https://doi.org/10.4038/sljastats.v24i1.8091

Sri Lankan Journal of Applied Statistics, Vol (24-1)

A Study on Factors Associated with Child Sexual Abuse and Recognizing the Severity: Special Reference to Galle District

L. H. K. Dilshan[1,*], N. Withanage[2], and N. V. Chandrasekara[1]

[1]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[2]Department of Statistics, University of Sri Jayawardenepura, Sri Lanka

*Corresponding author: kavindudilshan19971029@gmail.com
https://orcid.org/0009-0004-8167-1291

Received: 14th May 2023/ Revised: 12th June 2023/ Published: 30th June 2023

ABSTRACT

Child Sexual Abuse has been a global epidemic with devastating consequences. One in four girls and one in six boys have been experienced some form of sexual abuse in their tender age in the world. According to Police statistics, Child Sexual Abuse (CSA) cases is growing in recent years in Sri Lanka too. Galle is among the four districts where the reported child abuse cases high and the reported CSA complaints are increasing extraordinarily. Also, no previous research has been done in the Southern part of the country regarding the crisis of CSA. Hence, main objective of this study was to determine the key risk factors that could affect CSA in Galle Police Division, and to develop suitable regression and machine learning models to predict the severity of CSA. Data on 225 CSA cases reported to Police Child and Women Bureau of Galle Police Division during the period 2017 – 2020 were used in this study. According to chi-square test of association, out of twenty-one risk factors which were found from literature and knowledge of domain experts, sixteen variables showed significant relationship with the response variable, severity of CSA. Traditional Ordinary Logistic Regression (OLR) model was performed to predict severity of CSA and to detect key risk factors to CSA with two different data selection methods. Next, machine learning techniques: Decision Tree, SVM, and PNN were trained to classify severity of CSA. Random over-sampling technique was used to overcome the class imbalanced problem persists in the dataset. Finally, bagging technique was executed to conserve robustness of models and to improve performance. The OLR model classified the severity of CSA with 68.85% accuracy. Machine learning techniques,

Decision Tree, SVM and PNN model classified the severity of CSA with an accuracy of 82.15%, 77.68% and 85.25% respectively. PNN model performed with higher accuracy than the other fitted models. The results obtained from this study can be used to take precautions and to arrange awareness sessions for adults to reduce CSA in Galle Police Division. Also, the study can be extended to the whole island to reduce CSA and to make it a better place for children.

Keywords: Child Sexual Abuse, Ordinal Logistic Regression, Machine Learning Techniques, Oversampling, Bagging

1 Introduction

Child Maltreatment has been a universal issue with serious life-long consequences. Child Sexual Abuse (CSA) is the most common type of maltreatment in the world nowadays. One in four girls and one in six boys in the world have been experienced some form of sexual abuse in their childhood Zierler et al. (2018). Children may experience a wide range of short-term and long-term emotional, psychological, and physical problems because of being sexually abused. CSA is a silent and violent epidemic, silent because the victims and guardians are always reluctant to come out in public and look for justice Hébert et al. (2009), and violent because there are serious harmful effects on the casualty Bierre et al. (2003) , Browne et al. (1986). The WHO states that 1 in 5 women and 1 in 13 men report having been sexually abused as a child aged 0-17 years.

As reports of CSA feature in newsfeeds from across the island continuously with terrible frequency, it is obvious that Sri Lanka too is facing a nationwide crisis of CSA. Galle is among the four districts where the reported child abuse cases are high and the reported child abuse complaints are increasing drastically Veenema et al. (2018). Hence this study is based on the reported child sexual abuse cases in Galle Police Division during the period 2017-2020 based on Police Child and Women Bureau records.

In the modern world, parents spend a very busy life, and they might have forgotten their responsibilities as parents. So, the good connection between parents and children can be broken down with the loss of a protective and loving home environment. As a result of this, some children have distanced themselves from their parents and are tempted to find love and care from other people. Hence children are sadly misused by adults whom they rely upon for protection. With the loss of parents' protective environment, some children might become the victims of cruel people. Also, some people tend to abuse the children due to various reasons such as the lack of education, drug addiction, mental disorders, busy lifestyles, addiction to pornography videos Chen et al. (2018). Therefore, the number of CSA cases recorded also increased rapidly in Sri Lanka over the last few years. Therefore, detecting the key risk factors affecting CSA, and building suitable models to predict the severity of CSA are

very important to take actions and to develop policies to reduce the number of CSA cases in the future.

In Sri Lanka, many researchers have focused only on the characteristics and descriptive statistics of CSA. There are some studies that detect the key risk factors using chi-square tests and logistic regression Amararatne et al. (2016) , Chandrasiri et al. (2017) , Sathiadas et al. (2018). But none of the studies have used ordinal logistic regression models nor machine learning techniques to detect the key risk factors of CSA and to predict the severity of CSA in Sri Lanka. Although there are some studies that describes the characteristics of CSA in other provinces in Sri Lanka, up to my knowledge there is no study conducted in the Southern part of Sri Lanka in the past. Therefore, this study is conducted to determine the key risk factors affecting the CSA, to develop suitable models to classify the severity of CSA in Galle Police Division using, Ordinal Logistic Regression (OLR), Decision Trees (DT), Support Vector Machines (SVM), Probabilistic Neural Networks (PNN), and to make recommendations to overcome the crisis of CSA.

2 Methodology

2.1 Data Source

This study is considered on 225 CSA cases reported to Police Child and Women Bureau of Galle Police Division during the period 2017 – 2020. According to Police reports, a CSA can be categorized into three ordered categories namely: Not Fatal, Child Sexual Exploitation, and Fatal. The Chi-square test of association was used to test the significant relationship between response variable severity of CSA and each of twenty-one factors which were found from the literature and knowledge of domain experts. Those significant factors were considered as predictors in various models. The Figure 1 represents the proportions for each severity of Child Sexual Abuse reported in the sample.

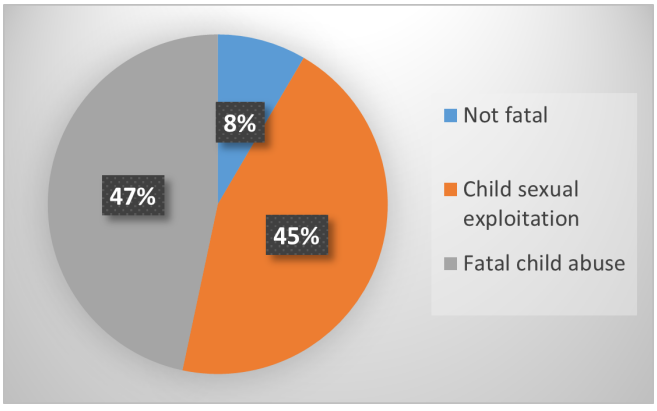

Fig 1: Chart of proportions of severity of CSA

According to the statistics of Police Child and Women Bureau in Galle Police Division, 47% of the child sexual abuse cases were fatal child abuse and 45% of the cases were child sexual exploitation. Meanwhile, the minority of the cases were not fatal.

2.2 Development of the design of the study

2.2.1 Ordinal Logistic Regression

Ordinal logistic regression is another extension of logistic regression that is used to predict an outcome variable with ordered multiple categories given one or more predictor variables. Logit model uses the ordering of the categories in forming logits. and they are simpler models with simpler interpretations and more powerful than (baseline) multinomial model (11). Ordinal Logistic Regression assumes that there is no multi-collinearity among predictor variables and the proportional odds assumption - i.e., the effects of any explanatory variables are consistent or proportional across the different thresholds.

2.2.2 Decision Trees

A decision tree is a supervised machine learning algorithm, used to represent the decisions and decision making visually and explicitly. These tree algorithms are tree-like structures that utilize the if-else statements to predict a result based on given data. The tree can be explained by four entities, namely root node, decision nodes, leaves, and arcs. The best splitting variable is kept as the root node at the top of the tree. The decision nodes are the nodes where the data is split, and each node is labelled with an attribute. The leaves which are at the bottom of the tree give the final outcomes or decisions, and each leaf node is labelled with a class.

2.2.3 Support Vector Machines

Support vector machines are supervised learning models that can be used for analysing data with classification and regression analysis. But it is widely used in classification tasks. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks.

2.2.4 Probabilistic Neural Network

A probabilistic neural network (PNN) is a radial basis neural network, which is widely used in classification and pattern recognition problems. Because of ease of training PNN has become a very effective tool for solving many classification problems. Also, it has the advantage of not being locked into

a local minimum. However, there is an outstanding issue of determining the network size, the locations of pattern layer neurons as well as the value of the smoothing parameter.

2.2.5 Class Imbalanced Problem

The class imbalance problem occurs when there are many more instances of some classes than others. Most of the classification algorithms are focusing more on classification of majority class records correctly while ignoring or misclassifying minority class records. The minority class samples are those that rarely occur but are very important with respect to many real-world problems.

2.2.6 Random Over Sampling

One approach to address the imbalance in a dataset is randomly oversample the minority class. Random oversampling involves duplicating records in the minority class. Although these records do not add any new information to the model, new records can be synthesized from the existing records. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique.

2.2.7 Bagging

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within dataset. Bagging significantly raises the stability of models with the improvement of accuracy. In bagging, the data is divided into a training set and testing set with replacement. After these large training and testing samples are generated, these models are then trained independently. Then the average or majority of those model predictions yield a more accurate estimate.

2.2.8 Data Splitting Methods

Different data splitting methods were considered for dividing the entire dataset into training and testing sets.

- Method 1: 80% of data (184) were used for model fitting from total cases, and the rest 20% (41) for testing.

- Method 2: Considering the class ratio of the response variable (the ratio of Not Fatal: Child Sexual Exploitation: Fatal Child Abuse is 8: 45: 47) of the original dataset, 80% (180) of the data was randomly selected for the training set and rest 20% (45) for the testing set.

- Method 3: This was an improved version of Method 2 with the use of over-sampling technique. By using the over-sampling technique and considering the class imbalance ratio between Not Fatal, CSE, and Fatal 240 of the data was randomly selected to build the model and 61 of the data was randomly selected to test the model.
- Method 4: This was an improved version of Method 3 with the use of bagging technique. Each individual model has been trained independently from the randomly chosen training samples in 10000 simulation runs. The dataset that solved the class imbalance problem in Method 3 was randomly divided into training and testing set with replacement in those 10000 simulations runs. Then each model was trained independently for each iteration with significant predictor variables. Then, the overall accuracy and classification accuracy of each class was stored in a vector for each iteration, and they are aggregated to make a collective decision. Finally, the average of those accuracies was taken to yield a more accurate estimate.

Each classification model was implemented considering each of the above data splitting methods separately.

2.2.9 Performance Measures

Accuracy is one very common method for evaluating classification models. Accuracy of predictions for a classification problem can be calculated as the number of correct predictions divided by the total number of cases.

Confusion Matrix
A confusion matrix illustrates the accuracy by summarizing prediction results on a classification problem. It gives insight not only into the errors being made by the classifier but more importantly the types of errors that are being made. For a given n response categories a confusion matrix D is a m x n matrix, where $C_{i,j}$ indicates the number of tuples from D that were assign to class $C_{i,j}$. It is obvious that the best solution will have only zero values in off-diagonal. The confusion matrix contains details about actual and predicted classifications performed by a classification problem.

Model Selection
Considering significant variables identified from the Chi-Square test, OLR was performed for all the data splitting methods. For selecting the "best" variables to the regression model stepwise backward selection method was applied. Then the multicollinearity and proportional odds assumption were checked for OLR Model. Only the predictors with severe multicollinearity were removed from the model with the use of Chi-Square test of Association. The best model was selected after considering the Deviance, Pseudo R Square, Pearson Chi-Square Goodness of fit statistics. After selecting the best model,

adequacy of the model was examined and classified the severity of CSA using the testing set.

The machine learning techniques DT, SVM, PNN were also developed to predict the severity of CSA in addition to the traditional statistical methods. The accuracy of each technique was compared with the help of overall and class accuracies calculated using the confusion matrices. Among them, the best model with the highest classification accuracy was selected to predict the severity of CSA in Galle Police Division.

3 Results and Discussion

Out of the 21 predictor variables, sixteen variables such as, area, child's age, gender, whether mother lives with child, willingness of child, frequency of abuses, place of incident, relationship to perpetrator, perpetrator's age, education level of perpetrator, perpetrator's job, marriage status, whether perpetrator has children, the number of children he has, drug addiction of perpetrator and reason including love affair, homosexual, excessive libido categories showed a significant relationship with the response variable according to the chi-square test of association under 10% level of significance.

OLR was fitted using the significant factors and considering aforementioned all the data splitting methods. Method 1 and Method 2 showed the overall accuracies of 73.17% and 77.18% respectively. But none of the observations in 'Not Fatal' category were not correctly classified in these two methods. This issue was further investigated, and it was observed that the data are suffering from class imbalanced problem. Random over sampling was used to overcome this problem and OLR Model 03 was fitted with the dataset that solved the class imbalanced problem as the Method 3. For selecting the "best" variables to the regression model stepwise backward selection method was applied. Area, gender, reason, frequency of abuses, place, perpetrator's job, and whether he has children identified as the significant predictor variables from OLR Model 03. Then the multicollinearity and proportional odds assumption were checked for OLR Model 03. Only the predictors with severe multicollinearity were removed from the model with the use of Chi-Square test of Association. Since OLR Model 03 satisfied the proportional odds assumption, OLR Model 03 was considered as the best model after considering the Analysis of Deviance, Pseudo R Square, Pearson Chi-Square Goodness of fit and the classification accuracy. Table 1 illustrates the confusion chart for the testing dataset of OLR Model 03.

Table 1 shows the performance of the OLR Model 03. 14 of the 20 'Not Fatal', 12 of the 20 'CSE', and 16 of the 21 'Fatal' are classified correctly by the model. Overall, the OLR Model 03 correctly predicts 68.85% of the CSA

Table 1: Confusion Chart for testing dataset (OLR Model 03)

	Expected values			
Predicted values	Not Fatal	CSE	Fatal	
Not Fatal	14	3	1	
CSE	6	12	4	
Fatal	0	5	16	
Accuracy	70%	60%	76.2%	Overall Accuracy =68.85%

cases. The equations for the OLR Model 03 are given below.

$$\begin{aligned}\text{logit}(P(Y \leq \text{Not\_Fatal})) &= -1.0346 - 1.4432 \cdot \text{Area}_{\text{Ahangama}} \\ &- 2.0824 \cdot \text{Area}_{\text{Hiniduma}} - 1.7826 \cdot \text{Area}_{\text{Akmeemana}} \\ &+ 2.1977 \cdot \text{Gender}_{\text{Male}} + 4.2580 \cdot \text{Reason}_{\text{Homo sexual}} + 1.0042 \cdot \text{Reason}_{\text{Love affair}} \\ &- 1.4549 \cdot \text{FreofAbuses}_{\text{Two times}} - 2.1072 \cdot \text{FreofAbuses}_{\text{Three times}} \\ &- 2.4492 \cdot \text{FreofAbuses}_{\text{More than three times}} + 1.4753 \cdot \text{Place}_{\text{In perpetrator's house}} \\ &+ 1.0727 \cdot \text{Place}_{\text{Outside}} - 1.4725 \cdot \text{Perpetrator'sJob}_{\text{Employed}} \\ &- 0.8309 \cdot \text{Perpetrator'sJob}_{\text{Unemployed}} - 2.3496 \cdot \text{WhetherHeHasChildren}_{\text{yes}}\end{aligned} \tag{1}$$

$$\begin{aligned}\text{logit}(P(Y \leq \text{CSE})) &= -1.0346 - 1.4432 \cdot \text{Area}_{\text{Ahangama}} \\ &- 2.0824 \cdot \text{Area}_{\text{Hiniduma}} - 1.7826 \cdot \text{Area}_{\text{Akmeemana}} \\ &+ 2.1977 \cdot \text{Gender}_{\text{Male}} + 4.2580 \cdot \text{Reason}_{\text{Homo sexual}} + 1.0042 \cdot \text{Reason}_{\text{Love affair}} \\ &- 1.4549 \cdot \text{FreofAbuses}_{\text{Two times}} - 2.1072 \cdot \text{FreofAbuses}_{\text{Three times}} \\ &- 2.4492 \cdot \text{FreofAbuses}_{\text{More than three times}} + 1.4753 \cdot \text{Place}_{\text{In perpetrator's house}} \\ &+ 1.0727 \cdot \text{Place}_{\text{Outside}} - 1.4725 \cdot \text{Perpetrator'sJob}_{\text{Employed}} \\ &- 0.8309 \cdot \text{Perpetrator'sJob}_{\text{Unemployed}} - 2.3496 \cdot \text{WhetherHeHasChildren}_{\text{yes}}\end{aligned} \tag{2}$$

It was found that the odds of being in a higher severity of CSA for Ahangama, Hiniduma, Akmeemana area are expected to decrease by 0.2362, 0.1246 and 0.1682 respectively compared to Galle while other variables hold constant. The model indicated that the odds of being in a higher severity of CSA for male child is expected to increase by 9.0043 compared to female child while other variables hold constant. Further, the odds of being in a higher severity of CSA for the reason of homosexual and love affair are expected to increase by 70.6685 and by 2.7297 respectively compared to the excessive libido while

other variables hold constant. Moreover, it was found that the odds of being in a higher severity of CSA for the place of incident is perpetrator's house and outside are expected to increase by 4.3723 and by 15.3283 respectively compared to the place of incident is in her/his own house while other variables hold constant. Additionally, the model concluded that the odds of being in a higher severity of CSA for the cases of perpetrator is employed and unemployed are expected to decrease by 0.2294 and by 0.4357 respectively compared to the cases of perpetrator is student while other variables hold constant. One of the findings of this model was that the odds of being in a higher severity of CSA for the cases where perpetrator has children is expected to decrease by 0.0954 compared to the cases where perpetrator hasn't children while other variables hold constant. Method 3 was implemented as the improved version of Method 2 with the use of bagging technique. OLR Model 04 with bagging technique as the method 03 predicted the severity of CSA with an overall accuracy of 67.18%. The accuracy of the classes 'Not Fatal', 'CSE', and 'Fatal' are 79.54%, 56.35% and 66.45% respectively. Machine learning methods were also developed to predict the severity of CSA. Table 2 summarizes the overall and class accuracies for each technique and data splitting method.

DT was trained with each of four data splitting methods. But DT showed somewhat lower classification accuracy compared to OLR models and still DT showed very low accuracy for the minority class, 'Not Fatal' (Method 1 and Method 2). Next, DT Model 03 was fitted with the dataset that solved the class imbalance problem and the classification accuracy was improved to the value of 72.13%. DT Model 04 was implemented using the bagging technique. The overall accuracy of the DT Model 04 was 82.15% and it exhibits better performance compared to the DT Model 03.

Next, SVM model was trained with respect to four data splitting methods. SVM performed as same as OLR models, but better than the DT models for the first two data splitting methods. The accuracies of the 'Not Fatal' class regarding SVM models were not too good. Then, data splitting method 3 was used to train the SVM Model 03 with different kernel functions, and it showed the better accuracy of 77.05% with the Linear Kernel Function. Then SVM Model 04 was performed using the Bagging technique and the overall accuracy was 77.68%.

Finally, the PNN model was trained with different spread parameters and PNN showed low accuracy for the data splitting method 1 compared to other techniques. Like OLR models, PNN models were unable to predict the minority class accurately. None of the observations were correctly classified as 'Not Fatal' for the first two methods. Consequently, PNN was again trained with the dataset that solves the class imbalance problem and showed the best accuracy with the 0.95 spread parameter. Figure 2 shows the optimum network architecture of the PNN network.

The PNN model was built with sixteen input variables with 0.95 spread. There are 240 neurons in pattern layer indicating the observations in the training set. In summation layer, there are 3 neurons representing the categories

Table 2: Classification accuracies of the models OLR, DT, SVM, and PNN with all data splitting methods for test set

	Model	Overall Accuracy	Accuracy of the class not fatal	Accuracy of the class CSE	Accuracy of the class fatal
OLR	Method 1	73.17%	0%	81.82%	75%
	Method 2	77.78%	0%	90%	81%
	Method 3	68.85%	70%	60%	76.2%
	Method 4	67.18%	79.54%	56.35%	66.45%
DT	Method 1	68.29%	33.3%	72.7%	68.8%
	Method 2	68.89%	0%	65%	85.7%
	Method 3	72.13%	70%	75%	71.4%
	Method 4	82.15%	94.07%	80.58%	73.26%
SVM	Method 1	73.17%	33.3%	77.2%	75%
	Method 2	77.78%	0%	90%	81%
	Method 3	77.05%	70%	85%	76.2%
	Method 4	77.68%	81.39%	81.84%	70.73%
PNN	Method 1	60.98%	0%	77.7%	56.3%
	Method 2	73.33%	0%	80%	81%
	Method 3	85.25%	100%	75%	81%
	Method 4	81.25%	98.37%	78.77%	68.49%

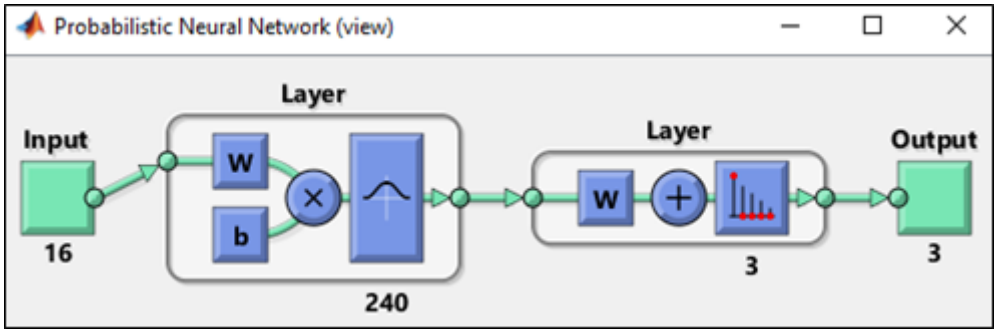

Fig 2: Network architecture of the PNN Model 03

of response variable (Severity of CSA). PNN predicted the severity of CSA with the best accuracy of 85.25% among all the techniques and data splitting methods.

Next, PNN Model 04 was performed using the Bagging technique to preserve the robustness of the models and to improve the performance. The overall accuracy of the PNN Model 04 was shown as 81.25%. The class accuracy for 'Not Fatal', 'CSE', and 'Fatal' are 98.37%, 78.77% and 68.49% respectively.

4 Conclusion

The main objective of this study was to determine the key risk factors and reasons affecting the child sexual abuse cases in Galle Police Division, and to develop suitable models to predict the severity of CSA. After fitting the suitable models to achieve these objectives, the final conclusions were drawn.

- Area, Gender, Reason, Frequency of Abuses, Place, Perpetrator's job, and whether he has children were identified as the most influential key risk factors to a CSA in Galle Police Division by the final OLR model.
- OLR model classified the severity of CSA with 67.18% accuracy.
- Machine learning techniques, the DT, SVM and PNN model classified the severity of CSA with an accuracy of 82.15%, 77.68% and 81.25% respectively for bagging technique.
- The PNN model performed with higher accuracy than the other fitted models for almost all the methods.
- Bagging Technique can be implemented to improve the performance of classification models.
- The results obtained from this study can be used to take precautions and to arrange awareness sessions for parents and adults to reduce CSA in Galle Police Division. Also, the study can be extended to the whole island to reduce CSA and to make it a better place for children.

References

Hébert M., Tourigny M., Cyr M., McDuff P. and Joly J. (2009) Prevalence of childhood sexual abuse and timing of disclosure in a representative sample of adults from Quebec, *The Canadian Journal of Psychiatry*, 54(9) : 631-636.

Browne A. and Finkelhor D., 1986. Impact of child sexual abuse: a review of the research,*Psychological bulletin*, 99(1) : p.66.

Bierre, J. and Elliot, D.M., 2003. Prevalence and psychological sequelae of childhood physical and sexual abuse in a general population sample of men and women. *Child Abuse Negl*, 27, pp.1205-22.

Amararatne, R.R.G.S. and Vidanapathirana, M., 2016. Child sexual abuse in Puttalam, Sri Lanka: a medico-legal analysis.

Chandrasiri, M., Wijewardena, D., Lanerolle, S., Chandrasiri, S., Wijewardena, K. and Cooray, R., 2017. Child sexual abuse presenting to district general hospital, Chilaw. *Ceylon medical journal*, 62(1).

Sathiadas, M.G., Viswalingam, A. and Vijayaratnam, K., 2018. Child abuse and neglect in the Jaffna district of Sri Lanka–a study on knowledge attitude practices and behavior of health care professionals. *BMC pediatrics*, 18, pp.1-9.

Veenema, T.G., Thornton, C.P. and Corley, A., 2015. The public health crisis of child sexual abuse in low and middle income countries: An integrative review of the literature. International journal of nursing studies, 52(4), pp.864-881.

Chen, L.P., Murad, M.H., Paras, M.L., Colbenson, K.M., Sattler, A.L., Goranson, E.N., Elamin, M.B., Seime, R.J., Shinozaki, G., Prokop, L.J. and Zirakzadeh, A., 2010, July. Sexual abuse and lifetime diagnosis of psychiatric disorders: systematic review and meta-analysis. In Mayo clinic proceedings (Vol. 85, No. 7, pp. 618-629). Elsevier.

Zierler, S., Feingold, L., Laufer, D., Velentgas, P., Kantrowitz-Gordon, I. and Mayer, K., 1991. Adult survivors of childhood sexual abuse and subsequent risk of HIV infection. American journal of public health, 81(5), pp.572-575.