster size is informative uses a generalized estimating equation (GEE) that is weighted inversely with the cluster size. Wang *et al.* (2010) also showed estimates of GEE are biased, while CWGEE gives unbiased estimates. However GEE estimates also give unbiased estimates when cluster size is used as a covariate, though in this case the parameter estimates will be different and will have a different interpretation (Faes, 2004).

3.2.4 Survival Analysis

Survival analysis is the model used to describe and fit the analysis of data in the form of times from a well-defined time origin to the occurrence of some particular event or end-point (time to event). However all the events might not hold at the same time giving rise to censoring.

In the case of recurrent events, correlation can come from event dependence i.e., the occurrence of one event may make further events more or less likely which results within subject correlation. Therefore, during statistical modeling the within subject correlation should be considered (Lu and Liu, 2008).

An appropriate univariate analysis should be done in any statistical analysis before complex models. Similarly, in survival analysis it is important to do the Kaplan-Meier curves for categorical variables to get insight into the possible assumptions (like proportionality of the groups, i.e. curves of the groups are parallel). The primary selection of any potential predictor variables is their scientific meaning to the field and next was retaining and excluding from the final model. That is, a possible explanatory variable from a set of variables could be selected depending on the clinical or biological importance. A cox proportional model for continues and log-rank test for categorical variables with a cut-off point of p-value 0.25 was used. That is predictors with p-value less than 0.25 were considered in the final model. The Kaplan-Meier estimator, Cox Proportional Hazards model uncorrected and corrected for the robust variance were fitted.

3.3 Model Selection Criteria

The task of choosing a parsimonious model that fits the data set better out of all possible models which use full likelihood was accomplished by obtaining the AIC values for each model and is defined as (Akaike, 1973): AIC=-2 (log-likelihood) - 2k. The log likelihood being for a particular model and k is the number of its parameters. The model with the smallest AIC value was considered as a better model. Since GEEs is not a fully likelihood approach, the QIC can be used to compare models given one is nested in the other. 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 Used

The SAS statistical package version 9.2 was used for data management and statistical analysis also excel 2010 and R version 2.15.0 were used only for data manipulation. All the tests were done at the 5% level of significance unless otherwise stated.

4 Results

4.1 Exploratory Data Analysis

The refined data base combined the microbiological result (susceptibility tested only for *E. coli*) data base and antimicrobial consumption pattern data base (only trimethoprim or trimethoprim/sulfonamide) and considered all urine samples and the respective prescription date was assumed to be July 1, 2004 for patients who were not prescribed from July 1, 2004 to Dec 31, 2005 resulted in 10 547 observations/records from 7 621 patients. Table 3 shows the distributions of different characteristics of these observations/records; a patient, antibiotics and if patients died during the follow up. From the susceptibility tests of the results, 30.5% were resistant (including intermediate results) and 69.5% were susceptible. The combination trimethoprim and sulfonamides was mostly tested (87.2%) and only trimethoprim in the minority (12.8%).

Table 3: Distributions (percentage) of susceptibility results for *E. coli* , for trimethoprim and trimethoprim/sulfonamide, and patient status.

Item	Frequency	Percentage (%)
Susceptibility result		
Resistant	3 035	30.5
Susceptible	6 932	69.5
Antibiotics (susceptibility of *E.coli* tested on refined data)		
Trimethoprim	1 278	12.8
Sulfonamides + Trimethoprim	8 689	87.2
Patient died in 2005		
Alive at the end of study (2005)	6 351	83.3
Died in 2005	1 270	16.7
Time category prior to trimethoprim use		
$\leq$ 30 days	301	2.9
(30-60] days	154	1.5
(60-90] days	118	1.1
(90-180] days	318	3.0
$>$ 180 days	9 656	91.5

Also 83.3 % patients were alive at the end of the study period while 16.7% died in 2005.

From table 4, the mean age of the participant patients was 79 years (it is closer to the minimum than maximum could indicate most of the participant patients were from 65 to 84 years, table 5) and the variability of age was found to be 7.5 years. The mean number of prescription (trimethoprim or/and trimethoprim/sulfonamide) a patient was prescribed 0.2 times with standard deviation of 0.8. Since most of patients were never prescribed within 6 months prior to the urine sample the mean is not good (could be affected by extreme values), also the standard deviation is larger than the mean. The mode is more stable (not affected by extreme values and it is the most frequent number of prescription for a patient) than mean and it was 0.

Trimethoprim use refers to the prescription of trimethoprim or trimethoprim/sulfonamide within the six months prior to the sample date. In the data base we could only look back until the 1st of July 2004. In 8.5% of the observations/records, patients used trimethoprim or trimethoprim/sulfonamide prior to the sample date and 91.5% they did not used prior to 6 months of urine sample. However, this was missing because we did not know for sure if (s)he took trimethoprim or trimethoprim/sulfonamide in the last 6 months (at most only known from July 1, 2004 to December 31, 2005) and the respective antibiotics and prescription date were assumed to be trimethoprim or trimethoprim/sulfonamide and July 1, 2004 respectively.

Table 4: Summary statistics for age, number of prior prescriptions of trimethoprim or/and trimethoprim/sulfonamide patients with *E. coli*.

Variable	Mean	Std Dev	Minimum	Maximum
Age	79	7.5	65	103
Prior_Trimeuse	0.2	0.8	0	9

Of all patients, 75.1% were between 65 to 84 years old and age above 84 years accounted for 24.9% (table 5). Most of the participating patients were women who covered 57.9% from 65 to 84 years and 21.8% above 84 years. Overall women covered 79.7% while men covered 20.3% (table 5).

Table 5: Distributions (percentage) of patients with *E. coli* by age group and sex .

Patient sex	Age 65 to 84	Age>84	Total
Men	994 (17.2)	178 (3.1)	1 172 (20.3)
Women	3 346 (57.9)	1 260 (21.8)	4 606 (79.7)
Total	4 340 (75.1)	1 438 (24.9)	5 778 (100.00)

Figure 1 displays the number of susceptible and resistant results with respect to the number of records a patient has in the refined database (results which can be considered correlated, results from same patient for different samples and resulting count called repeated samples," cluster size"). The plot shows almost equivalent amount of proportions of susceptible results (resistant: light blue painted and susceptible light purple painted) for trimethoprim or trimethoprim/sulfonamides, especially in the lower number categories. This might be an indicator as cluster size has no effect on susceptible results and considered as uninformative. Because this is an insight it needs to be checked formally (using statistical models).

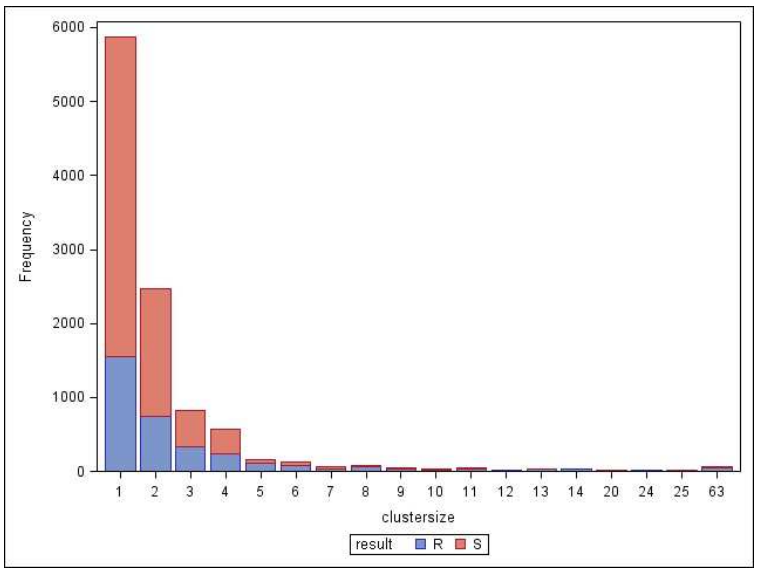

Figure 1: Frequency distribution of susceptibility results by number of repeated measurement for patients with *E. coli*.

There were 7 621 patients and 5 868 (55.7%) with only one observation. The maximum number of observations per patient was 63 from 63 patients; 0.6%. Overall, 526 (4.9%) patients contributed seven or more observations.

4.2 Statistical Models

4.2.1 Logistic Regression Analysis

Six months prior to the urine sample, the number of trimethoprim or trimethoprim/sulfonamide prescriptions (Prior_Trimeuse) and amount of daily defined dose (DDD) consumed were suspected to be highly correlated, and it was found to be 0.84 (they go in same direction as rate of trimethoprim prescription increased the amount of consumed DDD (C_DDD) also increased) and significantly different from zero (p-value <0.0001). The highly correlated nature of these two variables might resulted in multicollinearity (both independent variables explaining the response in the same way) leading to the instability of the parameter estimates. To keep either of them, from clinical view it is easier for the patient and the clinician to remember the number of prescriptions than amount of dose, this was supported by fitting simple logistic regression and the smaller AIC was kept. Therefore, Prior_Trimeuse (model AIC=193.8) was retained and C_DDD (model AIC= 246.4) was dropped from consecutive modelings.

Though a univariate logistic regression was used to select when there were many unidentified possible explanatory variables, in this study it was fitted assuming independence of observations to select first candidate variables. Antibiotic use either in hospital or pharmacy was dropped since it was not associated (p-value=0.6567) with the susceptibility results. Also age was not associated (p-value=0.2567) based on Hosmer and Lemeshow (2000) cutoff. This might be because all patients were elderly and with no substantial difference in immunity or other host related factors. Nevertheless age was retained in the consecutive modeling for biological reasons not to reject early at 25% level of significance. None of the two way meaningful interactions, age with sex and sex and prior trimethoprim use (Prior_Trimeuse), were associated to the susceptibility result. For further analysis at a 5% level of significance was used to retain further variables in the model. A different link function existed for logistic regression, *logit* (AIC value, 195.3), *complementary log-log* (AIC value, 193.5), and *probit* (AIC value, 193.8). Based on the smaller AIC value *complementary log log* link was a candidate, however it had no much improvement in AIC when compared to *logit* link. Hence the

logit link was used with its nice interpretability. The possibly influential factors for *E. coli* to be resistant for trimethoprim or/and trimethoprim/sulfonamide and logistic regression model is:

$$logit(Result = R) = \beta_0 + \beta_1 * prior\_trimeuse + \beta_2 * sex + \beta_3 * timecat1 + \beta_4 * timecat2 + \beta_5 * timecat3 + \beta_6 * timecat4$$

Table 6 displays the parameter estimates and standard errors with their respective p-values for multiple logistic regression, which does not take into account the correlated nature of the response. Patient sex was highly significant, also prior_trimeuse and timecat (time category) had a significant effect on susceptibility result.

Table 6: Parameter estimates and standard errors for simple logistic regression for patients with *E. coli*.

Effect	Parameter	Estimate	Standard error	p-Value
Intercept	β_0	-0.7048	0.0470	<.0001
Prior_Trimeuse	β_1	0.4763	0.0750	<.0001
Patient_Sex	β_2	-0.2169	0.0523	<.0001
$\leq$ 30 days	β_3	1.3792	0.1752	<.0001
(30–60] days	β_4	1.1822	0.2394	<.0001
(60–90] days	β_5	0.5823	0.2316	0.0119
(90–180] days	β_6	0.3733	0.1613	0.0206

4.2.2 Generalized Estimating Equations

Independent and exchangeable working correlation assumptions which were possible and meaningful candidates were used. The reference category for patient sex was male and for categorized time prior to urine sample date and prescription date was more than 180 days, other variables in the model were quantitative.

Results for independent working correlation were close to exchangeable working correlation assumption. Table 7 displays the result for GEE with exchangeable working correlation assumption.

Table 7: parameter estimates and standard errors exchangeable working correlation assumptions for full data for patients with *E. coli*.

Effect	Parameter	Estimate	Standard error	p-Value	Odds Ratio
Intercept	β_0	-0.7738	0.0547	<0.0001	-
Prior_Trimeuse	β_1	0.2805	0.1404	0.0456	1.32
Patient_Sex	β_2	-0.2330	0.0610	0.0001	0.79
$\leq$ 30 days	β_3	1.2762	0.1993	<0.0001	3.58
(30–60] days	β_4	0.8688	0.2893	0.0027	2.38
(60–90] days	β_5	0.5244	0.2406	0.0293	1.69
(90–180] days	β_6	0.4464	0.2160	0.0388	1.56

As different working correlation assumptions existed to fit the marginal model (population level), there were different methods to select model for further interpretations. One thing to be considered is based on model parsimony, i.e. in this case both the working correlations, independent and exchangeable, are simple further consideration was made to consider either working correlations. The model based and empirical (robust) standard errors of the parameter estimates were compared and the exchangeable working correlations' were found to be closer compared to the independent working assumption. Hence further interpretation was made on the exchangeable working correlation (correlation between any pair of responses, susceptibility result, for a patient are the same).

The time gap between prescription date and urine sample date in days (categorized), the number of trimethoprim and/or trimethoprim/sulfonamide analogues used prior to six months (Prior_Trimeuse) and patient sex were significant at 5% level of significance on the risk of trimethoprim and/or trimethoprim/sulfonamide and families (J01E class) to become resistant for *E. coli*.

The estimated odds ratio was obtained by taking exponent of the regression parameter estimate. The estimated OR for patient sex was 0.79; the estimated odds of females being resistant decreased by 21% when compared to males. i.e. the odds of being resistant for a susceptibility result for females is 0.79 times males keeping other variables fixed.

For each unit increase in prior use of trimethoprim and/or trimethoprim/sulfonamide, the

estimated odds of being resistant for susceptibility test increase by 32% (OR=1.32). Also the chance of becoming resistant increased as the time gap in days between urine sample and prescription dates got shorter. The chance of being resistant when the difference was within 30 days was 3.58 times higher in comparison with when difference was more than 180 days (6 months) with a 95% confidence interval (2.42, 5.30) keeping other variables fixed.

The estimated odds ratio was obtained by taking exponent of the regression estimate, the CI endpoints for the ORs were obtained by taking exponent of the CI endpoints for the corresponding regression parameter and the confidence interval for the regression parameters was found by (estimate$\pm$1.96*standard error).

The informativeness of the number of repeated measurement per subject, "cluster size", was checked using it as an explanatory variable and it was not significant and found to be ignorable (appendix, table 12), as expected from the exploratory data analysis (see figure 1). The cluster size, ranging from 1 to 63, was also categorized ($size \leq 1$, $1 < size \leq 4$ and $size > 4$) was also not significant.

Even though, "Cluster size" was not informative, the aim was to identify possible influential factors and to study this association for a randomly selected patient from whole population. Incase of non ignorable number of repeated measurements per subject, WCGEE and "cluster size" as a covariate could be used (Faes, 2004). Cluster Weighted GEE (CWGEE) was fitted how the effect of the covariates change as compared to the GEE. Table 8 shows the result for cluster weighted GEE with independent working correlation. The estimates of the parameters are stable and also in the same direction. However the strength increase in the CWGEE as compared to the GEE except time difference of (60–90] and ≥ 180 days. The interpretation of parameter estimates in CWGEE is different from GEE with "cluster size" as a covariate. 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 resistant urine sample result from a patient with a specific number of repeated measurements.

Table 8: parameter estimates and standard errors for CWGEE for patients with *E. coli*.

Effect	Parameter	Estimate	Standard error	p-Value
Intercept	β_0	-0.8140	0.0555	<.0001
Prior_Trimeuse	β_1	0.4580	0.1234	0.0002
Patient_Sex	β_2	-0.2225	0.0619	0.0003
$\leq$ 30 days	β_3	1.4700	0.2493	<.0001
(30–60] days	β_4	1.1475	0.0.3237	0.0004
(60–90] days	β_5	0.2397	0.3172	0.4499
(90–180] days	β_6	0.4260	0.2221	0.0551

4.3 Survival Analysis

One or more urine samples were taken from a patient and then look back (as a retrospective study) (i.e. urine sample date was later than prescription date) whether the patient had any prescription of trimethoprim or/and trimethoprim/sulfonamide and calculate the time difference which was closest to the urine sample date. If the patient had prescription then it was an event, if not censored. Figure 2 showed possible profiles for a patient, the open circle indicates censored observations and lines without the open circles are events.

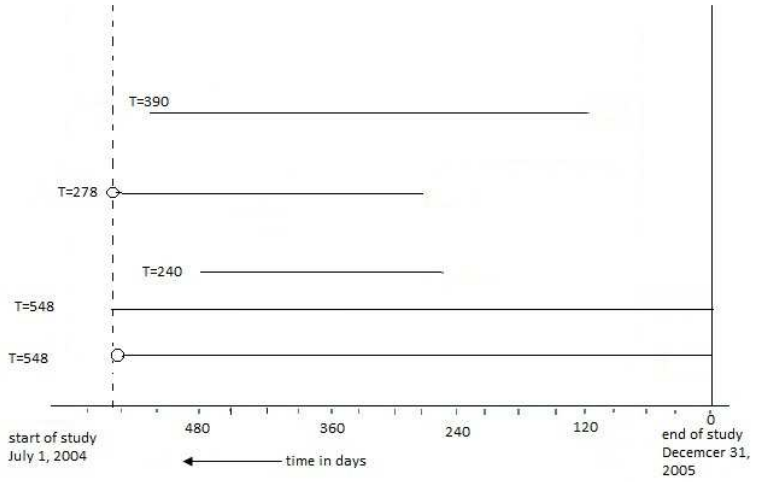

Figure 2: Profile plots for censoring and event times for patients with *E.coli*.

4.3.1 Kaplan-Meier Analysis

Analysis was started from a simple Kaplan-Meier (product limit) estimation. Log rank test was used to check whether there was difference on survival time between sex strata (men and women), similarly for susceptibility result strata: (resistant and susceptible). The survival estimates of men was different from women (χ^2=13.7, p-value=0.0002), there was also a difference for susceptibility result (susceptible vs resistant) (χ^2=263.8, p-value < 0.0001). However inference was not possible to made from Kaplan-Meier estimates since it assumed every observation was a single patient ignoring the repeated measurement. Though it was important for insight into the data.

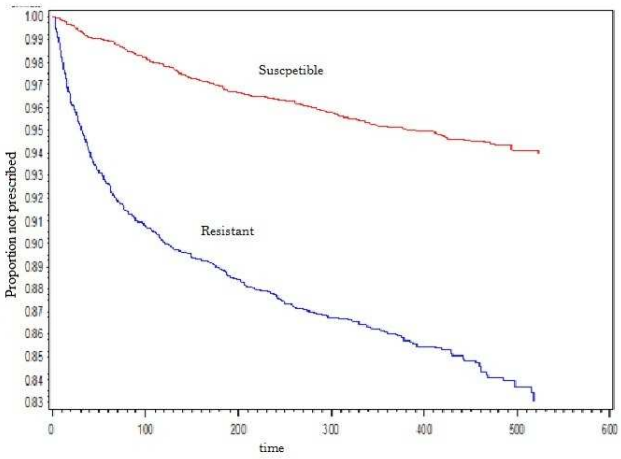

Figure 3: Kaplan-Meier curve for susceptibility result for patients with *E. coli*.

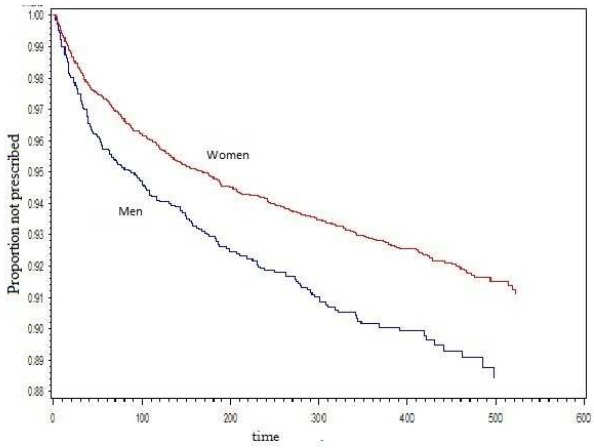

Figure 4: Kaplan-Meier curve for sex of a patients with *E. coli*.

To get the percentile distributions was not possible since the censored observations had a large time. Figures 3 and 4 showed the Kaplan-Meier curves for susceptibility results and patient sex respectively. They indicated that the susceptible results had a smaller proportion as there was prescription prior to urine sample compared to resistant. Similarly men's had a higher proportion of prescription than women's prior to urine sample at all time points. However the curves seem to stay parallel which could be an indication for holding proportional assumption.

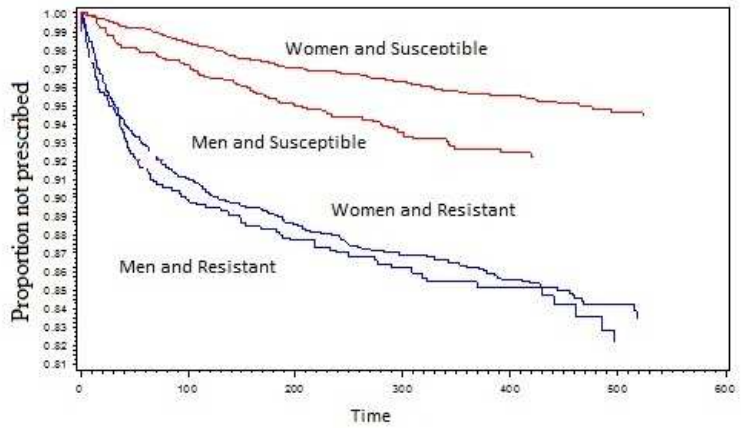

Figure 5: Kaplan-Meier curve for sex by susceptibility result of a patients with *E. coli*.

Figure 5 displays the Kaplan-Meier estimates of susceptibility results for men and women. As seen from figure 3 and 4, the behavior of the curves were similar. Also men who were resistant had higher proportion of prescription than women with resistant susceptibility result at a given time. Similarly, men with susceptible result had a higher prescription than women with susceptible result prior to urine sample.

4.3.2 Proportional Hazards Model

From the Kaplan-Meier estimates, the curves look parallel for sex strata and also susceptibility result strata, the cox proportional hazards model was fitted, susceptible result and women as a reference, and results were displayed in table 9.

Table 9: parameter estimates and standard errors for cox proportional hazards model for patients with *E. coli*.

Parameter	Estimate	Standard error	p-Value	Hazard ratio
Result	1.1435	0.0751	<.0001	3.14
Patient_Sex	0.2579	0.0862	0.0028	1.29

The parameter estimates were given in log (natural logarithm) scale. The hazard of a patient to had prescription prior to urine sample was significantly different for resistant and susceptible as well for men and women. However in this scenario the correlated nature of the data was ignored which underestimated standard errors of the parameters leading to incorrect conclusions (Therneau and Grambsch, 2000). Table 10 shows, a susceptible result and women as a reference, the cox proportional hazards model corrected for the standard errors of the parameters (robust variance). The parameter estimates remain same, however the standard errors increased.

Table 10: parameter estimates and standard errors for cox proportional hazards model corrected for variance (robust variance) for patients with *E. coli*.

Parameter	Estimate	Standard error	p-Value	Hazard ratio
Result	1.1435	0.0953	<.0001	3.14
Patient_Sex	0.2579	0.1226	0.0353	1.29

The estimate of a susceptible result and women (in log scale) were positive, indicating a higher hazard of being prescribed. That is, the hazard of women to had been prescribed prior to urine sample was 1.29 times men holding other covariate constant. The hazard of being prescribed is 3.14 times larger for patients that were susceptible compared to patients that were resistant holding patient sex constant.

5 Discussions and Conclusions

This study was conducted to assess the possible influential and associated variables on resistance of *E. coli* against trimethoprim and trimethoprim/sulfonamides. To answer any study objective, understanding the nature of the data was the crucial step. In this case data had a repeated nature imposing correlation among these measurements with binary response. Therefore, a model which takes into account the correlated nature had to be fitted. The generalized estimating equation (GEE) approach proposed and developed by Zeger and Liang (1986) was used for data collected in longitudinal, nested, or repeated measures designs to estimate more efficient and unbiased regression parameters.

GEE and cluster weighted GEE were fitted. The parameter estimates from weighted GEE and unweighted GEE analysis would be expected to differ substantially only if there were large differences in the number of repeated measurement,"cluster size", and/or proportions of susceptibility results;resistant and susceptible were different.

The main finding of this study as that the number of prescriptions of trimethoprim and trimethoprim/sulfonamide in a patient increased the probability of *E.coli* found in urine samples to be resistant. Another variable associated with *E. coli* resistance was patient sex. Men had a higher chance (21% more than women) of developing resistance against trimethoprim or trimethoprim/sulfonamide. Also the time between prescription and urine sample had a significant effect on resistance of *E.coli* against trimethoprim or trimethoprim/sulfonamide. That is the larger the difference the less probable it will be susceptible, however the shorter the time the higher chance the *E. coli* isolate to be resistant. The worst scenario for susceptibility result to become resistant was from 0 to 30 days. In some articles (Robert and Edward, 1999 and Peter, 2004) the way how antibiotics were administered contribute to the development of resistance, however in this study it had no association for trimethoprim or trimethoprim/sulfonamide. The number of repeated measurement for the scenario *E. coli* were very highly unbalanced, however it had no effect on the susceptibility result.

From survival analysis, the hazard of being prescribed for patients who had a susceptible result was higher than the resistant results. Similarly, the hazard of being prescribed

trimethoprim or/and trimethoprim/sulphonamide for women patients was larger than men patients.

In summary; sex, prior trimethoprim or/and trimethoprim/sulfonamide use (positively associated with resistant result), and time between prescription date and urine sample date (positively associated with resistant result) were predictive factors that the *E.coli* isolate to be resistant.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd ed.* New Jersey:John Wiley & Sons, Inc.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: B. N. PETROV and F. CSAKI, eds. *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267281.

Bauer, A.W., Kirby, W.M.M., Sherris, J.C., and Turck, M. (1966). Antibiotic susceptibility testing by a standardized single disk method. *Am. J. Clin. Pathol*, **36**:493–496.

Bishop, M.C. (2004). "Uncomplicated Urinary Tract Infection." *EAU Update Series*, **2(3)**: 143–150.

Burnham, K., and Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic Approach, 2nd ed.* New York: Springer-Verlag.

Chang, L.S., and Shortliffe D.L. (2006). Pediatric Urinary Tract Infections. *Pediatr Clin N Am*, **53**:379–400.

Clinical Laboratory Standards Institute (CLSI) (2006). Performance standards for antimicrobial disk susceptibility tests; Approved standard-9th ed. CLSI document M2-A9. 26:1. Clinical Laboratory Standards Institute, Wayne, PA.

Coenen, S., and Dvorakova, K. (2006). European Surveillance of Antimicrobial Consumption (ESAC): outpatient antibiotic use in Europe. *J Antimicrob Chemother*, **58**:401–407.

Collett, D. (2003). *Modeling Survival Data in Medical Research,. 2nd ed.* London: Chapman & Hall.

Donnan, P.T., Wei,L., Steinke, D.T., Phillips, G., Clark, R.,Sullivan, F.M., MacDonald, T.M., and Davey, P.G. (2004). Presence of bacteriuria caused by trimethoprim resistant bacteria in patients prescribed antibiotics: multilevel model with practice and individual patient data. *BMJ*, **328**:1297.

Drekonja, M.D., and Johnson, R.J. (2010). *Urinary Tract Infections 2nd ed.* New York:Springer.

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

Goossens, H., Ferech, M., Stichele, R. V., and Elseviers, M. (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet*, **365**:579–87.

Gyssens, I.C. (2011). Antibiotic policy. *Int J Antimicrob Agents*, **38**:Suppl:11–20.

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

Jorgensen, J.H., and Ferraro, M.J. (2009). Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices. *Oxford Journal*, **49 (11)**: 1749–1755.

Kirby, W. M. M., Bauer, A. W, and Perry, D. M. (1959). Single disc antibiotic sensitivity testing of Staphylococci. *A.M.A. Arch. Intern. Med*, **104**:208–216.

Lior, C., and Lars, B. (2004). Background for Different Use of Antibiotics in Different Countries. *CID*, **40**:333.

Lu, L., and Liu, C. (2008). Analysis of Correlated Recurrent and Terminal Events Data in SAS ®. Statistics & Analysis.

McCullagh, P., and Nelder, J.A. (1983). *Generalized Linear Models.* London: Chapman & Hall.

Molenberghs, G., and Lesaffre, E. (1999). Marginal modeling of multivariate categorical data. *Statistics in Medicin*, **18**: 2237–2255.

Molenberghs, G., and Verbeke, G.(2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Mysorekar, I.U., and Hultgren, S.J (2006). Mechanisms of uropathogenic *Escherichia coli* persistence and eradication from the urinary tract. *Proc Natl Acad Sci USA*, **38 (103)** :14170–14175.

Naber G.K., Wagenlehner, E.F., and Weidner, W. (2011). Principles of Bacterial Urinary Tract Infections and Antimicrobials. *Springer Specialist Surgery Series*, 91–103. DOI: 10.1007978-1-60327-239-1_10.

Peter, D.C. (2004). Antimicrobial Use in the Treatment of Calf Diarrhea. *J Vet Intern Med*, **18**:8–17

Robert, O., and Edward, S.W. (1999). Urinary Tract Infections in Adults. *Am Fam Physician*, **59(5)**:1225–1234.

Therneau, T.M., and Grambsch, P.M. (2000). *Modeling Survival Data, Extending the Cox Model*. New York: Springer.

Vellinga, A., Tansey, S., Hanahoe, B., Bennett, K., Murphy, A.W. and Cormican, M. (2012). Trimethoprim and ciprofloxacin resistance and prescribing in urinary tract infection associated with Escherichia coli: a multilevel model. *J Antimicrob Chemother*, doi: 10.1093jacdks222.

Wang, M., Kong, M. and Datta, S. (2010). Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Stat Methods Med Res*, **20**, 347–367.

White, D. G. P. D., Alekshun, M.N., McDermott, P.F. and Levy, S.B. (2005). Frontiers in antimicrobial resistance : a tribute to Stuart B. Levy. Washington, DC, *American Society for Microbiology*.

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

Williams D.A.(1982). Extra-binomial variation in logistic linear models. *Appl Statist*, **31**:144-148.

Appendix

Table 11: parameter estimates and standard errors for GEE with independent working correlation for patients with *E. coli*.

Effect	Parameter	Estimate	Standard Error	p-value
Intercept	β_0	-0.7048	0.0680	$<.0001$
Prior_Trimeuse	β_1	0.4763	0.0862	$<.0001$
Patient_Sex	β_2	-0.2169	0.0755	0.0041
$\leq$ 30 days	β_3	1.3792	0.2351	$<.0001$
(30–60] days	β_4	1.1822	0.2856	$<.0001$
(60–90] days	β_5	0.5823	0.2615	0.0260
(90–180] days	β_6	0.3733	0.1731	0.0311

Table 12: parameter estimates and standard errors for GEE with "cluster size" as a covariate accounting informativeness of cluster size for patients with *E. coli*.

Effect	Parameter	Estimate	Standard Error	p-value
Intercept	β_0	-0.7979	0.1420	$<.0001$
Prior_Trimeuse	β_1	0.3374	0.1338	0.0117
patient_Sex	β_2	-0.2174	0.0745	0.0035
$\leq$ 30 days	β_3	1.4559	0.2369	$<.0001$
(30–60] days	β_4	1.2416	0.2935	$<.0001$
(60–90] days	β_5	0.6465	0.2680	0.0159
(90–180] days	β_6	0.3395	0.1840	0.0650
clustersize	β_7	0.0482	0.0632	0.4453

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Modelling the Persistence of Trimethoprim and Sulfonamide Resistance in Escherichia coli Bacteria

Richting: **Master of Statistics-Biostatistics**
Jaar: **2012**

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,

Leykun, Melkie Chernet

Datum: **14/09/2012**