Article

Differences in Household size, Employment Status and Ability to pay for the service, are Associated with Distance Travelled for Inpatient Care in Kenya

Ngugi Mwenda [1,*], Ruth Nduati [2], Mathew Kosgey [1] and Gregory Kerich [1]

[1] School of Science and Aerospace Studies, Department of Mathematics, Physics and Computing, Moi University, Kenya; mkosgei12@gmail.com (M.K.); Kerichgregoryy@gmail.com (G.K.)

[2] Department of Pediatrics, University of Nairobi, Kenya; ruth_nduati2000@yahoo.com

* Correspondence: samwenda87@gmail.com

Background: Distance to a health facility for inpatient care in developing countries has been a huge hindrance towards achievement of the Sustainable Development Goal (SDG) three. The United Nation encourages countries to research on access to inpatient care, so as to form health policies based on data.

Methods: Data on four hundred and eighty-one participants of all ages from forty-seven counties in Kenya who sought inpatient care in Kenya in 2018 were analyzed. Distance to a health facility was captured as a continuous variable, and was self-reported by the respondent. The response exhibited a discrete mass at zero and continuous characteristic, therefore a Tweedie distribution was adopted for modelling. Due to the correlation nature of clustered data, we embraced the Generalized Estimating Equations (GEE) approach with an exchangeable correlation. Since no standard software was available to analyze this problem, we developed a R functions. We assessed best model fit using the Quasi Information Criteria (QICu) and R^2 criteria, in which the lowest value for the former and the highest for the later are preferred.

Findings: Differences in employment, ability to pay for the service and household size are associated with distance covered to access government facilities.

Interpretation: Poor people tend to have large households and are more likely to live in rural areas and slums, thus are forced to travel for longer distance to access the inpatient care. Compared to unemployed, the employed could have better socio economic status and possibly live within reach of the inpatient health facilities, therefore travel less distances to access. Longer distances are associated with high payments, signifying some form of specialized treatment care due to complexity of the medical cases, that are expensive to treat.

Keywords: distance; inpatient care; SDG's; Kenya; Tweedie; clustered

Introduction

Inpatient care is defined as a case where an individual is hospitalized for more than twenty-four hours, and reflects a more serious health problem. There were estimated 1.2 million Kenyans who required this services in 2013, and the number is predicted to increase exponentially in the upcoming decade [1]. Among those seeking care, various factors are key in predicting distance travelled. For example, differences in wealth index would determine the distance travelled. For example, those on the richer wealth quintile can afford to pay fare to any facility, pay insurance premiums that can guarantee admission at any facility or having the financial strength to pay cash at any given facility of choice. In contrast, those on the poorer quintiles have less choice to make on the type of facility for

care as they are limited by finances. In as much as the government facilities are much more affordable, and are the best choice for care, most are miles away and out of reach from places of residence.

In an effort to mitigate this, the government has increased the set-up of as many inpatient services as possible, which include upgrading of health facilities which used to offer outpatient services only by equipping them with machines that are needed to provide inpatient care. However, most of this upgrade would require essential services such as water and electricity, accessible road and housing which limits some facilities for upgrade. Some of this facilities are found in the rural areas and slums which serve a large number of people, which means their upgrade would be a milestone for the residents. What this means is that the poor who live in this areas continue experiencing difficulty when they need inpatient care. As a result, potential ease of access policies for the inpatients service should be invented.

Distance to access inpatient services can determine the well-being of a population, and is potentially linked to an individual survival. For example, there has been a link of longer distances with worse health outcomes, including more length of stay in hospital and non-attendance at follow up [2], and to the worst, patients fatality [3]. Study conducted in Zambia found that longer distances and the lack of geographic access to the much needed obstetric care by pregnant mothers, explain why there are still fatalities due to deliveries without skilled care[4], and in Tanzania there is an increased child mortality due to difficulties in access to a health facilities [5].

In contrast, closer or shorter distance to a facility were associated with utility and better health outcomes in sub Saharan Africa [6]. In an event of emergency, distance could be a defining factor on a patients survival, with longer distance predicting higher mortalities [5]. Studies across some developing countries, like Bangladesh [7], Kenya [8], Nigeria [9, 10], Afghanistan [11] and Burkina Faso[6] points out the importance of distance, in predicting a health outcome.

While a correlation of distance and decay exist, with those further associated with underuse and those closer associated with proper use, there is little evidence to show how this translates to the heath outcome. Therefore, distance travelled to acquire the much needed inpatient services require more investigation, and thus forms the basis of this paper.

Our analysis focuses on the secondary analysis of the Kenya Household Health Expenditure and Utilization Survey, collected in 2018, and the question of interest was; *Q68. What distance did <name> cover in Kms to get to the inpatient facility?*

This paper builds on our previous work on non-normal responses analysis under GEE by Mwenda *et al* [12] and adopts the approach by Taryn Swan [13] and Kurz Christoph [14]. Kurz fitted health care utilization cost data using a Tweedie distribution, but his works were based on generalized linear model, meaning correlation was not considered. In contrast, Taryn and Mwenda *et al* considered a decaying correlation with time, but with different applications to rainfall and health data respectively.

Our data is based on the clusters, which exhibit patient to patient correlation characteristic, meaning if we use the previous methods (correlation decay) we will have wrong results. Therefore, due to clustered correlation nature of our data within counties, our work proposes a new approach to model the response using the Tweedie distribution by considering what we refer to as '*decay distance with constant correlation*' and thus propose to use the exchangeable correlation structure under a GEE framework.

Our main goal is seeking to understand what covariates are best associated with distance travelled for inpatient care in Kenya. Since there is no software that can analyze this as far as we are concerned, we aim to create R function with a Tweedie distribution and exchangeable correlation structure under the GEE frame work. Due to difficulty in linking inpatient admissions and accessibility, we rely on the self-report from respondents on the distance they covered to access the health facility.

Method

Statistical Literature review

Tweedie distributions have been applied widely in modelling non-normal response data with discrete mass at zero since they are flexible to incorporate skewness without any data transformation. Most of the methods suggested in literature for analysis of such data, majorly consider data transformation [15], two-way analysis [16] and Bayesian methods [17]. However, these methods are not efficient for our approach due to correlation nature of our data within clusters. Approaches proposed for analyzing non-normal data in the GLM framework have limitation of ignoring the correlation, which may exist amongst the subjects who belong to the same cluster and also, the methods require a specification of a full likelihood. This means, if you miss-specify the likelihood then you get wrong results. Our proposed approach, to use the quasi likelihood methods, have the advantage that they only require specification of how the mean relates to the covariates. They are also very flexible in that in an event of misspecification of the correlation structure the estimates are still plausible. Moreover, our methods are easy to modify and adapt.

Previous evidence suggests influence of covariates on distance to health facilities, however research gap on selection of best fitting covariates remain. The aim of this paper is to investigate what combination of covariates will influence the distance a Kenyan citizen will travel to seek inpatient care. This new work enhances the application of Tweedie distribution to understand the influence of given set of covariates on distance.

Statistical analysis and Selection of variables

Data from the Kenya Household Health Expenditure Utilization Survey **(KHHEUS)** inpatient data was obtained, cleaned and coded. We investigated 13 covariates against our dependent variable, distance. Type of residence was categorized into two (rural and urban). Five wealth index quintiles ranging from richest to the poorest was constructed as a score from ownership of different household assets using the Principal component analysis as described by Filmer and Prictchett [18]. Grouping of the education was adopted due to justification and methods provided by Rippin *et al* [19]. We had four categories; Those who never went to school (Those under 3 years of age and those who responded, "Never went to School"), Lower education (pre-primary, primary and informal (madrassa)) intermediate (secondary, vocational and college) and higher education (university degree and above). Grouping of the age groups for employment followed the definition of employment rate by age group defined by the OECD. We divided into four age groups categories. Category 1 considered those aged below 15 and above 65, as they are considered unable to work, Category 2 considered age group 15-24years as those entering labor market following education, Category 3 considered age group 25-54 as those in prime working lives and finally category 4 considered age group 55-64 as those passing the peak of their career and approaching retirement [20].

We categorized household size into three categories small (1-3members) medium (4-6 members) and large (7+ members). We also categorized duration of admission into fewer days (1-5), medium days (6-20) and more (21+). Access to insurance was classified as yes or no. Headship of a household was classified into two, male and female headed. Amounts paid were categorized into three groups, low paying (1-3000KES), medium pay (3001-10,000KES) and higher pay (10,000+ KES).

Our dependent variable, distance was assumed to be 0 km for any value captured and is less than 2 km, considered by other studies in Kenya [21].

Data Availability

The datasets are freely available at Kenya National Bureau of Statistics (KNBS). Interested researchers should register an account to access the datasets at https://www.knbs.or.ke/kenada/. The authors confirm that others would be able to access these data in the same manner as the authors and that the authors did not have any special access privileges that others would not have.

Statistical Methods

In this study a Tweedie distribution is used to model predictors of distance for inpatient care. Justification to use the Tweedie distribution is provided by **Figure 1**, which shows that distance as the dependent, has a discrete mass at zero and a continuous characteristic. In addition, **Table 1** shows

that the data is right skewed with a skewness value 4.80. We analyzed the distance with the covariates to determine the best combinations to explain any existing association.
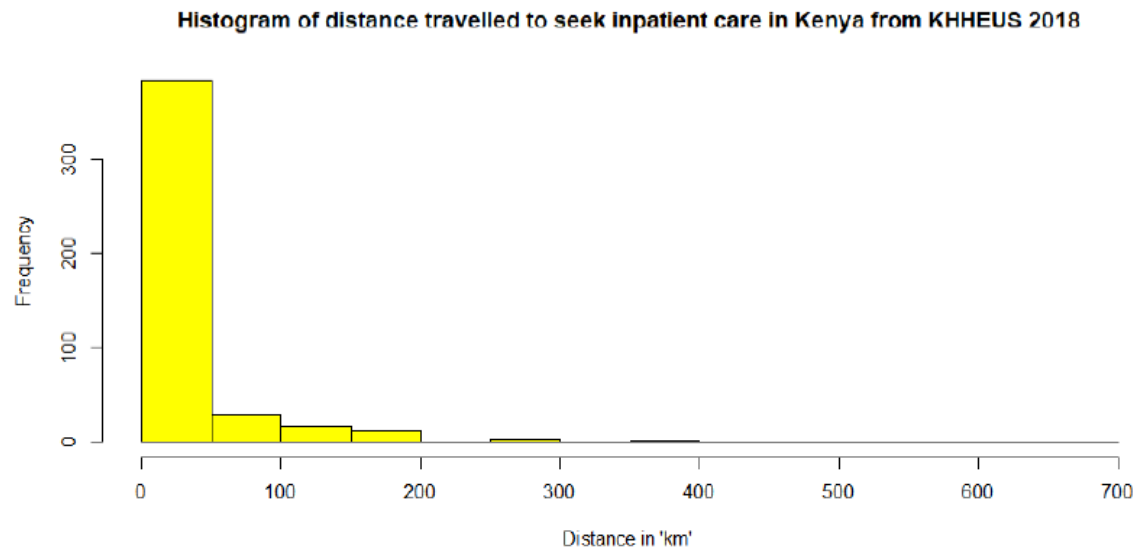

Figure 1. Frequency distribution of Distance travelled to access inpatient care in Kenya KHHEUS 2018.

Table 1

Statistic	Distance travelled (km)
Minimum	0
1st Quarter	3
Median	10
Mean	32.1
3rd Quarter	30
Maximum	700
Shape	Right Skewed
Skewness	4.80

Table 1. Summary statistics of the distance travelled for inpatient care.

Detailed methodology is in Appendix 3.

Multiple regression analysis under GEE framework using the Tweedie distribution and Exchangeable correlation structure

Here, we adjusted for household head gender, education, age, household size and wealth index in the regression model. This we used the exchangeable correlation under GEE approach using Tweedie distribution. We checked the model fit to select the best model using the QICu proposed by Hardin and Hilbe ([22] page 171-172) which is an extension of QIC proposed by Pan [23], following the AIC developed for GLM [24]. QICu imposes penalty based on model complexity to ensure few covariates are used to achieve model parsimony.

Data which are in .sav format, were imported into R statistical software version 3.6.3 [25] for cleaning, reformatting, recoding and analysis.

Results

We present 10 competing models with distance that demonstrates the best fitting model with the lowest QICu as seen in **Appendix 2**. We used the approach of backward selection [26] as a proxy

to identify the best predictors for distance under a GLM. However, our model output and interpretation is only based on distance adjusted for the respective covariates in the GEE framework. We added the covariates into the model and computed their QICu and R^2, we then removed the covariates one by one and checked whether the changes were improving the model fit.

The reported model, which is model 7 from **Appendix 2,** with the best fitting covariates is written as

$\log(\mu) = 3.093 + 0.222\,paid.middle + 1.226 paid.highest - 0.523 employed +$

$0.092 householdsize.mideum + 0.471 householdsize.large$

1. Where μ is the expected distance traveled to access inpatient care.

2. Ability to pay (middle, highest) takes a value 0 or 1 depending on which category is being assessed. Group (lowest) is the reference category.

3. Employment (yes for employed) takes a value 1 if the respondent said they were employed

4. Household size (medium and large) takes a value 0 and 1 depending on what is being assessed. The reference category (small household size) doesn't appear in the equation.

This model reported an $\phi = 6.12$, $\alpha = 0.045$, $R^2 = 9.96\%$ and $QICu = 13158.23$ with a $p = 1.64, 95\% CI\,(1.59, 1.68)$

To interpret the coefficients, which are captured in logarithmic form, we need to take the exponentials. From the given model, all factors kept constant, the population average distance to a government inpatient center in Kenya is approximately, exp(3.093) which is 22.04 km.

Compared to the least pay category (1-3000KES), the citizens in the middle pay category (3001-10,000 KES), travelled 1.24 times, while the ones who paid the highest category 3.40 times.

The employed covered half the distance to a health facility for inpatient care compared to the unemployed, 0.59times.

Compared to small household sizes (1-3), medium household sizes (4-6) travelled 1.096 times while the largest household sizes (7+) covered 1.60 times compared to small houses

Discussion

This work has demonstrated the use of an alternative technique for clustered and correlated non-normal responses which depict a discrete mass at zero, under generalized estimating equations. The paper presents the best set of covariates in predicting distance travelled by Kenyans to access inpatient care from the forty-seven counties. Data from each county are representative, and the pooled data contributed a substantial information about distance for inpatient care. A set of potential covariates were investigated to better understand their effects.

The model, without covariates showed that on average, a Kenyan seeking inpatient care travelled a distance of 22.04KM. Though a national average, the cost in terms of fare will differ substantially, in that the road terrain which is the most preferred for accessing hospitals is very different in Kenya. Some roads are all weather but others are seasonal, meaning during the rainy times, accessibility is greatly hampered. A health system performance is described using equity, measures or assessed using the health service distribution, access and utilization [27]. The access is mostly determined by cost and distance. This means irrespective of availability of a service in a hospital, if not utilized by the target group, then its fully utility cannot be actualized. The aim of SDG 3 is to *'Ensure healthy lives and promote wellbeing for all at all ages'* and this work shows the importance of distance in measuring this goal.

Most inpatient care is usually critical and require specialized attention by a medical expert, therefore distance to access could determine survival or fatality. Though some studies have not linked accessibility to use [29], there has been evidence that ease of access could potentially save life, as some life threatening conditions are worsened by accelerated duration to see the physician. For example, when a patient suffers a heart attack, how fast they get to the hospital could determine their survival.

In our inpatient count data, some patients were admitted to health centers operated by the government. Although not advanced very well to handle extreme cases, they are well equipped to handle some cases like admissions for child birth. However, in case of any complications from the child birth, in eventuality the woman needs caesarean services, then it means they have to be referred to a larger facility that can offer that.

What this means is that major conditions would still need referrals to major hospitals thus access for inpatient care needing this services at the lower level hospitals will still remain a challenge.

Noor *et al* [28] had investigated access to any government health center, and reported distance of less than 10km. Although they focused on only 4 districts in Kenya, they were diverse and could be used as a proxy for national estimate of the distance. However, their focus on the smaller health centers, are mostly used for outpatient care, could be misleading when predicting for inpatient access.

However, it is important to note that distance in Kenya is very hard to predict, because of the differences in terrain and road types (Tarmacked, marram, earth), thus the low value of the R^2 of 9.9% is reasonable.

Distance for inpatient care is a cornerstone to the Kenyan government as it tries to achieve the universal health coverage. The main goal of universal health coverage (UHC) is to ensure that every citizen has access to quality healthcare services, however, this can only be achieved if the distance to access is reasonable and achievable. What this means is that for the government to achieve the general objective, then it needs to do more on opening up access for inpatient services. It is also evident that as far as the access to the inpatient services is generalized, Kenya has a unique geographical terrain. For example, it may be easier to access a facility that is further away in an urban area than fewer km in the rural area. This is because, most urban cities have good road network, making access easy. If at all the government wants to increase access, then more needs to be done to improve road network infrastructure in the rural areas.

Our results, have shown the evidence that high costs are associated with longer distance travelled to access inpatient care. We can interpret this in two ways. First, the cost incurred could signify an expensive procedure or care. Two, those with the ability to pay could choose facility which are far, though expensive, and the required care isn't complicated, as they could be having more confidence in the hospitals of better care.

Those who had the highest ability (10,001 + KES and above) to pay seem to travel more distance for inpatient care up to 3.40 times compared with those with lowest (1-3,000 KES) ability, and those in the middle (3,001-10,000 KES) seems to travel 1.24 times more. Ability to pay give a person freedom to choose any facility they are comfortable for inpatient care and mostly, higher hospital figures are associated with complex medical care needs and procedures. For example, a caesarean will cost more than normal delivery, although all will need the inpatient care. The differences will arise in that a referral hospital could be more suited to handle caesarean from a woman with pre-eclampsia that other hospitals. This is because sophisticated medical equipment's would be required for the procedure, and are mostly found in large referral hospitals. Therefore, a patient with stronger financial muscles will travel longer distance for the procedure and pay higher medical cost. The one with low ability to pay will check in at the closest and most affordable facility.

Our result points out that Kenyans travelled longer distance for complex medical procedures or for better services which may not necessarily be found in the closest inpatient health facility. Others could have travelled longer distance for privacy of the disease. For example, a person could be more comfortable being admitted for inpatient care for a sexually transmitted diseases (STD) in a hospital further away from home. However, our results may not be comprehensively conclusive, and a further investigation on why those who travel longer paid more for the services should be established.

In Kenya, after adapting the new constitution in 2010, health services were devolved with an aim of proper and closer management at the low levels. However, due to limit in budgets and other indirect effects like road and electricity, there has been a slow growth in terms of the hospitals upgrades for inpatient care. For example, setting up an inpatient care deep in the rural set up without a good road network or proper supply of electricity would be meaningless in terms of serving the people. The result of this is that the facilities will mostly be set up in centers where such services can

be accessed easily. This means people in remote areas will still have to travel for longer distance to get this inpatient services.

Inequalities in employment opportunities also determine the distance travelled to access inpatient care. It is evident that employed will cover half the distance the unemployed covers, results supported by Allin et al [30]. This means that the unemployed have less financial ability will be forced to use the facilities that are within reach, and may be limited by lack of sufficient money to access better facilities for specialized treatment, in case they require the same. Additionally, the employed have an advantage of affording to live where larger inpatient facilities are found. For example, you expect to find large referral and specialized hospitals in capital cities or large towns so that they are accessible and serve many people. The facilities also need to be easily accessed, connected to uninterrupted supply of water and electricity. Most employed category of people would choose to live in the places where such services can be found.

Family size was the last covariate that determines distance travelled, with middle and larger household services travelling longer distances compared to small household sizes. This differences could be explained by the fact that larger sizes are mostly found in rural areas and slums compared to urban areas. It is easy to raise a large family in rural since things like food and accommodation are cheap and affordable. Mostly, people in rural areas live in their ancestral land where they also farm most of their food, thus no cost implication on housing and food. This shows evidence that there is need to improve inpatient facilities in the rural areas and slums. Without a strong policy focus to support equal support in accessing the inpatient services in Kenya, prioritizing the rural areas and slums, open up job opportunities, and encourage smaller families, then the dream of achieving UHC will still remain a dream.

This work is the first to estimate distance for inpatient care in Kenya, analyzing all the responses from the forty-seven counties. This provides the best estimate and evidence on which policies to formulate. This area has been under-studied with many researchers focusing on both inpatient and outpatient care, however such analysis could lead to wrong conclusion and bad formulation of policy. For example, Noor et al [28] reported distance to access of less than 10 km, which means if policy are based on this, then it means every person in fact has access to health care. However, as stated earlier, most of this facilities are for outpatient meaning if policies are based on this, then we may not improve the much needed inpatient care that requires sophisticated and complex procedures as well as the doctors.

A further advancement of this research would be to focus on individual analysis of the 47 counties. This is out of scope of this work that has only focused on the population average. This is an open room for further research analyzed per county. Although some studies have found no association with use [29] better access regardless of the quality is still influenced by distance.

A drawback from previous researches is analyzing distance using summaries reports. This is where the researchers don't dig deep into the data and only reports the average from the summary statistics. From **Table 1** alone, we could have reported a median of 10 km, which may be misleading as we didn't factor in skewness and correlation that exist in the data, which if analyzed could give more insight. This shows that our advanced statistical analysis has provided more meaningful interpretation to the data by factoring both skewness and correlation.

This analyses have limitations. Firstly, the data had a lot of missing information which made it very difficult to impute. So we only used data that were full. New complex statistical methodologies for predicting non normal data need to be developed.
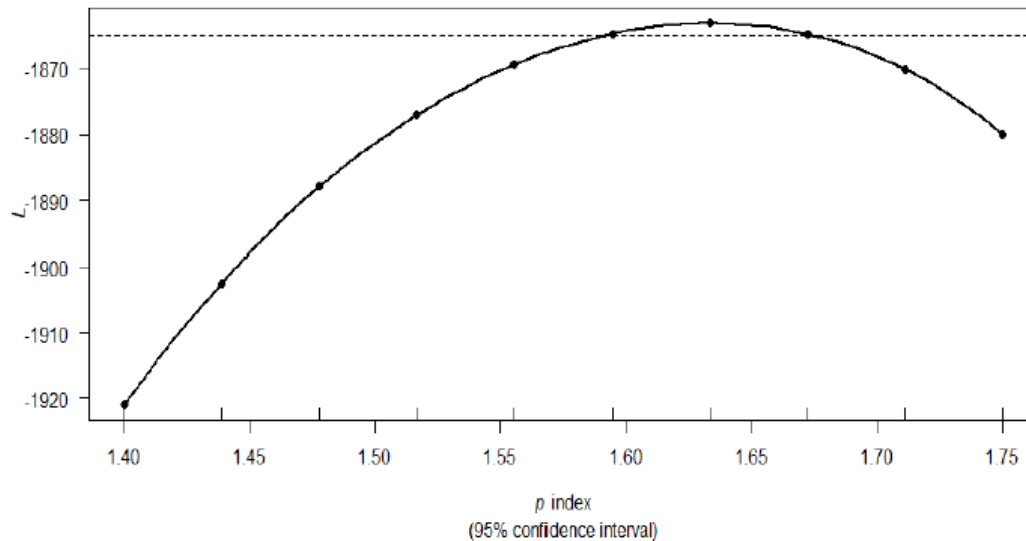
Figure 2. The profile log-likelihood plots for the distance travelled for inpatient care in Kenya using gender of the household head, household size and Education as covariates. The plot estimated the *p* as 1.63 (1.59,1.67), with the dots representing 95%. The solid line a cubic-spline smooth interpolation joining the points.

Conclusion

We have demonstrated a new approach of handling correlated non-normal data, and created an r code **Appendix 1** and GitHub link (https://github.com/samwenda/Tweedie-with-Exchengable-Correlation). Our approach has demonstrated how distance decay can affect access to much needed health services. Our approach can be used in other datasets which have a discrete mass at zero and have correlation within clusters. Our approach has also shown a new way of calculating the denominator as provided by Hardin and Hilbe book page 91 as shown in **Appendix 2**.

Policies targeting on having more government facilities at the rural areas and slums should be targeted since that is where we have large populations. Also, policies targeting smaller families that are manageable should be encouraged to ensure people have families who they can afford to give the best health care. Jobs availability will also increase flexibility on the choice of a facility. Sophisticated services should be brought close to the person on the ground to ensure they don't travel longer for the much needed services.

Appendix 1: Tables and Figures

Table 1. Summary statistics of the distance travelled for inpatient care.

Statistic	Distance travelled (km)
Minimum	0
1st Quarter	3
Median	10
Mean	32.1
3rd Quarter	30
Maximum	700
Shape	Right Skewed

Skewness	4.8

Table 2. Models selection using QICu and R^2

..Model number	Covariates	QICu	R^2	Variance power P(95%CI)	No of covariates
10	Ability to pay, employment status, household size, wealth index, education level, age	13304.7	10.39	1.63(1.58,1.67)	6
9	Ability to pay, employment status, household size, wealth index, education level	13306.16	10.41	1.63(1.59,1.67)	5
8	Ability to pay, employment status, household size, wealth index	13317.38	9.7	1.62(1.59,1.67)	4
7	Ability to pay, employment status, household size	**13158.23**	9.96	1.64(1.54,1.68)	3
6	Ability to pay, employment status	13280.7	9.5	1.64(1.59,1.68)	2
5	Have insurance, place of residence	12733.1	0.19	1.67(1.63,1.71)	2
4	Ability to pay	13066.33	8.38	1.64(1.60,1.68)	1
3	Place of residence	12773.2	0.17	1.67(1.63,1.71)	1
2	Household size	12755.65	0.54	1.67(1.63,1.71)	1
1	Employment	12698.31	1.4	1.67(1.63,1.67)	1

Model 7 was selected as the best model, even though model 9 and 10 had higher R^2 values. We selected the model with the best balance between the QICu and the R^2, in which model 7 fits as the best parsimonious model, with the least covariates with acceptable QICu and R^2 values

Table 3. Run time in seconds in calculating the denominator of the scale parameter.

Method	Run time in seconds
Hardin and Hilbe	0.033
Proposed method 1	0.0319
Proposed method 2	0.0289
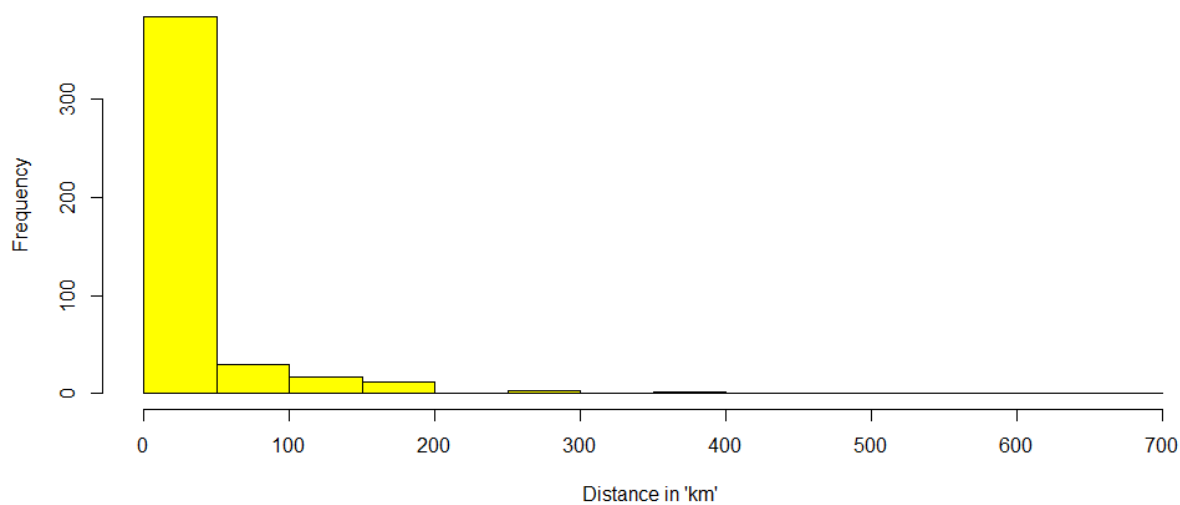
Figure 1. Frequency distribution of Distance travelled to access inpatient care in Kenya KHHEUS 2018.
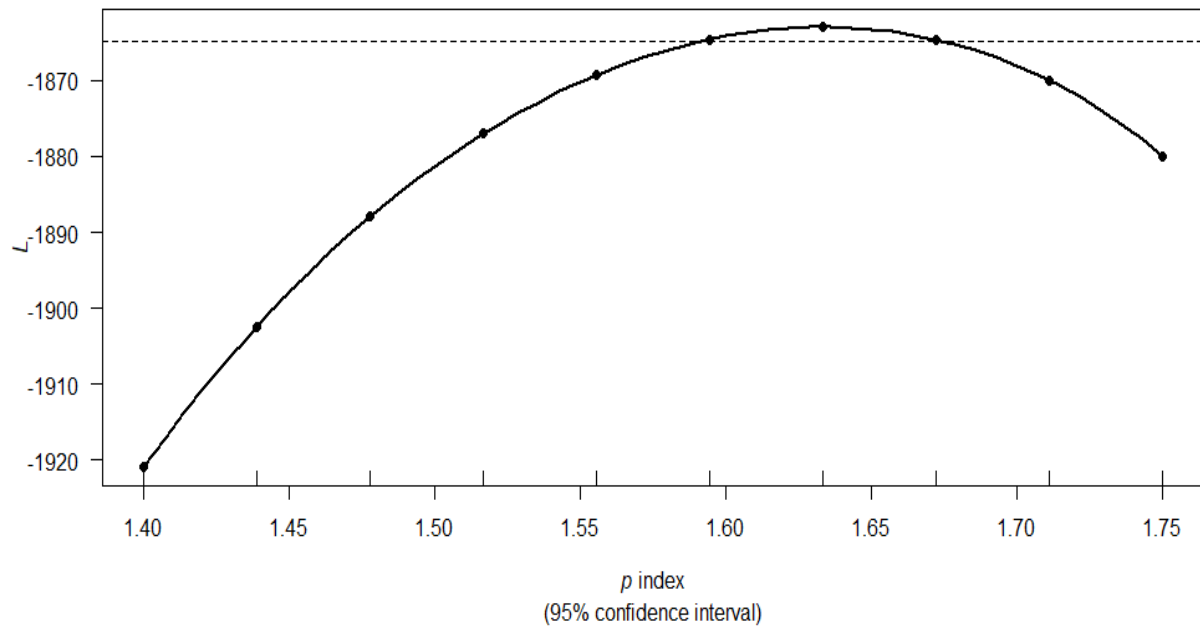
Figure 2. The profile log-likelihood plots for the distance travelled for inpatient care in Kenya using gender of the household head, household size and Education as covariates. The plot estimated the *p* as 1.63 (1.59,1.67), with the dots representing 95%. The solid line a cubic-spline smooth interpolation joining the points.

Appendix 2: Models selection using QICu and R^2

Model number	Covariates	QICu	R^2	Variance power P(95%CI)	No of covariates
10	Ability to pay, employment status, household size, wealth index, education level, age	13304.7	10.39	1.63(1.58,1.67)	6
9	Ability to pay, employment status, household size, wealth index, education level	13306.16	10.41	1.63(1.59,1.67)	5
8	Ability to pay, employment status, household size, wealth index	13317.38	9.7	1.62(1.59,1.67)	4
7	Ability to pay, employment status, household size	**13158.23**	9.96	1.64(1.54,1.68)	3
6	Ability to pay, employment status	13280.7	9.5	1.64(1.59,1.68)	2
5	Have insurance, place of residence	12733.1	0.19	1.67(1.63,1.71)	2
4	Ability to pay	13066.33	8.38	1.64(1.60,1.68)	1
3	Place of residence	12773.2	0.17	1.67(1.63,1.71)	1
2	Household size	12755.65	0.54	1.67(1.63,1.71)	1
1	Employment	12698.31	1.4	1.67(1.63,1.67)	1

Appendix 3: Methodology

The best approach of modelling a mixture of discrete and continuous data that accounts for the discrete mass at zero are the Tweedie distributions. The Tweedie distribution don't have densities that can be expressed in closed form, however, they are flexible enough to encompass discrete

distribution such as poison and continuous distribution such as gamma, in what we refer to as compound Poison-Gamma distribution.

Following Hasan and Dunn [**Error! Reference source not found.**], for $y \in Q$ for some function $c(y, \phi)$ and known function θ and $\gamma(.)$ and dispersion parameter $\phi > 0$, the Tweedie family belongs to EDM defined as a probability function of the form

$f(y; \theta, \phi) = c(y, \phi)\exp[1/\phi\{y\theta - \gamma(\theta)\}]$

$c(y, \phi)$ cannot be written in closed form, and is known as the normalizing constant that ensures that the total integral of Y over the domain Q sums to one.

Following Dunn and Smyth [**Error! Reference source not found.**], $Y \sim ED(\mu, \phi)$ indicate that Y belongs to an EDM with mean given by $E[y] = \mu = \gamma'(\theta)$ and variance given by $var[Y] = \phi\gamma''(\theta)$, and the mean relates to the variance through a mean-variance relationship given by $var[Y] = \phi V(\mu)$ with $V(\mu)$ referred to as the variance function.

The Tweedie's are special cases of EDM that obey a fixed mean-variance relationship of $Var(Y) = \phi\mu^p$ for $\mu = E(Y)$ and constant ϕ and $p \notin (0,1)$. They are denoted as $Tw_p(\mu, \phi)$ with an index parameter p, dispersion parameter ϕ and mean μ. Altering the value of the index parameter p, the Tweedie's can be reduced to some other common distributions which are also EDM members. When $p = 0$, the normal distribution, $p = 1$, the poisson distribution, $p = 2$ gamma distribution and when $p = 3$ we have the inverse Gamma. To note is that for values greater than 2, then Tweedie's have no closed form. Our focus is on distribution with the p index expressed as $1 < p < 2$.(1 and 2 are not included) known as poison-gamma distribution, which have a discrete mass at zero and are continuous.

An important property of Tweedie under GEE, is its ability to accommodate both correlation and right skewedness which is a characteristic of our continuous data. This approach is used in this paper and it complements the work of Taryn Swan, who used the AR(1) correlation structure.

Tweedie regression models allows relation of the mean of distance to the selected covariates. This allows the mean of distance to be modelled as a linear function of covariates using the log link given by $\log(\mu_i) = \beta_0 + \beta_1 x_i +, \dots, \beta_n x_n$ where β_j vectors are regression coefficients that corresponds to the x_j vectors of covariates, all fitted based on Tweedie EDM. Means are calculated to assess the relationship between covariates and the distance.

To fit the models, we need to estimate the index parameter p and the $\beta' s$ from the Tweedie distribution using the GLM framework. This could be computationally difficult but the R package Tweedie [**Error! Reference source not found.**] and statmod [**Error! Reference source not found.**] fit this easily. The calculated index parameter as calculated by the software is shown by **Figure 2** in **Appendix 2**. To estimate the $\beta' s$, we apply the approach of McLuggar and Nelder [**Error! Reference source not found.**] called the Iterative reweighted least square (IRLS) method in the GLM's. These $\beta' s$ are then updated under the GEE framework. As follows

Following Hardin and Hilbe [22], quasi-likelihood estimating equations for GLM without any preference for mean and variance function, with 47 clusters are expressed as

$\Psi(\beta) = \sum_{i=1}^n \left\{ x'_{ik} T \frac{1}{V(\mu_i)} \frac{y_i - \mu_i}{a(\phi)} \right\} = 0$ for $k = 1, 2, \dots, 47$ where T is a diagonal matrix of derivatives $\frac{\partial \mu_i}{\partial \eta_i}$, $V(\mu_i)$ is an $n_i \times n_i$ diagonal matrix which can be decomposed into $V(\mu_i) = T[V(\mu_i)]^{\frac{1}{2}} I_{n_i \times n_i} T[V(\mu_i)]^{\frac{1}{2}}$. The observation in the clusters are treated as independent, but our focus is on the population average for which both the mean and the variance are averaged over all the clusters.

Following Liang and Zeger,[**Error! Reference source not found.**], the identity matrix in the above is replaced with a more general correlation matrix since the variance matrix for data which is correlated doesn't have a diagonal form, as follows $V(\mu_i) = T[V(\mu_i)]^{\frac{1}{2}} CM_i(\alpha) T[V(\mu_i)]^{\frac{1}{2}}$,where the correlation matrix $CM_i(\alpha)$ is estimated by vector α. Proper specification of the $CM_i(\alpha)$ then the $\hat{\beta}$ are consistent and asymptomatically normal. To achieve more efficiency, then there is need to include a hypothesized correlation structure within the clusters. The structure that is suitable for our clustered data without any time dependence is the exchangeable given by $CM_{ik}(\alpha) = \begin{cases} 1 \text{ if } i = k \\ \alpha \text{ if } i \neq k \end{cases}$

Finally, we developed the R code that is used to estimate the β after calculating the best fitting model using the $QICu$ and the R^2 as shown by **Appendix 4**

Appendix 4: R code for the Analysis

#Analysis of inpatient care data in Kenya using GEE and exchangeable correlation under tweedie distribution

#using GEE

rm(list = ls()) # clear all data from memory

options(scipen=999)# gives output without scientific notations

#loading the packages for use

#for stepwise selection of covariates

#for data manipulation

#for working with the tweedie as a response

library(dplyr)

library(haven)

library(stats)

library(statmod)

library(leaps)

library(tweedie)

library(labelled)

library(foreign)

library(purrr)

library(stringr)

#start by reading the data

inpatient <- read_sav("folder_location/khheus_c3.sav")

attach(inpatient) #attach the dataset

#select 1=national refferal, 2=county government hospitals, 5 government health centeres

df1<-inpatient %>% filter(q53==1| q53==2 | q53==5)

#any distance below 2km is coded as zero

df1$dist<-ifelse(df1$q68_dist_km<=2,0,df1$q68_dist_km)

#group the ages into inactive 0-14 and 65 and above, then 15-24, 25-24, 55-64

df1$new_age_work_group<-ifelse(df1$q6bx<15|df1$q6bx>64,1,

                ifelse(df1$q6bx>=15&df1$q6bx<=24,2,

                    ifelse(df1$q6bx>=25 & df1$q6bx<55,3,4)))

##highest education level

df1$school_cat<-ifelse(df1$q8==8|df1$q8==9|df1$q8==10,1,

            ifelse(df1$q8==1|df1$q8==2|df1$q8==7,2,

```
                    ifelse(df1$q8==3|df1$q8==4,3,4)))


###days admitted, categorize 1-5 days, 6-20 days and 21 days and above

df1$number_day_admit_cat<-ifelse(df1$q51>=1 & df1$q51<6,1,

                    ifelse(df1$q51>5 & df1$q51<21,2,3))


###employment versus not employed

df1$emp_cat<-ifelse(df1$q11x2==1|df1$q11x2==2,1,0)

##grouping of the monies paid 1-3,000, 3,001-10,000 and above 10,000

df1$paid_cat<-ifelse(df1$q60_11_grp==1|df1$q60_11_grp==2,1,

            ifelse(df1$q60_11_grp==3|df1$q60_11_grp==4,2,3))


#select the variables for use from the large data set and rename them

new_dfa<-df1                                                        %>%
select(q68_dist_km,dist,county,resid,q51,hhhead,number_day_admit_cat,q59,q65,

            q71a,wlth_index,q4,q6b,q6bx,new_age_work_group,q7,

            agegrp1,agegrp2,hhsize_sum, hhsize_group, q8,school_cat, q11,q11x,q11x2,

            emp_cat,q72orig,q63new1,q60_11_sum,q60_11_grp,paid_cat)%>%

   rename(days_admitted=q51,name_paid_INP_service=q59,satisfied=q65,dist_1=q68_dist_km,

        accompanied_to_hosp=q71a, gender=q4,age_1=q6b,age_2=q6bx,school_lev_status=q7,

        age_group_5=agegrp1,age_group_10=agegrp2,
highest_educ_level=q8,main_employment_status=q11,

        employment_1=q11x, employment_2=q11x2,access_insurance=q72orig,

        had_cash=q63new1, total_paid=q60_11_sum, total_paid_group=q60_11_grp)


#identify any missing data

sapply(new_dfa, function(x) sum(is.na(x)))

#filter data with missing

d1<-new_dfa %>% filter(!is.na(hhsize_group)) %>% filter(!is.na(paid_cat)) %>% filter(!is.na(dist))

#create a new dataset with the covariates of choice and the indepedent.

#Weuse the backwards method using the

d2<-d1 %>% select(dist,county,resid,number_day_admit_cat,

            satisfied,wlth_index,new_age_work_group,hhsize_group,

            school_cat,emp_cat,access_insurance,had_cash,paid_cat)
```

```
#----------------------------------------------------Gives a proxy for best variable combination

d3 <- d2[-2]  #remove the county since it is our ID

head(d3)

null=lm(dist~1, data=d3)

full=lm(dist~., data=d3)

step(null, scope=list(lower=null, upper=full), direction="forward")


step(full, data=d3, direction="backward")

step(null, scope = list(upper=full), data=d3, direction="both")

#----------------------------------------------------------------------------------------Ends here


attach(d2)

####power tweedie starts here

var.power=tweedie.profile(dist~factor(emp_cat)+factor(paid_cat)+factor(hhsize_group),

              p.vec=seq(1.4,1.75,length=10),

              do.ci=TRUE, method="saddlepoint", data = d2)

var.p=var.power$p.max

var.p

var.power$ci

modGlm<-glm(dist~factor(emp_cat)+factor(paid_cat)+factor(hhsize_group),

        family=tweedie(var.power=var.p,link.power=0),x=TRUE, data=d2)


##-------------------------------------------starts here

fitted.v<-modGlm$fitted.values

beta.v=modGlm$coefficients

phi_old = var.power$phi.max

n=length(dist)

r=modGlm$rank

deviance1=sum(tweedie.dev(dist,fitted.v,var.p))

deviance2=100*deviance1

ep = 1e-8
```

```r
while (abs(deviance2 - deviance1)/(0.1 + abs(deviance1)) > ep) {

p.residuals=(dist-fitted.v)/sqrt(fitted.v^p)

p.residuals

new.phi<-sum(p.residuals^2)/(n-r)

new.phi


d1$resd2<-p.residuals

res_total_n_for_alph<-d1 %>%

  group_by(county) %>%

  summarise(resd = sum(combn(resd2, 2, FUN = prod)),n = n())%>%  mutate(sampl_n = n*(n-1)/2)

res_total_n_for_alph


sum_res_alpha<-sum(res_total_n_for_alph$resd)

sum_res_alpha

sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)

sum_n_alpha

alpha<-(1/new.phi)*(1/(sum_n_alpha-r))*sum_res_alpha

alpha


i=0

R <- matrix(NA, n,n)

for (i in 1:n){

  for (j in 1:n){

    if (i==j){

      R[i,j] <- 1.

    } else {

      R[i,j] <- alpha

    }

  }

}

R

A<-diag(fitted.v)^(p/2)

V = (A %*% R  %*% A)
```

```
g<-modGlm$model

xmat=modGlm$x

D = matrix(nrow=n,ncol=r)

for(i in (1:r)){

  D[,i]=fitted.v*xmat[,i]

}

C=t(D) %*% solve(V) %*% D

 B=t(D) %*% solve(V) %*% (dist-fitted.v)

beta1=beta + (solve(C) %*% B)

fitted.v<-exp(t(beta.v) %*% t(xmat))

fitted.v<-as.vector(fitted.v)

deviance1<-dev

deviance2<-sum(tweedie.dev(dist,fitted.v,var.p))

}

quasi.likeli<-sum((dist*fitted.v^(1-var.p)/(1-var.p))-((fitted.v^(2-var.p))/(2-var.p)))

QICu<-(-2*quasi.likeli)+(2*r)

QICu
```

```
marginal.sum=(1/n)*sum(dist)

numerator=sum((dist-fitted.v)^2)

denominator=sum((dist-marginal.sum)^2)

R.sq = 1 - (numerator / denominator)

R.sq
```

##---Ends here

Appendix 5: Proof for alternative calculation of the denominator for Exchangeable correlation for the scale parameter

Proof by induction

The denominator of an exchangeable correlation on page 98 in Hardin and Hilbe [22] book, to calculate the alpha is given by

$$\sum_{i}^{n_i} 0.5 * n_i(n_i - 1) - p$$

We propose that the same can be implemented using combinations. We suggest the following

$$\sum_{i}^{n_i} n_i combn, 2 - p$$

for $n_i > 1$ where n_i are intergers

Proof by induction

For balanced clusters, assume the data given by Hardin and Hilbe page 66. The data has 2 subjects with 4 time periods.

To calculate the denominator then it can be shown that

0.5*4(4-1)+0.5*4(4-1)=12-p

Similarly

4combination,2 =6 *2=12-p

Hence the proof.

We gain more efficiency in run time when we use our proposed approach than when we use Hardin and Hilbe approach as seen in **Table 3**

##R code for the proof using Hardin and Hilbe Approach

library(dplyr)

fun_M1<- function() {

id<-c(1,1,1,1,2,2,2,2)

t<-c(1,2,3,4,1,2,3,4)

y<-c(4,5,6,7,5,6,7,8)

x<-c(0,1,0,1,0,1,0,1)

fd<-data.frame(id,t,y,x)

m1<-glm(y~x,x=TRUE, data = fd)

y.res=m1$residuals

n=length(fd$id)

new.phi2<-sum(y.res^2)/(n-r)

res_total_n_for_alph<-fd %>%

  group_by(id) %>%

  summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>% mutate(sampl_n = 0.5*n*(n-1))

sum_res_alpha<-sum(res_total_n_for_alph$resd)

sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)

alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha

alpha2 }

start_time <- Sys.time()

fun_M1()

end_time <- Sys.time()

end_time - start_time

##R code for the proof using our proposed method 1

library(dplyr)

```
fun_M1<- function() {

id<-c(1,1,1,1,2,2,2,2)

t<-c(1,2,3,4,1,2,3,4)

y<-c(4,5,6,7,5,6,7,8)

x<-c(0,1,0,1,0,1,0,1)

fd<-data.frame(id,t,y,x)

m1<-glm(y~x,x=TRUE, data = fd)

y.res=m1$residuals

n=length(fd$id)

new.phi2<-sum(y.res^2)/(n-r)

res_total_n_for_alph<-fd %>%

  group_by(id) %>%

  summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>%  mutate(sampl_n = choose(n, 2))

sum_res_alpha<-sum(res_total_n_for_alph$resd)

sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)

alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha

alpha2 }

start_time <- Sys.time()

fun_M1()

end_time <- Sys.time()

end_time - start_time
```

R code for the proof using our proposed method 2

```
library(dplyr)

fun_M1<- function() {

id<-c(1,1,1,1,2,2,2,2)

t<-c(1,2,3,4,1,2,3,4)

y<-c(4,5,6,7,5,6,7,8)

x<-c(0,1,0,1,0,1,0,1)

fd<-data.frame(id,t,y,x)

m1<-glm(y~x,x=TRUE, data = fd)

y.res=m1$residuals

n=length(fd$id)

new.phi2<-sum(y.res^2)/(n-r)
```

```
res_total_n_for_alph<-fd %>%

  group_by(id) %>%

  summarise(resd = sum(combn(y.res, 2, FUN = prod)),n = n())%>%  mutate(sampl_n = n*(n-1)/2)

sum_res_alpha<-sum(res_total_n_for_alph$resd)

sum_n_alpha<-sum(res_total_n_for_alph$sampl_n)

alpha2<-(1/new.phi2)*(1/(sum_n_alpha-r))*sum_res_alpha

alpha2 }

start_time <- Sys.time()

fun_M1()

end_time <- Sys.time()

end_time - start_time
```

References

[1]   GOK, "2013 kenya household health expenditure and utilisation survey," Ministry of Health, Tech. Rep., 2014.

[2]   C. Kelly, C. Hulme, T. Farragher, and G. Clarke, "engAre differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? a systematic review," *engBMJ open*, vol. 6, no. 27884848, pp. e013059–e013059, Nov. 2016. [Online]. Available: https://-www.ncbi.nlm.nih.gov/pmc/articles/PMC5178808/

[3]   M. Karra, G. Fink, and D. Canning, "Facility distance and child mortality: a multi-country study of health facility access, service utilization, and child health outcomes," *Int J Epidemiol*, vol. 46, no. 3, pp. 817–826, 2017. [Online]. Available: https://doi.org/10.1093/ije/dyw062

[4]   S. Gabrysch, S. Cousens, J. Cox, and O. M. R. Campbell, "The influence of distance and level of care on delivery place in rural zambia: A study of linked national data in a geographic information system," *PLOS Medicine*, vol. 8, no. 1, p. e1000394, Jan. 2011. [Online]. Available: https://doi.org/10.1371/journal.pmed.1000394

[5]   D. Kadobera, B. Sartorius, H. Masanja, A. Mathew, and P. Waiswa, "The effect of distance to formal health facility on childhood mortality in rural tanzania, 2005-2007." *Global health action*, vol. 5, pp. 1–9, Nov. 2012.

[6]   A. Schoeps, S. Gabrysch, L. Niamba, A. Sié, and H. Becher, "The effect of distance to health-care facilities on childhood mortality in rural burkina faso," *Am J Epidemiol*, vol. 173, no. 5, pp. 492–498, Jan. 2011. [Online]. Available: https://doi.org/10.1093/aje/kwq386

[7]   R. K. Biswas and E. Kabir, "engInfluence of distance between residence and health facilities on non-communicable diseases: An assessment over hypertension and diabetes in bangladesh," *engPloS one*, vol. 12, no. 28545074, pp. e0177027–e0177027, May 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/-PMC5436657/

[8]   V. Escamilla, L. Calhoun, J. Winston, and I. S. Speizer, "The role of distance and quality on facility selection for maternal and child health services in urban kenya," *Journal of Urban Health*, vol. 95, no. 1, pp. 1–12, 2018. [Online]. Available: https://doi.org/10.1007/s11524-017-0212-8

[9]   R. Stock, "Distance and the utilization of health facilities in rural nigeria," *Social Science & Medicine*, vol. 17, no. 9, pp. 563 – 570, 1983. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-0277953683902988

[10]   T. T. Awoyemi, O. A. Obayelu, and H. I. Opaluwa, "Effect of distance on utilization of health care services in rural kogi state, nigeria," *Journal of Human Ecology*, vol. 35, no. 1, pp. 1–9, Jul. 2011. [Online]. Available: https://doi.org/10.1080/09709274.2011.11906385

[11]   N. Nic Carthaigh, B. De Gryse, A. S. Esmati, B. Nizar, C. Van Overloop, R. Fricke, J. Bseiso, C. Baker, T. Decroo, and M. Philips, "Patients struggle to access effective health care due to ongoing violence, distance, costs and health service performance in afghanistan," *Int Health*, vol. 7, no. 3, pp. 169–175, Dec. 2014. [Online]. Available: https://doi.org/10.1093/inthealth/ihu086

[12]   N. Mwenda, R. Nduati, M. Kosgey, and G. Kerich, "Time effects of bacterial vaginosis on infant morbidities in kenya assessed using modified skewed generalized estimating equations," jun 2020.

[13]   T. Swan, "Generalized estimating equations when theresponse variable has a tweedie distribution:an application for multi-site rainfall modelling," phdthesis, University of Southern Queensland, 2006.

[14]   C. F. Kurz, "Tweedie distributions for fitting semicontinuous health care utilization cost data," *BMC Medical Research Methodology*, vol. 17, no. 1, p. 171, 2017. [Online]. Available: https://doi.org/10.1186/s12874-017-0445-y

[15]   S. Manikandan, "engData transformation." *engJournal of pharmacology & pharmacotherapeutics*, vol. 1, pp. 126–7, Jul 2010.

[16]   S. Su, N. C. Dzupire, P. Ngare, and L. Odongo, "A poisson-gamma model for zero inflated rainfall data," *Journal of Probability and Statistics*, vol. 2018, p. 1012647, 2018. [Online]. Available: https://doi.org/10.1155/-2018/1012647

[17]   B. Swallow, S. T. Buckland, R. King, and M. P. Toms, "engBayesian hierarchical modelling of continuous non-negative longitudinal data with a spike at zero: An application to a study of birds visiting gardens in winter," *engBiometrical journal. Biometrische Zeitschrift*, vol. 58, no. 25737026, pp. 357–371, Mar. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4964939/

[18]   D. Filmer and L. H. Pritchett, "engEstimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of india." *engDemography*, vol. 38, pp. 115–32, Feb 2001.

[19]   H. L. Rippin, J. Hutchinson, D. C. Greenwood, J. Jewell, J. J. Breda, A. Martin, D. M. Rippin, K. Schindler, P. Rust, S. Fagt, J. Matthiessen, E. Nurk, K. Nelis, M. Kukk, H. Tapanainen, L. Valsta, T. Heuer, E. Sarkadi-Nagy, M. Bakacs, S. Tazhibayev, T. Sharmanov, I. Spiroski, M. Beukers, C. van Rossum, M. Ocke, A. K. Lindroos, E.

Warensjo Lemming, and J. E. Cade, "Inequalities in education and national income are associated with poorer diet: Pooled analysis of individual participant data across 12 european countries," *PLOS ONE*, vol. 15, no. 5, p. e0232447, May 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0232447

[20]   OECD, "Employment rate by age group(indicator)," Sep. 2020.

[21]   E. Mwaliko, R. Downing, W. O Meara, D. Chelagat, A. Obala, T. Downing, C. Simiyu, D. Odhiambo, P. Ayuo, D. Menya, and B. Khwa-Otsyula, ""not too far to walk": the influence of distance on place of delivery in a western kenya health demographic surveillance system," *BMC Health Services Research*, vol. 14, no. 1, p. 212, 2014. [Online]. Available: https://doi.org/10.1186/1472-6963-14-212

[22]   J. W. Hardin, *EnglishGeneralized estimating equations. Hardin, Hardin,*. Boca Raton, Fla.: CRC Press, 2013.

[23]   W. Pan, "Akaike's information criterion in generalized estimating equations," *Biometrics*, vol. 57, no. 1, pp. 120–125, Mar. 2001. [Online]. Available: https://doi.org/10.1111/j.0006-341X.2001.00120.x

[24]   H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe, and G. Kitagawa, Eds. New York, NY: Springer New York, 1998, pp. 199–213. [Online]. Available: https://doi.org/10.1007/978-1-4612-1694-0_15

[25]   R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[26]   Z. Zhang, "engVariable selection with stepwise and best subset approaches." *engAnnals of translational medicine*, vol. 4, p. 136, Apr 2016.

[27]   S. Thaddeus and D. Maine, "engToo far to walk: maternal mortality in context." *engSocial science & medicine (1982)*, vol. 38, pp. 1091–110, Apr 1994.

[28]   A. M. Noor, A. A. Amin, P. W. Gething, P. M. Atkinson, S. I. Hay, and R. W. Snow, "engModelling distances travelled to government health services in kenya." *engTropical medicine & international health : TM & IH*, vol. 11, pp. 188–96, Feb 2006.

[29]   R. C. Nesbitt, T. J. Lohela, S. Soremekun, L. Vesel, A. Manu, E. Okyere, C. Grundy, S. Amenga-Etego, S. Owusu-Agyei, B. R. Kirkwood, and S. Gabrysch, "engThe influence of distance and quality of care on place of delivery in rural ghana." *engScientific reports*, vol. 6, p. 30291, Aug 2016.

[30]   S. Allin, C. Masseria, and E. Mossialos, "engMeasuring socioeconomic differences in use of health care services by wealth versus by income," *engAmerican journal of public health*, vol. 99, no. 19150899, pp. 1849–1855, Oct. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741508/