Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2024;10(3): 342-359

Original Article

Bayesian Approach in Modeling Prostate Cancer

Job Lusweti Sirengo*, Drinold Aluda Mbete

Department of Mathematics, School of Natural Science, Masinde Muliro University of Science and Technology, Kakamega, Kenya.

ARTICLE INFO

Received 20.06.2024
Revised 19.07.2024
Accepted 25.08.2024
Published 15.12.2024

Key words:
Risk factors;
Random effects;
Ordinal Logistic;
Bayesian analysis

ABSTRACT

Introduction: Prostate cancer is an emerging health problem in Sub-Saharan Africa. It is often diagnosed at an advanced stage due to lack of access to screening and diagnostic facilities.
Methods: This study therefore aimed to model the effects of risk factors on the outcome of prostate cancer screening using Generalized Bayesian ordinal logistic regression with random effects then compare the results obtained with the model without random effects. The study further used Mean Squared Errors to establish if the estimates for the two models were different.
Results: The findings in this study indicate that aged individuals have high chances of having prostate cancer at the early, late or advanced stage. The individual with traces of family history and hereditary breast & ovarian cancer syndrome are also most likely to be in late or advanced stage of prostate cancer.
Conclusion: From the findings aged individuals, having traces of family history and individuals with hereditary breast & ovarian cancer history, should make sure they understand all symptoms of prostate cancer so that incase of any signs they immediately seek for screening services. In addition, the Ministry of Health should create awareness training and increase screening facilities, this will also encourage for early screening and detection of prostate cancer. The models with presence of random effects were considered best since they had lowest Widely Applicable Information Criterion values in each category.

Introduction

The World Health Organization (WHO) report of 2018,[1] indicates that cancer was responsible for an estimated global death of 9.6 million, with 70% of deaths occurring in low and middle-income countries. Further, about 20% of all cancers across the world relate to chronic infections. In Sub-Saharan Africa, cancer is an emerging health problem and it was estimated that, up to 15% of these diseases had viral etiology where high incidences were in developing countries than the rest of the world.[2] Cancer was ranked as the third cause of mortality after infectious and cardiovascular diseases in Kenya.[3, 4] Further, prostate cancer was ranked as the most common cancer in males at 17.3% with majority of men presenting for

*.Corresponding Author: sirengojob@gmail.com

treatment in advanced stages of the disease.[5]

Prostate cancer is a significant health problem in Sub-Saharan Africa and it is often diagnosed at an advanced stage due to the lack of access to screening and diagnostic facilities, resulting in a poor prognosis.[6] The exact causes of prostate cancer are not known, but several risk factors have been identified which include; age, family history, ethnicity, diet, lifestyle, and exposure to environmental toxins. These risk factors are not well understood, but studies suggest that genetics and lifestyle factors play a major role.[7] The strongest risk factors for prostate cancer are age and African American race/ethnicity. Family history is also an important risk factor for prostate cancer, although only a small proportion of cases will be due to high-penetrance genes.[8]

This study considered the risk factors on each level of prostate cancer outcome. The risk factors assessed include; age, traces of family history, weight control and breast/ovarian syndrome. The random effects helped the study identify the unmeasured risk factors. Random effects models are always considered more generalizable than fixed effects models.[9] By allowing for entity-specific effects to vary across different groups, random effects models capture a wider range of variation in the population. This can enhance the external validity of the results, making them more applicable to broader contexts.

Literature Review

The screening recommendations for prostate cancer, problems encountered in detection, patients' interest and patients' knowledge or informed decision making were important to our study. Partin et al[10] assessed the effect of video and pamphlet interventions on patient prostate cancer knowledge of screening, decision-making in participation, preferences, and behaviours. They used randomized controlled trial of four Mid-Western Veterans Affairs medical facilities. Randomization of patients was done on mailed pamphlet, mailed video, or usual control. The assessment by phone two weeks survey post intervention were; correct responses to questions on prostate cancer natural history, treatment efficacy, the *(PSA)' s* predictive value, and expert disagreement about the Prostate Specific Antigen, whether screening was discussed with provider, screening preferences, and Prostate Specific Antigen testing rates. It was found out that the mean knowledge index scores were higher for video than pamphlet. Further, Bowen et al[11] determined the extent of informed decision making for prostate cancer screening in a defined population. The results pointed out to the need for increasing informed decision making about prostate cancer screening which was of great importance to our study. When there was informed decision making, there would be a majority of people turning up for screening services thus leading to early control measures.

Brant et al[12] carried out research on screening for prostate cancer by using random effects models. They used a mixed linear model in prediction of prostate cancer using Prostate Specific Antigen data. It was found that, 86.8 percent or 88.3 percent were classified correctly by using the longitudinally collected Prostate Specific Antigen measurement, depending on whether or not a distinction was made between local and metastatic cancer. This study intended to estimate the significant risk factors so as to determine the major risk factors

that men considered before seeking for prostate cancer screening services. Some of these risk factors were known not to fully measure their effects on the screening outcomes for prostate cancer, thus, a generalized Bayesian ordinal logistic regression model with and without random effects was used in the analysis. Model parameter estimates for each screening outcome were given. Equally, the model with the lowest widely applicable information criterion (WAIC) was considered as the best model.

Ugwu & Zewotir,[13] used mixed effects logistic regression models for complex survey data on malaria rapid diagnostic test results. The effect of malaria in Nigeria was still worrisome and had remained a leading public health issue in the country. In 2016, Nigeria was the highest malaria burden country among the 15 countries in sub-Saharan Africa that accounted for 80% global malaria cases. The purpose of this study was to utilize appropriate statistical models in identifying socio-economic, demographic and geographic risk factors that have influenced malaria transmission in Nigeria, based on malaria rapid diagnostic test survey results. Their study contributed towards re-designing intervention strategies to achieve the target of meeting the Sustainable Development Goals 2030 Agenda for total malaria elimination. The study adopted the generalized linear mixed models' approach which accounts for the complexity of the sample survey design associated with the data. The 2015 Nigeria malaria indicator survey data of children between 6 and 59 months are used in the study. From the findings of this study, the cluster effect is significant ($P < 0.0001$) which has suggested evidence of heterogeneity among the clusters. It was also found that the vulnerability of a child to malaria infection increases as the child advances in age. Other major significant factors were; the presence of anaemia in a child, an area where a child resides (urban or rural), the level of the mother's education, poverty level, number of household members, sanitation, age of head of household, availability of electricity and the type of material for roofing. Moreover, children from Northern and South-West regions were also found to be at higher risk of malaria disease and re-infection. Improvement of socio-economic development and quality of life was concluded to be paramount to achieving malaria free Nigeria. There is a strong link of malaria risk with poverty, under-development and the mother's educational level.

Ali et al[14] did research on the Sufficient Sample Size and Power in Multilevel Ordinal Logistic Regression Models. For most of the time, biomedical researchers have been dealing with ordinal outcome variable in multilevel models where patients are nested in doctors.We can justifiably apply multilevel cumulative logit model, where the outcome variable represents the mild, severe, and extremely severe intensity of diseases like malaria and typhoid in the form of ordered categories. Based on our simulation conditions, Maximum Likelihood (ML) method is better than Penalized Quasilikelihood (PQL) method in three-category ordinal outcome variable. PQL method, however, performs equally well as ML method where five-category ordinal outcome variable is used. Further, to achieve power more than 0.80, at least 50 groups are required for both ML and PQL methods of estimation. It may be pointed out that, for five-category ordinal response variable model, the power of PQL method is slightly higher than the power of ML method.

Rezapour et al,[15] applied Bayesian ordinal logistic model for identification of factors to

traffic barrier crashes: considering roadway classification. One of the main objectives of policy makers is to reduce crash severity due to high social impacts and economic loss associated with severe crashes. It was indicated that, one of the most efficient ways to achieve this objective is through identification of the contributory factors to severe crashes. Highway traffic barriers have been installed with the objective of protecting motorists who have drifted off the roadway. Although these traffic barriers save many lives, the crash severity for these crashes were disproportionally high. Only traffic barriers crashes were considered in this study to identify the factors for these types of crashes. Moreover, due to the importance of low volume crashes, especially in rural areas like Wyoming, this study investigated the effects of road classification on crash severity as well as how these effects impact the role of the contributory factors. Low volume roads often receive less attention in terms of road safety due to their low crash frequencies. A Bayesian approach was used to fit the models since this approach does not require large sample assumptions, it does not rely on approximations for estimating non linear functions of the parameters, and also it provides simpler interpretations for model unknowns. The factors identified by this study included the main effects of day of the week, seasonality, improper restraints as well as the interaction effects of low volume roads with shoulder width, road surface conditions, and lighting conditions. These interaction terms indicated that the effects of these contributory factors change with the traffic volume. Possible causes of the significant main and interaction terms are discussed in the manuscript.

Li et al,[16] did research on Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. They aimed at comparing different statistical software implementations of these models. The methods used were based on the individual patient data from 8509 patients in 231 centers with moderate and severe Traumatic Brain Injury (TBI) enrolled in eight Randomized Controlled Trials (RCTs) and three observational studies. They fitted logistic random effects regression models with the 5-point Glasgow Outcome Scale (GOS) as outcome, both dichotomized as well as ordinal, with center and/or trial as random effects, and as covariates age, motor score, pupil reactivity or trial. They then compared the implementations of frequentist and Bayesian methods to estimate the fixed and random effects. Frequentist approaches included R (lme4), Stata (GLLAMM), SAS (GLIMMIX and NLMIXED), MLwiN ([R] IGLS) and MIXOR, Bayesian approaches included WinBUGS, MLwiN (MCMC), R package MCMCglmm and SAS experimental procedure MCMC. Three data sets (the full data set and two sub-datasets) were analysed using basically two logistic random effects models with either one random effect for the center or two random effects for center and trial. For the ordinal outcome in the full data set also a proportional odds model with a random center effect was fitted. The results showed that the packages gave similar parameter estimates for both the fixed and random effects and when based on relatively sparse data set, i.e. when the numbers of level-1 (patient level) and level-2 (hospital level) data units were also about the same. However, the frequentist and Bayesian approaches showed somewhat different results. The software implementations differ considerably in flexibility, computation

time, and usability. There were also differences in the availability of additional tools for model evaluation, such as diagnostic plots. The experimental SAS (version 9.2) procedure MCMC appeared to be inefficient. They concluded that, on relatively large data sets, the different software implementations of logistic random effects regression models produce similar results. Thus, for a large data set there seems to be no explicit preference (of course if there is no preference from a philosophical point of view) for either a frequentist or Bayesian approach (if based on vague priors). The choice for a particular implementation may largely depend on the desired flexibility, and the usability of the package. For small data sets, the random effects variances are difficult to estimate. In the frequentist approaches the MLE of this variance was often estimated zero with a standard error that is either zero or could not be determined, while for Bayesian methods the estimates could depend on the chosen "noninformative" prior of the variance parameter. The starting value for the variance parameter may be also critical for the convergence of the Markov chain.

In this study, the effects of risk factors on prostate cancer screening outcome are modelled using Generalized Bayesian ordinal logistic regression with random effects. The results are then compared to the results obtained for the model without random effects by use of Mean Squared Error (MSE) and Widely Applicable Information Criterion (WAIC).

Materials and Methods

The research design was a retrospective study of patients who presented with prostate cancer from January 2020 – December 2022. The study used secondary data from Kenyatta National Hospital (KNH). Records of patients who were managed for prostate cancer were retrieved from the department and the oncology unit of the KNH. To assess the stage of cancer, imaging reports of CT scans or MRIs that were undertaken during staging assessment were used. To assess histological grade, the histological reports of prostate biopsy or prostatectomy specimens was used. Patients' records were interrogated to capture the relevant data for the study which was entered in standard questionnaire for eventual transfer to excel computer data sheet. The hospital records show that between the years 2020 – 2022 it has screened 704 people for prostate cancer. To determine the sample size to use in the study, the researcher used the sample size formula for population proportion proposed by Cochran,[17] to yield a representative sample for proportions.

$$n = Z^2 \frac{p(1-p)}{e^2}$$

Where n is the sample size, Z^2 is the standard error associated with chosen significance level (Z=1.96), e is the desired level of precision (margin of error 0.05), p is the estimated proportion patients. The expected p is 16% patients presenting with advanced cancer.[18] Therefore, the sample size used in the study is 207 as shown in the formula.

$$n = 1.96^2 \frac{0.16(1-0.16)}{0.05^2} = 207$$

Demographic Characteristics and descriptive statistics

The results in Table 1 shows that 88.7% of the patients screened for prostate cancer were

above 50 years of age while 11.3% were below 50 years. On occupation, the results shows that 75% of the patients screened for prostate cancer were employed compared to 25% who were unemployed. The results further shows that 35.3% of the patients screened for prostate cancer were in stage II and majority of these patients 25.5% were from Central region.

The data is mostly used by researchers in construction of statistical models such as multilevel models, hierarchical models or mixed effects models.[19] Since in this study the outcome of screening for prostate cancer are ordinal, the effects of the risk factors are based on the coefficient parameters θ_{kj} where *k=1,2,...K* the number of risk factors and *j=1,2,...J-1* is prostate cancer screening outcome.

Therefore, there exists *K×(J-1)* matrix of the coefficient parameters where *J-1=3*.

The generalized Bayesian ordinal regression model consisted a series of binary regression models for the four outcomes of prostate cancer. The unmeasured risk factors considered by individuals who turned out for prostate cancer screening were measured by a random effect. Let γ_j be a random intercept which pertained to the k^{th} risk factor of a random covariate X_{ik}. Thus the series of latent variables is defined by the following equation,

$$Z_{ij}=X_{ik}\,\theta_{kj}+\gamma_j+\epsilon_{ij};\ \epsilon_{ij}\sim N(0,\sigma^2) \text{ and } \gamma_j\sim N(\mu_{\gamma_j},\sigma_{\gamma^2}) \quad (1)$$

Here, γ_j denotes the random effect and θ_j;

Table 1. Demographic characteristics

		Frequency	Percent
Age	Above 50 years	181	88.7
	Below 50 years	23	11.3
	Total	204	100.0
Occupation	Employed	153	75.0
	Not employed	51	25.0
Duration of symptoms	1 week	15	7.4
	1 Month	48	23.5
	3 months	96	47.1
	More than 3 months	45	22.1
Stage of disease	Stage 1	33	16.2
	Stage II	72	35.3
	Stage III	68	33.3
	Advanced Stage	31	15.2
Region	Nairobi Region	44	21.6
	Western Region	30	14.7
	Central Region	52	25.5
	Coastal Region	34	16.7
	Eastern Region	27	13.2
	Rift Valley region	17	8.3

j=1,2,3 represented the set of coefficients for the model. From Equation 1, the parameters are grouped as; Z_j - the series of latent variables and (θ_j, μ_{θ_j}, σ_θ^2, α_j^*, γ_j, μ_{γ_j}, σ_γ^2) the set of model parameters. It was our purpose to iterate between each of the parameters conditional on the remaining parameters of (θ_j, μ_{θ_j}, σ_θ^2, α_j^*, γ_j, μ_{γ_j}, σ_γ^2 and Z_j) until they converge to generate the parameters for the model with random effects. These parameters were estimated and compared with the parameter estimates for the model without presence of random effects of the study done by[20] using mean squared error (MSE).

In Equation 1, it is clearly seen that the latent propensity of a Bayesian ordinal regression model follows a series of logistic distribution with conditional mean matrix $X_{ik}\ \theta_j+\gamma_j$. Therefore, there is a series of latent continuous random variables $Z_{ij}\sim N(X_{ik}\ \theta_j+\gamma_j, I)$ and thus, the variable Y_i is observed such that, $Y_i=j$ if $\alpha_{j-1}^* < Z_{ij} \le \alpha_j^*$ with $\alpha_1^*=-\infty$ and $\alpha_j^*=\infty$. Specifically, a latent propensity variable Z_{ij} is used as a basis for modeling the ordered ranking of prostate cancer screening outcome for the i^{th} individual.

The Conditional Distribution of θ_j

To estimate the coefficient parameters θ_j, the following priors are added;

$$\begin{aligned} \theta_j \mid \mu_{\theta_j}, \sigma_\theta^2 &\sim N\left(\mu_{\theta_j}, \sigma_\theta^2\right) \\ \pi\left(\mu_{\theta_j}\right) &\propto 1 \\ \sigma_\gamma^2 &\sim IG(\mathbf{u}, \mathbf{v}) \end{aligned} \tag{2}$$

The likelihood and the latent variables when the parameter μ_γ is added to the model is given by joint posterior dis tribution.[21]

$$\pi(\theta_j, \mu_{\theta j}, \sigma_\theta^2, \gamma_j, \mu_{\gamma j}, \sigma_\gamma^2, \alpha_j^*, Z_j | Y_i) \propto \prod_{i=1}^{N} \left[\left(\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(Z_{ij} - X_{ik}\theta_j - \gamma_j)^2\right) \right) \times \sum_{j=1}^{J} \left\{ I(Y_i = j) I(\alpha_{j-1}^* < Z_{ij} < \alpha_j) \right\} \right] \tag{3}$$

Therefore, the full conditional distribution of θ_j is given by,

$$\theta_j \mid \mu_{\theta_j}, \sigma_\theta^2, \gamma_j, \mu_{\gamma_j}, \sigma_\gamma^2, \alpha_j^*, Z_j, Y_i \sim N(\tilde{\theta}_j\ \Sigma_{\theta_j}) \tag{4}$$

To prove this equation the priors in Equation (1) and Equation (2) are used. Therefore,

The conditional distribution of γ_j

The other introduced parameters which represent the random effects are also estimated. Where the full conditional distribution of γ_j is given by,

$$\gamma_j \mid \theta_j, \mu_{\theta_j}, \sigma_\theta^2, \mu_{\gamma_j}, \sigma_\gamma^2, \alpha_j^*, Z_j, Y_i \sim N(\tilde{\gamma}_j, \Sigma_{\gamma_j})$$

$$\pi(\theta_j|\mu_{\theta_j},\sigma_\theta^2,\gamma_j,\mu_{\gamma_j},\sigma_\gamma^2,\alpha_j^*,Z_j,Y_i) \propto exp\left[-\frac{1}{2}\left\{(Z_j-X_k\theta_j-\gamma_j)(Z_j-X_k\theta_j-\gamma_j)+\frac{1}{\sigma_\theta^2}(\theta_j-\mu_{\theta_j})(\theta_j-\mu_{\theta_j})\right\}\right] \quad (5)$$

Let, $w = \left[(Z_j-\gamma_j)-X_k\theta_j\right]\left[(Z_j-\gamma_j)-X_k\theta_j\right]+\frac{1}{\sigma_\theta^2}(\theta_j-\mu_{\theta_j})(\theta_j-\mu_{\theta_j})$ then,

$$\begin{aligned} w &= (Z_j-\gamma_j)^T(Z_j-\gamma_j)-2\beta_j^{*T}X_k^T(Z_j^*-\gamma_j)+\theta_j^TX_k^TX_k\theta_j \\ &\quad +\frac{1}{\sigma_\theta^2}\theta_j^T\theta_j-2\theta_j^T\frac{\mu_{\theta_j}}{\sigma_\theta^2}+\frac{\mu_{\theta_j}^2}{\sigma_\theta^2} \\ w &= \theta_j^T\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)\theta_j-2\theta_j^T\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right) \\ &\quad +\left[\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\right. \\ &\quad \left.\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\right]-\left[\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)\right. \\ &\quad \left.\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\right]+(Z_j-\gamma_j)^T(Z_j-\gamma_j)+\eta_j^* \\ w &= (\theta_j-\tilde{\theta}_j)^T\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)(\theta_j-\tilde{\theta}_j)-\left[\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\right. \\ &\quad \left.\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right)\right]+(Z_j-\gamma_j)^T(Z_j-\gamma_j)+\eta_j^* \end{aligned} \quad (6)$$

Where

$$\Sigma_{\theta_j}=\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1} \text{ and } \tilde{\theta}_j=\left(\frac{I}{\sigma_\theta^2}+X_k^TX_k\right)^{-1}\left(X_k^T(Z_j-\gamma_j)+\omega_j^*\right) \quad (7)$$

(8)

To prove Equation (8), the following equation is used,

$$\pi\left(\gamma_j\middle|\theta_j,\mu_{\theta_j},\sigma_\theta^2,\mu_{\gamma_j},\sigma_\gamma^2,\alpha_j^*,Z_j,Y_i\right) \propto \exp\left[-\frac{1}{2}\left\{(Z_j-X_k\theta_j-\gamma_k)(Z_j-X_k\theta_j-\gamma_j)+\frac{1}{\sigma_\gamma^2}(\gamma_j-\mu_{\gamma_j})(\gamma_j-\mu_{\gamma_j})\right\}\right] \quad (9)$$

Also, let $w_2 = \left[(Z_j - X_k\theta_j) - \gamma_j\right]\left[(Z_j - X_k\theta_j) - \gamma_j\right] + \frac{1}{\sigma_\gamma^2}(\gamma_j - \mu_{\gamma_j})(\gamma_j - \mu_{\gamma_j})$, then,

$$\begin{aligned} w_2 &= (Z_j - X_k\theta_j)^T(Z_j - X_k\theta_j) - 2\gamma_j^T(Z_j - X_k\theta_j) + \gamma_j^T\gamma_j \\ &\quad + \frac{1}{\sigma_\gamma^2}\left\{\gamma_j^T\gamma_j - 2\gamma_j^T\mu_{\gamma_j} + \mu_{\gamma_j}^2\right\} \\ w_2 &= \gamma_j^T\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}\gamma_j - 2\gamma_j^T\left((Z_j - X_k\theta_j) + \nu_j\right) \\ &\quad + \left[\left((Z_j - X_k\theta_j) + \nu_j\right)\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\right. \\ &\quad \left.\left((Z_j - X_k\theta_j) + \nu_j\right)\right] - \left[\left((Z_j - X_k\theta_j) + \nu_j\right)\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}\right. \\ &\quad \left.\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\left((Z_j - X_k\theta_j) + \nu_j\right)\right] + (Z_j - X_k\theta_j)^T(Z_j - X_k\theta_j) + \kappa_j \\ w_2 &= (\gamma_j - \tilde{\gamma}_j)^T\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}(\gamma_j - \tilde{\gamma}_j) - \left[\left((Z_j - X_k\theta_j) + \nu_j\right)\right. \\ &\quad \left.\left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\left((Z_j - X_k\theta_j) + \nu_j\right)\right] + (Z_j - X_k\theta_j)^T(Z_j - X_k\theta_j) + \kappa_j \end{aligned} \quad (10)$$

Where,

$$\Sigma_{\gamma_j} = \left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1} \text{ and } \tilde{\gamma}_j = \left\{\left(1 + \frac{1}{\sigma_\gamma^2}\right)I\right\}^{-1}\left((Z_j - X_k\theta_j) + \nu_j\right) \quad (11)$$

The conditional distribution of $\sigma_{\gamma_j^2}$

With defined prior for $\sigma_{\gamma_j}^2$ to the model with the presence of random effects the fully conditional distribution shown in the following equation is also derived,

$$\sigma_{\gamma_j}^2 \mid \gamma_j, \mu_{\gamma_j}, \theta_j, \mu_{\theta_j}, \sigma_{\theta_j}^2, \alpha_j^*, Z_j \sim IG\left(\mathbf{u}_{\sigma_{\gamma_j}^2},\ \mathbf{v}_{\sigma_{\gamma_j}^2}\right) \quad (12)$$

Since $\sigma_{\gamma_j}^2 \sim IG(\mathbf{u}, \mathbf{v})$ such that IG(u,v) is inverted gamma distribution with parameters u and v then,

$$\pi(\sigma_{\gamma_j}^2 \mid \mathbf{u}, \mathbf{v}) = \frac{\mathbf{v}^{\mathbf{u}}}{\tilde{\Gamma}(\mathbf{u})}(\sigma_{\gamma_j}^2)^{-\mathbf{u}-1}exp\left(-\frac{\mathbf{v}}{\sigma_{\gamma_j}^2}\right) \quad (13)$$

The posterior distribution of $\sigma^2_{\gamma_j}$ is then solved as shown below,

$$\begin{aligned}\sigma^2_{\gamma_j}|\gamma_j,\mu_{\gamma_j},\theta_j,\mu_{\theta_j},\sigma^2_{\theta_j},\alpha^*_j,Z_j \quad &\propto \quad \pi(\gamma_j|\mu_{\gamma_j},\sigma^2_{\gamma_j})\times\pi(\sigma^2_{\gamma_j}) \\ &\propto \quad \prod_{j=1}^{n_j}\left\{\frac{1}{\sqrt{2\pi\sigma^2_{\gamma_j}}}\exp\left[-\frac{1}{2\sigma^2_{\gamma_j}}\left(\gamma_j-\mu_{\gamma_j}\right)^2\right]\right\} \\ &\quad \times\frac{\mathbf{v^u}}{\Gamma(\mathbf{u})}(\sigma^2_{\gamma_j})^{-\mathbf{u}-1}exp\left(-\frac{\mathbf{v}}{\sigma^2_{\gamma_j}}\right) \qquad (14)\end{aligned}$$

$$\sigma^2_{\gamma_j}|\gamma_j,\mu_{\gamma_j},\theta_j,\mu_{\theta_j},\sigma^2_{\theta_j},\alpha^*_j,Z_j \propto \left(\frac{1}{\sigma^2_{\gamma_j}}\right)^{\frac{J}{2}+\mathbf{u}+1}exp\left[-\frac{1}{\sigma^2_{\gamma_j}}\left\{\mathbf{v}+\frac{\sum_{j=1}^{J}(\gamma_j-\mu_{\gamma_j})^2}{2}\right\}\right]$$

Which was a kernel of inverse gamma distribution with,

$$\mathbf{u}_{\sigma^2_{\gamma_j}} = \frac{J}{2}+\mathbf{u} \text{ and } \mathbf{v}_{\sigma^2_{\gamma_j}} = \mathbf{v}+\frac{\sum_{j=1}^{J}(\gamma_j-\mu_{\gamma_j})^2}{2} \qquad (15)$$

Hence,

$$\sigma^2_{\gamma_j}|\gamma_j,\mu_{\gamma_j},\theta_j,\mu_{\theta_j},\sigma^2_{\theta_j},\alpha^*_j,Z_j \sim IG\left[\frac{J}{2}+\mathbf{u},\ \mathbf{v}+\frac{\sum_{j=1}^{J}(\gamma_j-\mu_{\gamma_j})^2}{2}\right] \qquad (16)$$

The conditional distributions of the latent variables Z_j is also derived. The latent variables Z_j is given by;

$$Z_j\,|\,\theta_j,\mu_{\theta_j},\sigma^2_{\theta_j},\gamma_j,\mu_{\gamma_j},\sigma^2_{\gamma_j},\alpha^*_j,Y_i=j \quad \sim \quad N\left(X_{ik}\theta_j+\gamma_j,\ I\right)$$

truncated at the left and right by α^*_{j-1} and α^*_j (17)

Assessment Criteria

Mean Squared Error

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the

observed and predicted values. The value of MSE approaches zero when the error in the model reduces. It is one of many ways to quantify the difference between values predicted by an estimator and the true values of the response.

In this research, the MSE of model parameters was used to determine if there was any difference in parameter estimates obtained for the model with presence of random effect and the model obtained by[20]. When the error of the model parameters was high or not equal to zero, the parameter estimates of the model with and without presence of random effects were different and therefore the Widely Applicable Information Criterion (WAIC) was used to determine the best model fit. The following equation is used to compute the MSE,

$$MSE = \frac{\sum_{k=1}^{K}(\beta_k - \theta_k)^2}{K} \quad (18)$$

where;

β_k $k = 1...K$ – are parameter estimates without presence of random effects in the study by[20].

θ_k $k = 1 \ldots K$ – are parameter estimates for the model with presence of random effects.

K – is the number of risk factors.

The Widely Applicable Information Criterion

In statistics, the widely applicable information criterion (WAIC), also known as Watanabe–Akaike information criterion, is the generalized version of the Akaike information criterion (AIC) onto singular statistical models that is more fully Bayesian than the Deviance Information Criterion (DIC).[22]

Like DIC, WAIC estimates the effective number of parameters to adjust for overfitting. The Widely Applicable Information Criterion is given by,

$$WAIC = -2(lppd\text{-}pWAIC) \quad (19)$$

In the Model, lppd is the log pointwise predictive density and pWAIC is the effective number of parameters.[23]

Results

The random effects in our model helped in measuring the unmeasured risk factors affecting individuals' outcome of prostate cancer. When there is more than one random effect per grouping factor, correlations between random effects are estimated. Table 2 shows the coefficient parameters of the risk factors for the model without random effects in the study [20] denoted by E_1. The coefficient parameters E_2 are for the model with random effects.

The estimates E_1 and E_2 are compared using Mean Squared Error. For positive coefficients E_1 and E_2, the higher values on the explanatory variable increase the chance that the respondent will be in higher category of the dependent variable than the current one. The negative coefficients signifies that the higher values on the explanatory variable increase the likelihood of being in the current or lower category.

The changes observed on the parameters of the first category that is 0 Vs 1, 2, 3 were; the intercept increased from 4.71 to 9.59, coefficient of age also changed from E_1=0.83,CI (-0.4,1.3) for the model without random effects to E_2=2.27,CI (-0.9,5.7) for the model with random effects indicating that aged people were more likely to have prostate cancer. The coefficients of family history in both models E_1=2.82,CI (0.6,6.5) and

E_2=2.97,CI (-0.7,7.0) were positive indicating that individuals with history of prostate cancer were likely to be positive. This was also witnessed in individuals with history of breast and ovarian syndrome in their families E_1=0.77,CI (0.3,1.3) and E_2=2.28,CI (0.1,5.1). Weight control E_1=-1.56,CI (-2.3,-0.9) and E_2=-1.22,CI (-3.9,2.2) which were negative coefficients in both models indicating that despite individuals managing their weight, they were likely to have prostate cancer.

In the second category that is, 0, 1 vs 2, 3; there was positive effect of age with E_1=2.42,CI (2.1,2.7) and E_2=2.97,CI (2.6,3.3) to the screening outcome. This means that elderly people had higher chances to be in late or advanced stage of prostate cancer. Traces of family history, weight control and hereditary breast/ovarian cancer had also positive effect on prostate cancer screening outcome as shown in Table 2. It means that individual with traces of family history and hereditary breast & ovarian cancer syndrome were also most likely to be in late or advanced stage of prostate cancer.

Figure 1, is the representations of marginal effects which provide a direct effect and easily interpreted answers for the second category of prostate cancer outcome, that is, 0, 1 Vs 2, 3. All the values move in the same direction indicating a positive effect of risk factors on the screening outcome of prostate cancer. Considering these risk factors, there were higher chances for individuals to be in late or advanced stages of prostate cancer.

The third category that is, 0, 1, 2 vs 4; the coefficient parameters are all positive for the model with presence of random effects. This means that aged individual, those with traces of prostate in their family members and also with history of hereditary breast/ovarian cancer syndrome were likely to be in advanced stage. The weight control had negative effects in the model without random effects with E_1=-0.03,CI (-0.4,0.1). Since there are more than one random effect per grouping factor, correlations between random effects are also estimated as shown in Table 3.

The study by Ugwu & Zewotir ,[13] contributed towards re-designing intervention strategies to achieve the target of meeting the Sustainable Development Goals 2030 Agenda for total malaria elimination. This study identified the major risk factors that are considered to be most likely to cause prostate cancer. The findings

Table 2. Estimate of Coefficients for the Model without and with Random Effects

Variables	0 Vs 1, 2, 3				0,1 Vs 2, 3				0,1, 2 Vs 3			
	E_1	95% CI	E_2	95% CI	E_1	95% CI	E_2	95% CI	E_1	95% CI	E_2	95% CI
Threshold	4.71	3.0, 5.9	9.59	4.5, 19.2	-1.99	-2.7, -1.8	1.60	0.2, 4.0	0.61	0.2, 1.4	1.70	0.4, 4.2
Age (X_1)	0.83	-0.4, 1.3	2.27	-0.9, 5.7	2.42	2.1, 2.7	2.97	2.6, 3.3	0.67	0.1, 1.0	1.58	0.7, 2.5
Traces of family History (X_2)	2.82	0.6, 6.5	2.97	-0.7, 7.0	2.83	2.6, 3.3	2.98	2.6, 3.3	1.08	0.7, 2.2	1.71	0.8, 2.6
Weight Control (X_3)	-1.56	-2.3, -0.9	-1.22	-3.9, 2.2	0.98	0.7, 1.2	2.93	2.6, 3.3	-0.03	-0.4, 0.1	1.65	0.8, 2.5
Breast and Ovarian Syndrome (X_4)	0.77	0.3, 1.3	2.28	0.1, 5.1	2.29	1.9, 2.5	3.06	2.7, 3.4	0.73	0.2, 0.9	1.94	1.2, 2.7

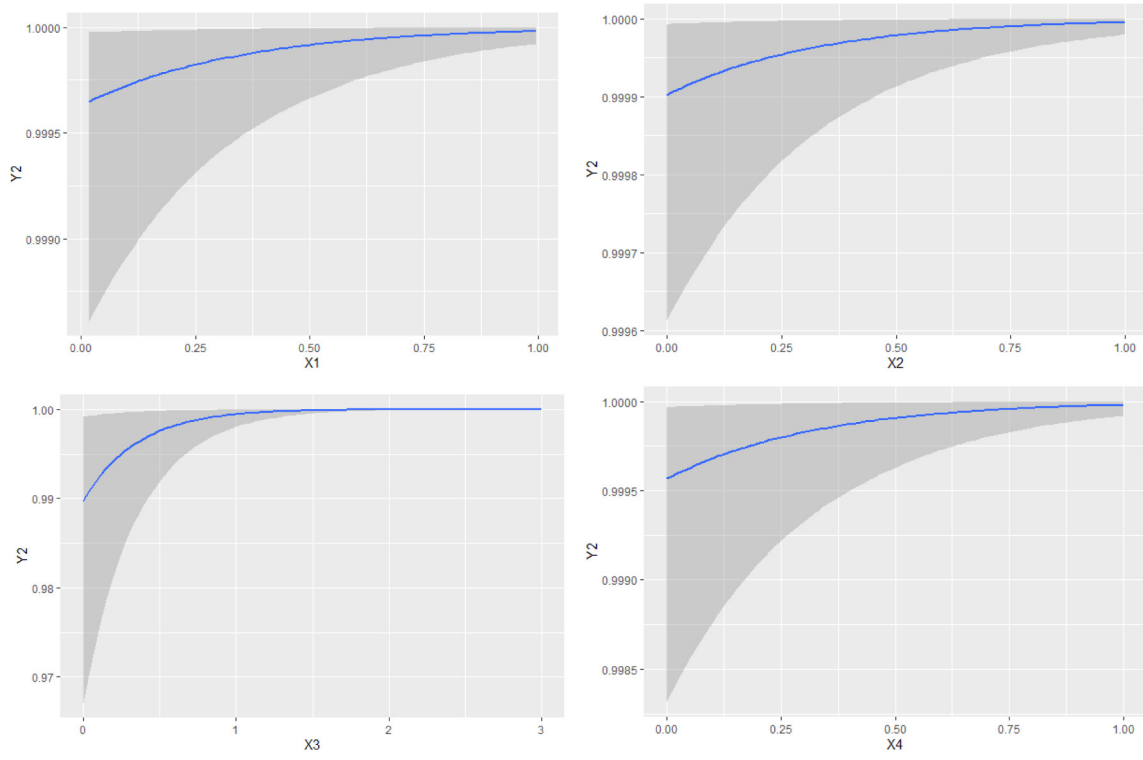

Figure 1. The Marginal Effects for Individuals in Early or Late Stages of Prostate Cancer

Table 3. The Random Effects on the Risk Factors Affecting Prostate Cancer Outcome

Variables	0 Vs 1, 2, 3			0,1 Vs 2, 3			0,1, 2 Vs 3		
	Estimate	l-95% CI	u-95% CI	Estimate	l-95% CI	u-95% CI	Estimate	l-95% CI	u-95% CI
sd;(Threshold)	1.35	0.05	3.96	2.12	0.09	5.72	1.70	0.06	6.12
sd;(X_1)	2.44	0.10	7.17	1.79	0.06	5.45	2.02	0.06	6.32
sd;(X_2)	1.90	0.05	5.94	9.55	2.27	20.64	6.80	0.83	15.17
sd;(X_3)	3.22	0.22	8.31	6.40	4.79	8.70	4.32	2.40	6.55
sd;(X_4)	1.44	0.05	4.40	2.22	0.11	5.91	1.27	0.03	3.95
cor; (Threshold, X_1)	-0.09	-0.83	0.69	0.04	-0.73	0.78	0.05	-0.73	0.78
cor; (Threshold, X_2)	-0.02	-0.78	0.78	0.15	-0.66	0.84	0.05	-0.73	0.78
cor;(X_1, X_2)	-0.04	-0.75	0.71	0.06	-0.72	0.77	0.03	-0.73	0.77
cor; (Threshold, X_3)	-0.03	-0.75	0.73	-0.05	-0.77	0.66	0.13	-0.72	0.87
cor;(X_1, X_3)	-0.04	-0.74	0.70	0.11	-0.70	0.80	0.12	-0.65	0.78
cor;(X_2,X_3)	-0.15	-0.82	0.67	-0.18	-0.87	0.61	-0.24	-0.90	0.60
cor;(Threshold,X_4)	-0.03	-0.75	0.72	-0.25	-0.87	0.62	-0.11	-0.81	0.71
cor;(X_1,X_4)	0.01	-0.75	0.75	-0.19	-0.86	0.69	-0.10	-0.81	0.70
cor;(X_2,X_4)	-0.02	-0.75	0.74	-0.02	-0.76	0.73	-0.07	-0.78	0.71
cor;(X_3,X_4)	-0.05	-0.76	0.72	-0.43	-0.90	0.40	0.08	-0.68	0.78

have shown that individuals who consider the risk factors are likely to be in all stages of prostate cancer. The findings contribute to the existing literature by creating awareness. This also helped the policy makers and ministry of health in guiding and encouraging individuals to seek for early screening.

The study that was conducted by Cerhan, J. R et al[24] on family history and prostate cancer risk in a population-based cohort of Iowa men indicated that 4.6% of the cohort reported a family history of prostate cancer in a brother or father, and this was positively associated with prostate cancer risk after adjustment for age or after multivariate adjustment for age, alcohol, and dietary factors. The risk was greater if a brother had prostate cancer than if a father had prostate cancer. Also at baseline, 9.6% of the cohort had a family history of breast and/or ovarian cancer in a mother or sister, and this was positively associated with prostate cancer risk. Men with a family history of both prostate and breast/ovarian cancer were also at increased risk of prostate cancer. These findings are similar to the findings of this study whereby bot family history and hereditary breast/ovarian had positive effect on prostate cancer outcome. Studies have also shown that prostate cancer incidence and mortality rates are strongly related to the age with the highest incidence being seen in elderly men.[25] Further, older patients are more likely to have high-risk prostate cancer at diagnosis and less likely to receive local therapy. Indeed, underuse of potentially curative local therapy among older men with high-risk disease may in part explain observed differences in cancer-specific survival across age strata.[24] This study found out that aged individuals were likely to be in late stages of cancer which are difficult to guarantee effective treatment.

In the study by Rodriguez, C et al,[27] it was concluded that obesity increases the risk of more aggressive prostate cancer and may decrease either the occurrence or the likelihood of diagnosis of less-aggressive tumors. Men who lose weight may reduce their risk of prostate cancer. In this study, weight control did not consistently or clearly show its effect on the outcome of prostate cancer because the effects varied across all the three categories.

Mean Squared Error of the model parameters

The model parameter estimates for the model with and without random effects were compared using the Mean Squared Error. The parameter estimates without random effects were assumed to be the actual values while the parameter estimates with the presence of random effects were assumed to be the predicted values.

Given a set of coefficient parameters $\beta_1, \beta_2, \ldots, \beta_K$ for the model without random effects and $\theta_1, \theta_2, \ldots, \theta_K$ the coefficients parameters of the model with presence of random effects, the MSE was computed as follows;

$$MSE = \frac{(\beta_1 - \theta_1)^2 + (\beta_2 - \theta_2)^2 + \cdots + (\beta_K - \theta_K)^2}{K} \quad (20)$$

Similarly, the Equation (20) was applied to the coefficient parameters estimated in this research. The coefficient parameters were different if the $MSE \neq 0$.

In our analysis the MSE of the parameters estimate for negative or positive outcome of prostate cancer was 5.66. The value was not equal to zero thus indicating that the two set of

parameters were not similar or equal. Similarly, comparing the coefficient parameter estimate values for the model representing individuals who were in the category 0, 1, Vs 2, 3, the MSE was 3.54. Also, the values for those in the category 0, 1, 2 Vs 3, gave a Mean Square Error equal to 1.34. The values were all not equal to zero. It was therefore clearly seen that, the parameters for the above model fit were all different.

The best model fit

Since the model parameters had MSEs greater than zero, the study went further to determine the best model fit by using Widely Applicable Information Criterion (WAIC) as shown in Table 4.Page Br

To get the best model fit, we compared the model performance with and without the random effects. The model with smaller WAIC was considered to be the best model. From the table, the models with presence of random effects had lowest WAIC values. These models were considered to be the best in each category.

Discussion

The study findings showed that aged individuals were more likely to have prostate cancer. In the second category of the study grouping, age also positively affected the screening outcome showing that elderly people had higher chances to be in late or advanced stage of prostate cancer. In addition, traces of family history, weight control and hereditary breast & ovarian cancer had positive effect on prostate cancer screening outcome. It means that individual with traces of family history and hereditary breast & ovarian cancer syndrome were also most likely to be in late or advanced stage of prostate cancer.

In order to get best model fits for the two models that is, with and without random effects, the three pairs of study models were compared using Widely Applicable Information Criterion (WAIC). The models with presence of random effects had lowest WAIC values hence they were considered to be the best models in each category.

Conclusion

Prostate cancer can be managed if individuals turn up for early screening and diagnosis. The outcome of screening depends on the age, historical background and hereditary breast & ovarian cancer. The study advises all aged individuals who experience symptoms of prostate cancer to continuously seek for screening services and follow the advice of medical practitioners on how to maintain good health to avoid the risk of prostate cancer.

Table 4. Widely Applicable Information Criterion (WAIC) of model fit

Category of prostate cancer Individuals outcome	Random Effects	WAIC
0 Vs 1, 2, 3	NA	333.8
0, 1 Vs 2, 3	NA	833.6
0, 1, 2 Vs 3	NA	751.6
0 Vs 1, 2, 3	Present	179.3
0, 1 Vs 2, 3	Present	227.5
0, 1, 2 Vs 3	Present	327.0

Considering the findings of this study, aged individuals, having traces of family history and individuals with hereditary breast & ovarian cancer history, have higher chances to be with prostate cancer and therefore need to understand all symptoms of prostate cancer. Once any signs or symptoms appear they are supposed seek for screening services at earliest time possible. For this to be effective, the Ministry of Health should create awareness training and increase screening facilities across the country. If all of these initiatives are considered and be implemented, they will encourage for early screening and detection of prostate cancer.

Conflicting interest

The authors declares that there is no conflicting interest

References

1. World Health Organization, et al (2018). Cancer fact sheet. 2018 http://www. who. int/ mediacentre/factsheets/fs297/en. Accessed March, 15.

2. World Health Organization, et al. (2017). Guide to cancer early diagnosis.

3. Macharia, L. W., Mureithi, M. W., and Anzala, O. (2019). Cancer in Kenya: types and infection-attributable. data from the adult population of two national referral hospitals (2008-2012). AAS Open Research, 1(25).

4. Makau-Barasa, L. K., Greene, S., Othieno-Abinya, N., Wheeler, S. B., Skinner, A., and Bennett, A. V. (2020). A review of Kenya's cancer policies to improve access to cancer testing and treatment in the country. Health Research Policy and Systems, 18(1):1–10.

5. Nairobi, K. (2018). Kenya National Cancer Screening Guidelines. Nairobi: Ministry of Health, 1-122.

6. Kingham, T. P., Alatise, O. I., Vanderpuye, V., Casper, C., Abantanga, F. A., Kamara, T. B., ... & Denny, L. (2013). Treatment of cancer in sub-Saharan Africa. The Lancet Oncology, 14(4), e158-e167.

7. Odedina, F. T., Akinremi, T. O., Chinegwundoh, F., Roberts, R., Yu, D., Reams, R. R., ... & Kumar, N. (2009). Prostate cancer disparities in Black men of African descent: a comparative literature review of prostate cancer burden among Black men in the United States, Caribbean, United Kingdom, and West Africa. Infectious agents and cancer, 4(1), 1-8.

8. Gann, P. H. (2002). Risk factors for prostate cancer. Reviews in urology, 4(Suppl 5), S3.

9. Clarke, P., Crawford, C., Steele, F., & Vignoles, A. F. (2010). The choice between fixed and random effects models: some considerations for educational research.

10. Partin, M. R., Nelson, D., Radosevich, D., Nugent, S., Flood, A. B., Dillon, N., Holtzman, J., Haas, M., and Wilt, T. J. (2004). Randomized trial examining the effect of two prostate cancer screening educational interventions on patient knowledge, preferences, and behaviors. Journal of general internal medicine, 19(8):835–842.

11. Bowen, D., Hannon, P., Harris, J., and Martin, D. (2011). Prostate cancer screening and informed decision-making: provider and patient perspectives. Prostate cancer and prostatic diseases, 14(2):155–161.

12. Brant, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E., and Carter, H. B. (2003). Screening for prostate cancer by using random-effects models. Journal of the Royal Statistical Society: Series A (Statistics in Society), 166(1):51–62.

13. Ugwu, C. L. J., & Zewotir, T. T. (2018). Using mixed effects logistic regression models for complex survey data on malaria rapid diagnostic test results. Malaria journal, 17(1), 1-10.

14. Ali, S., Ali, A., Khan, S. A., & Hussain, S. (2016). Sufficient sample size and power in multilevel ordinal logistic regression models. Computational and mathematical methods in medicine, 2016.

15. Rezapour, M., Wulff, S. S., Mehrara Molan, A., & Ksaibati, K. (2021). Application of Bayesian ordinal logistic model for identification of factors to traffic barrier crashes: considering roadway classification. Transportation letters, 13(4), 308-314.

16. Li, B., Lingsma, H. F., Steyerberg, E. W., & Lesaffre, E. (2011). Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC medical research methodology, 11, 1-11.

17. Cochran WG. (1963). Sampling Techniques. 2nd ed. New York: John Wiley and Sons, Inc.

18. Siegel DA, O'Neil ME, Richards TB, Dowling NF, Weir HK. Prostate Cancer Incidence and Survival, by Stage and Race/ Ethnicity — United States, 2001–2017.Morb Mortal Wkly Rep. 2020;69 (41):1473-1480.

19. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). sage.

20. Sirengo, J. L., Alilah, D. A., Mbete, D. A., & Keli, R. (2023). Estimation of Risk Factors Affecting Screening Outcomes of Prostate Cancer Using the Bayesian Ordinal Logistic Model. Journal of Probability and Statistics, 2023.

21. Lee, H., & Kyung, M. (2014). Korean Welfare Panel Data: A Computational Bayesian Method for Ordered Probit Random Effects Models. Communications for Statistical Applications and Methods, 21(1):45-60.

22. Watanabe, Sumio (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research. 11(12): 3571–3594.

23. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian inference. Bayesian Data Analysis. 3rd ed. Boca Raton: CRC, 6-7.

24. Cerhan, J. R., Parker, A. S., Putnam, S. D., Chiu, B. C., Lynch, C. F., Cohen, M. B., ... & Cantor, K. P. (1999). Family history and prostate cancer risk in a population-based cohort of Iowa men. Cancer Epidemiology

Biomarkers & Prevention, 8(1), 53-60.

25. Rawla, P. (2019). Epidemiology of prostate cancer. World journal of oncology, 10(2), 63.

26. Bechis, S. K., Carroll, P. R., & Cooperberg, M. R. (2011). Impact of age at diagnosis on prostate cancer treatment and survival. Journal of Clinical Oncology, 29(2), 235-241.

27. Rodriguez, C., Freedland, S. J., Deka, A., Jacobs, E. J., McCullough, M. L., Patel, A. V., ... & Calle, E. E. (2007). Body mass index, weight change, and risk of prostate cancer in the Cancer Prevention Study II Nutrition Cohort. Cancer Epidemiology Biomarkers & Prevention, 16(1), 63-69.