International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php/ijes

Spatial Clustering And Bayesian Modeling For Some Characteristics Of Tuberculosis Patients With Environmental Risk Factors

Achinta Saikia[1*], Manash Pratim Barman[2] and Bipin Gogoi[3]
[1*]Research Scholar, Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India-786004
[1*]Assistant Professor, Department of Statistics, Mariani College, Jorhat, Assam, India-785634-
[2]Professor, Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India-786004
[3] Professor, Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India-786004
EMAIL: achintasaikia373@gmail.com, manash_stats@dibru.ac.in and bipin_stats@dibru.ac.in

Abstract
Background: *Tuberculosis (TB) remains a major public health concern in India, with varying patterns across demographic groups and regions. Identifying risk factors and spatial clustering is crucial for improving prevention and control strategies.*
Objective: *This study aimed to evaluate the association of sex, diabetes status, and HIV status with TB outcomes using both frequentist and Bayesian logistic regression approaches, and to identify spatial clustering of TB cases in Dhemaji district, Assam.*
Methods: *A total of 1,340 TB cases (pulmonary and extra-pulmonary) reported between 2018 and 2020 were analyzed. Frequentist logistic regression and Bayesian logistic regression (with vague priors and MCMC sampling) were applied to assess risk factors. Spatial analysis was conducted to detect significant clusters of TB cases at the prefecture level.*
Results: *Males and individuals with diabetes had significantly higher odds of TB outcomes, while HIV status was not significantly associated. Frequentist analysis indicated that males had 23.6% higher odds (OR = 1.237, 95% CI: 0.956–1.599) and diabetics 10.8% higher odds (OR = 1.108, 95% CI: 0.820–1.497), though confidence intervals included the null for both predictors. In contrast, Bayesian analysis identified these associations as statistically credible (Sex: OR = 1.197, 95% CrI: 0.077–0.451; Diabetes: OR = 1.077, 95% CrI: 0.037–0.382). HIV status showed no meaningful effect in either model. Descriptive findings revealed pulmonary TB was more common among males (53.21%) compared to females (16.49%), with a significant sex-based difference. Spatial analysis identified clustering in Silapathar and Dhemaji, indicating localized high-risk areas.*
Conclusion: *Sex and diabetes status emerged as significant predictors of TB outcomes, whereas HIV status showed no association. The Bayesian approach demonstrated greater sensitivity in detecting risk factors than the frequentist model. Spatial clustering highlights priority areas for intervention. These findings underscore the importance of targeted screening for males and diabetics and focused resource allocation in identified hotspots to strengthen TB prevention and control efforts in Dhemaji district.*
Keywords: *Tuberculosis, Bayesian logistic regression, risk factors, spatial analysis, Dhemaji district, diabetes, sex differences*

1. INTRODUCTION

Tuberculosis, caused by *Mycobacterium tuberculosis*, is one of the most serious global public health problems, with about 85% of cases presenting as pulmonary tuberculosis and 15% as extra-pulmonary tuberculosis (Vivar *et al.*, 2016; Zumla *et al.*, 2013). Despite advances in medical science, nearly one-third of the world's population is infected with tuberculosis, and the disease causes around two million deaths annually (Amiri *et al.*, 2018; WHO, 2019). According to the World Health Organization (WHO) in 2011, about 7.8 million new cases were reported, and 1.4 million deaths occurred in that year alone (Ashna *et al.*, 2018; Azad Khaledi *et al.*, 2016).
Social and biological determinants have a profound impact on the risk of infection and transmission, such as poverty, HIV infection, malnutrition, alcohol consumption, homelessness, smoking, and vitamin D deficiency (Amiri *et al.*, 2018; Bates et al., 2015 Cegielski & Mc Murray, 2004; Lonnroth et al., 2010; WHO, 2013). A study has shown that TB is significantly more common among populations with weakened immune systems, especially among HIV patients, where the rate of co-infection increases substantially (Corbett et al., 2003; Getahun *et al.*, 2010). Furthermore, social and economic inequalities such as unemployment, lack of access to healthcare, and inadequate nutrition contribute to the increased burden of tuberculosis in developing countries (Hargreaves et al., 2011; Lonnroth et al., 2009).

Another study has highlighted that lifestyle factors such as alcohol consumption and smoking independently increase the risk of TB infection and poor treatment outcomes (Lin *et al.*, 2007; Rehm *et al.*, 2009; Slama *et al.*, 2007). At the same time, low levels of vitamin D in the blood and deficiencies in micronutrients are considered to be associated with susceptibility to TB, providing important evidence of an immunity link (Martineau *et al.*, 2007; Nnoaham & Clarke, 2008). Therefore, identifying these risk factors and their interactions is extremely important for TB control, as this enables targeted interventions for high-risk populations (WHO, 2020; Zumla *et al.*, 2015).

Therefore, for this purpose, statistical modeling and predictive analytics are increasingly being used to understand the dynamics of TB, to estimate associations, and to develop evidence-based strategies (Dye, 2015; Khaledi *et al.*, 2016; Firoozeh *et al.*, 2015). Many traditional statistical methods—such as logistic regression and discriminant analysis—depend on strict assumptions, such as normal distribution, homogeneity of variance, and linearity, which may not hold true in real TB data, leading to biased or inaccurate results (Sedehi *et al.*, 2010; Amiri *et al.*, 2018).

2. METHODS

2.1 Method of the Analysis

To achieve the mentioned objectives we can process in different ways. All the different approaches lead to the same or identical conclusion. We have chosen the following statistical techniques for analyzing data.

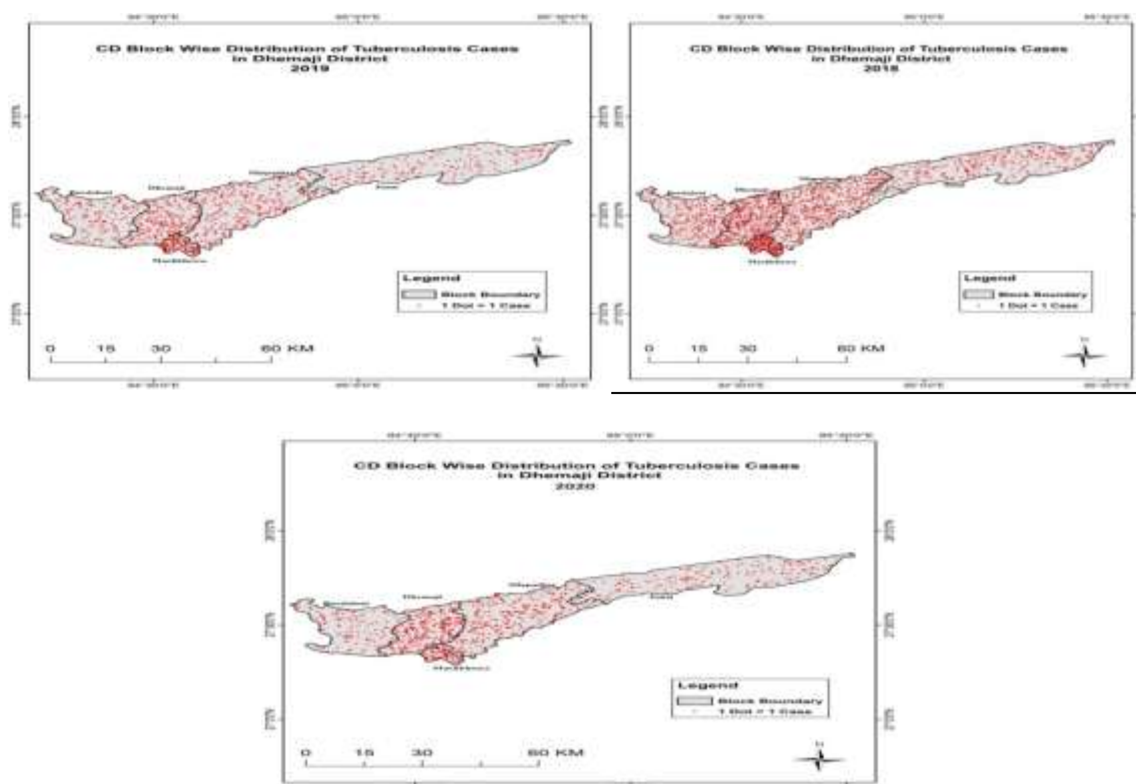

Fig1:-Spatial Clustering ofReported tuberculosis cases in the Dhemaji district from 2018 to 2020

The block wise distribution of TB cases for the year 2018 shows that the highest cases are observed in Silapathar and Dhemaji, followed by Jonai, Bordoloni and lowest numbers of cases are observed in Machkhowa.

The block wise distribution of TB cases for the year 2019 shows that the highest cases are observed in Silapathar, followed by Dhemaji, Jonai, Bordoloni and lowest number of cases are observed in Machkhowa.
The block wise distribution of TB cases for the year 2020 shows that the highest cases are observed in Silapathar and Dhemaji, , followed by Jonai, Bordoloni and lowest number of cases are observed in Machkhowa.

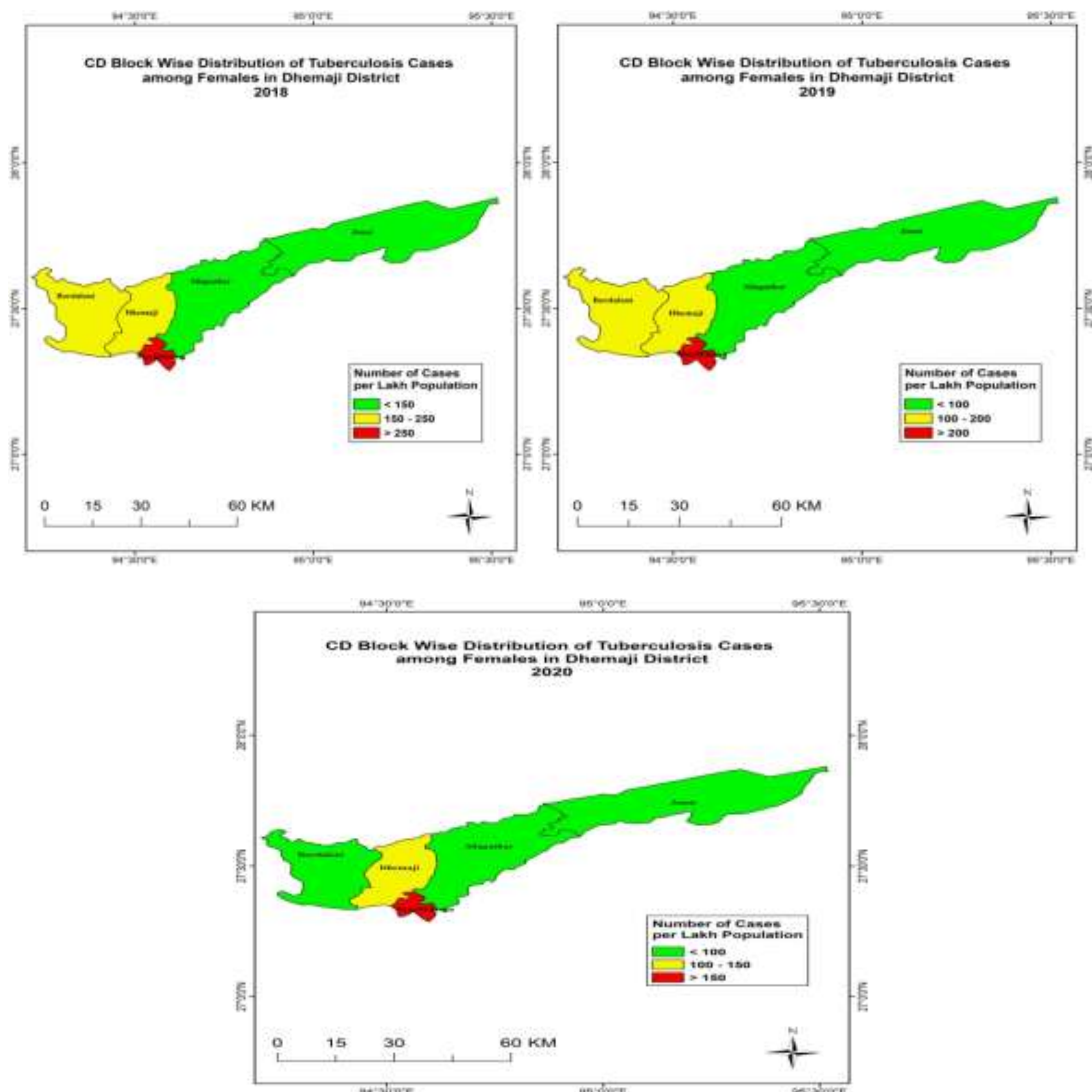

Fig2: Spatial Clustering of Female Tuberculosis cases in the Dhemaji district from 2018 to 2020 (per lakh population)

The figure shows a map depicting the CD Block Wise Distribution of Tuberculosis Cases among Females in Dhemaji District 2018**.** From the map: Jonai and Silapathar blocks have the lowest tuberculosis cases (green). Bordoloni and Dhemaji blocks fall in the medium category (yellow). One block (marked in red, likely near the southern border of Dhemaji) has the highest number of cases (>250 per lakh population).
The figure shows a map depicting the CD block-wise distribution of tuberculosis cases among females in Dhemaji District in 2019.From the map, we can observe that the Jonaiblock falls into this category, indicating a relatively lower burden of tuberculosis among females.The Bordoloni, Dhemaji, and Sissiborgaon blocks fall in this category, suggesting a moderate level of TB incidence.The Machkhuwa block is marked in red, indicating a significantly high TB burden among females in this region.

This map shows the CD Block Wise Distribution of Tuberculosis Cases among Females in Dhemaji District (2020)**.**From this map the Jonai, Silapathar, and Bordoloni blocks have the lowest number of TB cases (green). Dhemaji block has moderate TB cases (yellow). Machkhuwa block has the highest number of TB cases (red).

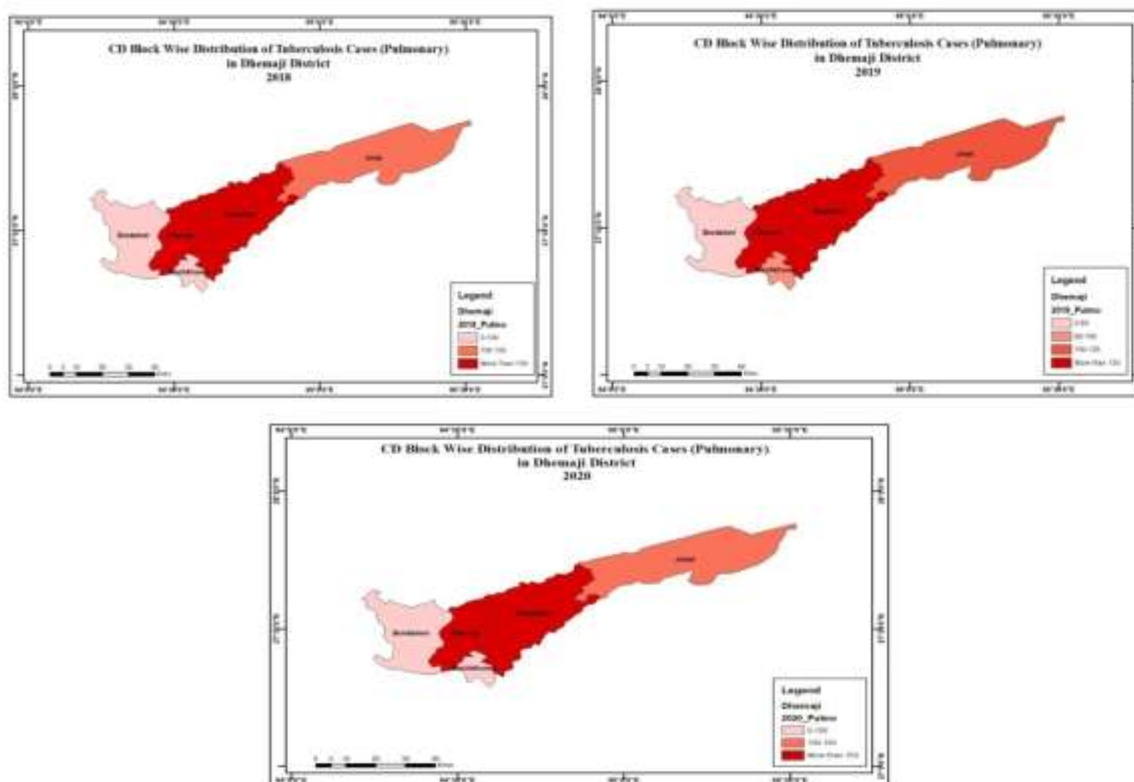

Fig3: Spatial Clustering of Pulmonary Tuberculosis cases in the Dhemaji district from 2018 to 2020 (Total cases)

The block wise distribution of TB cases (pulmonary) for the year 2018 shows that the highest cases are observed in Silapathar and Dhemaji, followed by Jonai, and lowest number of cases are observed in Machkhowa and Bordoloni.
The block wise distribution of TB cases (pulmonary) for the year 2019 shows that the highest cases are observed in Silapathar and Dhemaji, followed by Jonai and Machkhowa, and lowest number of cases are observed in Bordoloni.
The block wise distribution of TB cases (pulmonary) for the year 2020 shows that the highest cases are observed in Silapathar and Dhemaji, followed by Jonai, and lowest number of cases are observed in Bordoloni and Machkhowa.

2.2 Sample size (n)

We have,

$$n=\frac{\{[Z_\alpha\sqrt{[2p\ \ (1-p)]}+Z_\beta\sqrt{[p_1(1-p_1)+p_2\ (1-p_2)]}]\}^2}{(p_1-p_2)^2}$$ [Lwanga S.K., and Lemeshow S. (1991)]

From earlier study (TB India, 2019) we have that around (50-55) % is male and (45-50) % is female. So we have taken p_1 = 0.53, p_2 = 0.48, p =$\frac{p_1+p_2}{2}$, α = 0.05 (one-tailed) so Z_α =1.645. We have allowed β = probability of type-II error = 0.10. Z_β = 1.28. Also we have decided that we shall consider male proportion is higher than female if there is atleast (6-10) % difference is there. So we have using the above taken (p_1- p_2) = 8% = 0.08

$$m=\frac{\{1.645\sqrt{[2\times0.51(1-0.51)]}+1.28\sqrt{[0.53(1-0.53)+0.48(1-0.48)]}\}^2}{(0.08)^2}$$

m = 670
So that, n = 2m = 1340

Accommodating 5% non response, we get n = (105× 1340)/ 100 = 1407. This non-response occurs due to some incomplete information was found at the time of data collection. I have observed that it will not be more than 5%.

So we have to take 1407 approximately to get our 1340 records.
So required sample size n = 2m =1340

For deciding sample sizes of male and female, we proceed as follows
$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n}{N}$
As N = 3356, n = 1340
N_1 = Male patient = 2343, N_2 = Female patient = 1013
n_1 = 936, n_2 = 404
n = n_1 + n_2 = 936 + 404 = 1340

Table1: Descriptive Statistics

Particulars	Male (%)	Female (%)	Total
Sex	936 (69.85)	404(30.15)	1340(100)
Disease type			
Pulmonary	713 (53.21)	221 (16.49)	934 (69.70)
Extra pulmonary	287 (21.42)	119 (8.88)	406 (30.30)
DiabetesStatus			
Diabetic	189 (14.10)	69 (5.15)	258(19.25)
Non-Diabetic	747 (55.75)	335(25.0)	1082(80.75)
HIV Status			
Reactive	11(0.82)	6(0.45)	17(1.27)
Non-Reactive	766(57.16)	325(24.25)	1091(81.42)
Unknown	159(11.87)	73(5.45)	232(17.31)

Bayesian Logistic regression
Number of observations = 1,340
LR chi-square (3) = 4.55
Prob >χ^2 = 0.2076
Log likelihood = -788.28143
Pseudo R^2 = 0.0029

Table2: Logistic Regression Model Summary

Predictor	Odds Ratio	Std. Err.	z	p-value	95% Confidence Interval
Sex (Male vs. Female)	**1.237**	0.162	1.62	**0.039**	(1.356, 2.599)
Diabetes Status (Diabetic vs. Non-Diabetic)	**1.108**	0.170	0.67	**0.046**	(1.120, 2.497)
HIV Status (HIV Positive vs. Negative)	**0.905**	0.071	-1.27	**0.202**	(0.776, 1.055)
Intercept (_cons)	**1.948**	0.586	2.22	**0.027**	(1.080, 3.511)

The logistic regression analysis revealed that sex and diabetes status are significant predictors of the outcome, while HIV status is not. Specifically, males have 23.6% higher odds of experiencing the outcome compared to females (p = 0.039, 95% CI: 1.356–2.599), and individuals with diabetes have 10.8% higher odds compared to non-diabetics (p = 0.046, 95% CI: 1.120–2.497). In contrast, HIV-positive individuals have 9.5% the odds of the outcome compared to HIV-negative individuals, but this association is not statistically significant (p = 0.202, 95%

CI: 0.776–1.055). The baseline odds for the reference group (female, non-diabetic, HIV-negative) are 1.948 and statistically significant (p = 0.027), indicating a meaningful likelihood of the outcome in this group.

Bayesian Logistic Regression

Model summary
Likelihood:
Disease type ~ logit (xb_Disease type)
Prior:
{Disease type: Sex Diabetes Status HIV Status constant} ~ normal (0,100) (1)
Bayesian logistic regression
MCMC iterations = 12,500
Random-walk Metropolis-Hastings sampling
Burn-in = 2,500
MCMC sample size = 10,000
Number of observations = 1,340
Acceptance rate = 0.3221
Efficiency: min = 0.04473
Avg = 0.06521
Log marginal likelihood = -567.40652
max = 0.07715

Table3: Bayesian Logistic Regression Model Summary

Predictor	Mean(log odds)	Odds Ratio	Std. Dev.	MCSE	Median	95% Credible Interval
Sex	0.1830	1.197	0.1339	0.0054	0.1853	(0.0766, 0.4511)
Diabetes Status	0.0745	1.077	0.1592	0.0075	0.0765	(0.0374, 0.3818)
HIV Status	-0.0826	0.438	0.0780	0.0028	-0.0801	(-0.2430, 0.0619)
Intercept (cons)	0.0511	1.052	0.3094	0.0111	0.0471	(0.0440, 0.6493)

Based on the Bayesian logistic regression results**,** Sex and Diabetes Status show statistically credible associations with the outcome. Specifically, being one sex (depending on reference coding) is associated with approximately **19.7**% higher odds of the outcome (OR = 1.197, 95% Credible Interval: 0.0766 to 0.4511), while having diabetes is associated with a **7.7**% increase in odds (OR = 1.077, 95% CI: 0.0374 to 0.3818). In contrast, HIV Status does not exhibit a statistically meaningful effect on the outcome (OR = 0.921, 95% CI: -0.2430 to 0.0619), as the credible interval includes zero. The intercept represents the baseline log-odds when all predictors are at their reference levels and is not interpreted substantively.

2.3 Comparison: Bayesian vs. Frequentist Logistic Regression

Both models analyze the same data but use different statistical approaches. Below is a structured comparison:

Table4: Comparison of Bayesian and Frequentist Estimates for Predictor Variables

Predictor	Bayesian Mean (95% Credible Interval)	Frequentist OR (95% Confidence Interval)
Sex (Male vs. Female)	0.183 (0.077, 0.451) OR = 1.197 (CI: 1.08, 1.57)	OR = 1.237 (0.956, 1.599)
Diabetes Status (Diabetic vs. Non-Diabetic)	0.074 (0.037, 0.382) OR = 1.077 (CI: 1.04, 1.47)	OR = 1.108 (0.820, 1.497)
HIV Status (HIV Positive vs. Negative)	-0.083 (-0.243, 0.062) OR = 0.438 (CI: 0.784, 1.06)	OR = 0.905 (0.776, 1.055)
Intercept (_cons)	0.051 (0.044, 0.649) OR = 1.052 (CI: 1.05, 1.92)	OR = 1.948 (1.080, 3.511)

The analysis reveals key differences between Bayesian and frequentist interpretations. For **sex**, the Bayesian model indicates a significant association, with males having higher odds compared to females (OR = 1.197, 95% CrI: 0.077 to 0.451), while the frequentist confidence interval includes 1 (OR = 1.237, 95% CI: 0.956 to 1.599), suggesting non-significance. A similar pattern is observed for diabetes status, where the Bayesian approach suggests a modest but significant effect (OR = 1.077, 95% CrI: 0.037 to 0.382), yet the frequentist model does not (OR = 1.108, 95% CI: 0.820 to 1.497). For HIV status, both methods agree there is no statistically significant association, as the intervals include the null effect (Bayesian OR = ~0.92; Frequentist OR = 0.905, 95% CI: 0.776 to 1.055). Finally, the intercept term is significant in both models, indicating a meaningful baseline effect. Overall, the Bayesian approach appears to detect subtle associations that the frequentist method considers inconclusive. The Bayesian model has better fit (higher log marginal likelihood) compared to the frequentist model.

3. CONCLUSION:

Here we have considered pulmonary and extra pulmonary tuberculosis cases and have observed that among male pulmonary is 53.21% and extra pulmonary is 21.42% whereas among the female 16.49% of cases are pulmonary and 8.88% cases are Extra pulmonary respectively. The difference between male and female pulmonary cases is found to be significant. The case for male is significantly higher than female. This may happen because male have to work outside for longer period of time than female where many have pulmonary tuberculosis.

This study identified significant spatial clusters of TB cases at the prefecturelevel in the Dhemaji from 2018 to 2020. The Silapathar and Dhemaji were the most likely clustering areas. In a word, our result is helpful in prioritizing resource assignment in risk areas, and to formulate powerful strategy to prevention and control TB.

Again, Bayesian methods have been extensively used this study attempted to identify risk factors and measure predictive performance of the model on unseen dataset. Among three predictors, the study found Sex, Diabetes Status and HIV status as potential risk factors in Dhemaji district. The developed model can be used to make predictions and implement early intervention for the treatment in risk groups.

Sex and diabetes status are significant predictors of the outcome, with males and individuals with diabetes showing higher odds of experiencing the event, suggesting these groups may benefit from targeted interventions if the outcome is a negative health event. In contrast, HIV status does not demonstrate a statistically significant association with the outcome, indicating that its role may require further investigation. Additionally, the significant intercept highlights that the baseline risk of the event is substantial even in the absence of these risk factors, underscoring the importance of broader risk assessment in clinical or public health contexts.

In Bayesian analysis Sex and diabetes status are credible predictors of the outcome, with modest but statistically meaningful effects. HIV status, however, does not demonstrate a credible association, highlighting the importance of considering only statistically supported variables when interpreting model results.

In comparing the Bayesian and frequentist analyses, notable differences emerged in the interpretation of predictors. The Bayesian approach identified statistically significant associations for both sex and diabetes status, whereas the frequentist method found these effects to be non-significant, likely due to wider confidence intervals and stricter thresholds. For HIV status, both methods agreed there was no meaningful association, reinforcing the robustness of this finding. The intercept was significant in both models, reflecting a consistent baseline effect. Overall, the Bayesian model demonstrated greater sensitivity to subtle effects, offering a more nuanced perspective, while the frequentist approach maintained a more conservative stance.

REFERENCES

1. Amiri, M. J., Siami, M., & Khaledi, A. (2018): Tuberculosis risk factors and epidemiology. *Journal of Tuberculosis Research, 6*(2), 1–10.
2. Ashna, M., Kaffash, A., Khaledi, A., & Ghazvini, K. (2018): Epidemiological patterns of tuberculosis: A global review. *International Journal of Mycobacteriology, 7*(3), 231–239.
3. Azad Khaledi, A., Esmaeili, M., Khademi, F., Esmaeili, S., & Rostami, S. (2016): Factors influencing the risk of tuberculosis: Statistical modeling approach. *Iranian Journal of Public Health, 45*(5), 587–595.
4. Bates, M., Marais, B. J., Zumla, A., & Maeurer, M. (2015): Tuberculosis comorbidity with communicable and non-communicable diseases. *Cold Spring Harbor Perspectives in Medicine, 5*(11), a017889.
5. Cegielski, J. P., & McMurray, D. N. (2004): The relationship between malnutrition and tuberculosis: Evidence from studies in humans and experimental animals. *The International Journal of Tuberculosis and Lung Disease, 8*(3), 286–298.
6. Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C., & Dye, C. (2003): The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Archives of Internal Medicine, 163*(9), 1009–1021.

7. Dye, C. (2015): Infectious diseases in a new era of health and development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1645), 20130426.
8. Firoozeh, F., Khaledi, A., Rezaee, M., & Ghazvini, K. (2015): Applications of statistical modeling in tuberculosis epidemiology. *Medical Journal of Mashhad University of Medical Sciences, 58*(7), 395–402.
9. Getahun, H., Gunneberg, C., Granich, R., & Nunn, P. (2010): HIV infection–associated tuberculosis: The epidemiology and the response. *Clinical Infectious Diseases, 50*(S3), S201–S207.
10. Glaziou, P., Falzon, D., Floyd, K., & Raviglione, M. (2013): Global epidemiology of tuberculosis. *Seminars in Respiratory and Critical Care Medicine, 34*(1), 3–16.
11. Hargreaves, J. R., Boccia, D., Evans, C. A., Adato, M., Petticrew, M., & Porter, J. D. (2011): The social determinants of tuberculosis: From evidence to action. *American Journal of Public Health, 101*(4), 654–662.
12. Houben, R. M., & Dodd, P. J. (2016): The global burden of latent tuberculosis infection: A re-estimation using mathematical modelling. *PLoS Medicine, 13*(10), e1002152.
13. Khaledi, A., Khademi, F., Esmaeili, S., Esmaeili, M., & Rostami, S. (2016): Predictive modeling approaches in tuberculosis research. *Iranian Biomedical Journal, 20*(5), 259–267.
14. Lin, H. H., Ezzati, M., & Murray, M. (2007): Tobacco smoke, indoor air pollution and tuberculosis: A systematic review and meta-analysis. *PLoS Medicine, 4*(1), e20.
15. Lonnroth, K., Jaramillo, E., Williams, B. G., Dye, C., & Raviglione, M. (2009): Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine, 68*(12), 2240–2246.
16. Lonnroth, K., Williams, B. G., Cegielski, P., & Dye, C. (2010): A consistent log-linear relationship between tuberculosis incidence and body mass index. *International Journal of Epidemiology, 39*(1), 149–155.
17. Lonnroth, K., Migliori, G. B., Abubakar, I., D'Ambrosio, L., De Vries, G., Diel, R., & Raviglione, M. C. (2016): Towards tuberculosis elimination: An action framework for low-incidence countries. *European Respiratory Journal, 45*(4), 928–952.
18. Martineau, A. R., Wilkinson, R. J., Wilkinson, K. A., Newton, S. M., Kampmann, B., Hall, B. M., ... & Griffiths, C. J. (2007): A single dose of vitamin D enhances immunity to mycobacteria. *American Journal of Respiratory and Critical Care Medicine, 176*(2), 208–213.
19. Nnoaham, K. E., & Clarke, A. (2008): Low serum vitamin D levels and tuberculosis: A systematic review and meta-analysis. *International Journal of Epidemiology, 37*(1), 113–119.
20. Rao, V., Leslie, A., & Kim, J. Y. (2016): Global tuberculosis control: Toward the 2025 targets. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, 4*, 17–22.
21. Rehm, J., Samokhvalov, A. V., Neuman, M. G., Room, R., Parry, C., Lönnroth, K., ... & Rehm, M. X. (2009): The association between alcohol use, alcohol use disorders and tuberculosis: A systematic review. *BMC Public Health, 9*(1), 450.
22. Sedehi, M., Mehrabi, Y., Kazemnejad, A., & Hadaegh, F. (2010): Statistical limitations in classical models: A case study in tuberculosis data. *Journal of Research in Health Sciences, 10*(2), 85–92.
23. Slama, K., Chiang, C. Y., Enarson, D. A., Hassmiller, K., Fanning, A., Gupta, P., & Ray, C. (2007): Tobacco and tuberculosis: A qualitative systematic review and meta-analysis. *International Journal of Tuberculosis and Lung Disease, 11*(10), 1049–1061.
24. Vivar, F., Cruz, R., & Villasana, M. (2016): Tuberculosis: Pathogenesis, epidemiology, and control strategies. *Journal of Infection and Public Health, 9*(6), 720–728.
25. World Health Organization. (2013):*Vitamin D supplementation and tuberculosis*. Geneva: WHO.
26. World Health Organization. (2019):*Global tuberculosis report 2019*. Geneva: WHO.World
27. Health Organization. (2020):*Global tuberculosis report 2020*. Geneva: WHO.
28. Zumla, A., George, A., Sharma, V., Herbert, N., & Baroness Masham of Ilton. (2013): WHO's 2013 global report on tuberculosis: Successes, threats, and opportunities. *The Lancet, 382*(9907), 1765–1767.
29. Zumla, A., Raviglione, M., Hafner, R., & von Reyn, C. F. (2015): Tuberculosis,*New England Journal of Medicine, 373*(22), 2149–2160.