Logistic Regression Approach for Prediction of Gender Bias in the Prevalence of Tuberculosis in India

Vinay Kumar* and S. C. Malik

Department of Statistics, M.D. University, Rohtak, Delhi Road, University Secretariat, Rohtak - 124001, Haryana, India; vinay.stat@gmail.com, sc_malik@rediffmail.com

Abstract

Objectives: India is one of major contributors to global burden of tuberculosis which alone accounted for an estimated one quarter (26%) of all tuberculosis cases worldwide. The estimation of disease burden of tuberculosis is a challenge, considering its varied epidemiology and dynamics of transmission. As true disease burden cannot be estimated with count data therefore, statistical modelling techniques have been employed to analyze the disease burden in terms of prevalence of tuberculosis among males and females. Methods: An attempt has been made to predict gender bias in the prevalence of tuberculosis. The statistical models for both male and female tuberculosis positivity are developed on the information taken from NHFS (2015-16). The gender wise prevalence of tuberculosis has been compared by considering different indicators. In our analysis, the binary logistic regression model has been used by considering some factors to know their impact on prevalence of the tuberculosis. Results: Some of the variables under socioeconomic factors, Demographic Factors, Cultural factors and Health factors have shown decline their impact on prevalence of tuberculosis in males as compared females. However rests of the variables have the impact on the prevalence of tuberculosis without any variation. Conclusion: The study reveals that there are some factors which are responsible for prevalence of tuberculosis in India among male as well as in females and these factors are continuously contributing in increasing the prevalence of tuberculosis. Hence it is suggested that there is a need to redesign the policies to minimize the risk factors generated on the part of the factors having same impact on the prevalence of tuberculosis to avoid gender bias.

Keywords: Body Mass Index, Explanatory Variables and Logistic Regression, Risk Factors, Tuberculosis

1. Introduction

Tuberculosis (TB) is one of India’s most important public health problems and the country accounts for nearly one fifth of the global TB burden. Every year more than 20,000 people develop the disease, and more than 1000 die from TB globally. The medical and social dimensions of TB indicate that it is characterized by its close relation to poor socio-economic conditions. TB patients experience psychological and social sufferings and their basic rights may be negated. There are many problems faced by TB patients and social stigma has been always dominant amongst these problems. Social stigma is “an undesirable or discrediting attribute that an individual possesses, thus reducing that individual’s status in the eyes of society”. Because of this social stigma India remains far behind in managing this disease. The stigma attached to TB adds to the burden of disease for both men and women, and this burden is enhanced if the man and woman are adults. Both Man and woman face problems in context with their existing environment. On one side men have to deal with the stigma at their work place and at the community level, and on the other side women are faced with banishment within the household and in the surrounding area. The most crucial thing in their lives is that they are reticent in discussing their illness and participating in social func-

*Author for correspondence
Logistic Regression Approach for Prediction of Gender Bias in the Prevalence of Tuberculosis in India

2. Aim and Scope

In 1993, the WHO declared tuberculosis as a global emergency and the instituted WHO global tuberculosis control policy as a measure to help combat this epidemic. The logistic model also called growth model, had been used by various statisticians in different fields of specializations. It was used by Pearl and Reed to describe the growth of an albino rat and of a tadpole's tail. Berkson employed the logistic model for analysing bioassay data. Cox (1989) used the logistic model for handling quanta response data. Besides, have all used this model to classify observations into two or more groups. In this work, the logistic regression model is used to analyse gender bias in prevalence of tuberculosis. Here, for this unit level data of NFHS-4 has been utilized. The information in this survey is related to more than 90,000 households. In NHFS, the information regarding tuberculosis has been collected to know whether a household member is suffering from TB or not? The information on the background characteristics of households is considered for further statistical analysis and following fifteen independent variables that are listed by giving priority in the models based on their importance to prior studies.


Null values are assumed for males on variables 12-15, since these questions were included only on the women's survey in the original NHFS data.

The NHFS dataset has been analysed using a logistic regression modelling process. Since the most important dependent variable, TB positivity, is disaggregated into ‘positive’ and ‘negative responses from individuals, it can at most be treated as binomial data. The logistic regression technique is capable of predicting a binomial response using a series of independent traits, so this technique was the strongest possible analysis. Secondary models are made by filtering the NHFS dataset for respondents who would desire discretion about having TB (see Included Variables), and for those over the age of 35 since this category expresses the entirety of the gender discrepancy. First, a general model is developed to predict TB positivity, where gender is included as a variable. This is obtained by combining all the independent variables. However, few variables are proved insignificant. The list is pared down such that only the strongest variables are left in the final analysis. The individual (male vs. female) logistic regression models are obtained in a similar fashion, but in this case categorical dependent variables are set.

Lastly, a series of comparative models are created using the maximum set of independent variables to predict TB positivity for each of the following categories: total men, total women, and men over the age of 35, women over the age of 35, men who desired discretion about TB infection, and women who desired discretion about TB infection. These models provide β coefficients that can be compared at the scale level, and thus present a powerful measure of relative variance in TB trends between men and women. All the Data Analysis work has been carried out in SPSS 20.0 software.

3. Statistical Methodology

3.1 Logistic Regression Model

Logistic Regression is a mathematical modelling approach that can be used to describe the relationship between a binary dependent variable and a set of continuous and/or discrete independent variables. Logistic regression is by far the most popular modelling procedure used to
analyse epidemiology data when the outcome measure is dichotomous. The logistic model specifies that probably of disease depends on set of variable \( X_1, X_2, \ldots, X_n \) in following way

\[
P = P \left( \frac{y=1}{X} \right) = \frac{1}{1 + \exp \left( - \beta_0 - \beta_1 X_1 - \ldots - \beta_p X_p \right)}
\]

Where \( y \) denotes either the presence (\( y=1 \)) or absence (\( y=0 \)) of disease and \( X \) denotes a set of \( p \) variables which may represents the risk factors. The \( \beta \)'s are parameters that represent the effects of \( X \)'s in the risk of the disease. Then logistic regression model (logit transformation of \( P \)), where:

\[
\text{Logit}(P) = \log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k
\]

Where \( \beta_0 \) is the constant of the equation and \( \beta_i \) is the coefficient of predictor variables.

Each of the dependent variables are modelled using logistic regression, for both men and women: overall TB positivity, positivity over the age of 35, and positivity in those who desire discretion about TB infection. It should be noted that some variables are available only for women, such as use of solid fuels and ability to seek healthcare. It should be noted that in almost every case \( \mu_p \) is found to be larger than \( \mu_t \) roughly by a factor of ten, suggesting some degree of efficacy in predicting TB positivity based on these models.

### 4. Results and Discussion

#### 4.1 General Logistic Regression Models

The formulae below represent an overall model to predict TB positivity using as many of the above variables as possible. Two models are presented. In the former, all of the variables above are included; in the latter, variables in which \( p > 0.05 \) are removed to see if a more powerful prediction could be made. Accordingly, a value for overall predictability of TB is included for both models. For this predictive function, it is necessary to have values for as many variables as possible, and so a roughly representative subsample of 66,475 men and 76,300 women are selected based on the criteria of at most one value missing. In cases where values are missing, a series mean is assumed for the sake of computation. The function was computed to solve for \( p \) by the formula \( p = \left( 10^\gamma / (1 + 10^\gamma) \right) \), and the mean of \( p \) for TB-positive individuals was found (\( \mu_p \)). This latter mean represents the ability for the model to predict TB positivity, since models are more powerful as \( \mu_p \) approaches 1 (TB-positive = 1, TB-negative = 0). This may be compared to the value \( \mu_t \), which represents the mean of \( p \) for TB-negative individuals. Accordingly, this value is more powerful as \( \mu_p \) approaches 0.

Each model is presented in the following form:

**Dependent Variable**

\[
[y] = \log \left( \frac{p}{1-p} \right) = \text{Y-Intercept} + \text{Coefficient} \left[ \beta_1 X_1 + \ldots + \beta_k X_k \right]
\]

Parenthesized values represent confidence levels, where the maximum permitted level for the latter model is \( p < 0.05 \). Pearson goodness-of-fit (GoF) values are also taken for each model; their significance level is represented below each. The maximum permitted confidence level for Pearson is \( p < 0.05 \) for the latter model, and the maximum permitted number of iterations is 100. Variables in this section are indexed as the following:

- \( X_1 \): Gender,
- \( X_2 \): Age greater than 35,
- \( X_3 \): Literacy: literate vs. illiterate,
- \( X_4 \): Wealth index: middle/high vs. low,
- \( X_5 \): Use of public healthcare,
- \( X_6 \): Smokes cigarettes,
- \( X_7 \): Knowledge of TB transmission,
- \( X_8 \): Desire for discretion about TB infection,
- \( X_9 \): Membership in a scheduled tribe,
- \( X_{10} \): Body mass index: normal vs. low,
- \( X_{11} \): Anaemia level: normal vs. anaemic,
- \( X_{12} \): Use of solid fuels in the household,
- \( X_{13} \): Need for permission to seek healthcare,
- \( X_{14} \): Trouble procuring funds for medical treatment,
- \( X_{15} \): Final say on own healthcare (women only).

#### Inclusive General Model

\[
Y = \log \left( \frac{p}{1-p} \right) = -7.489 + 1.124 X_1 + .868 X_2 + .710 X_3 + .516 X_4 + .988 X_5 + .174 X_6 + .541 X_7 + .283 X_8 + .424 X_9 + 1.303 X_{10} + .608 X_{11} + .181 X_{12} - .919 X_{13} + .497 X_{14} + .273 X_{15}
\]

#### Reduced General Model

\[
y = \log(p/1-p) = -7.458 + .768 X_1 + .827 X_2 + .734 X_3 + .563 X_4 + .998 X_5 + .562 X_7 - .294 X_8 + .404 X_9 + 1.206 X_{10} + .565 X_{11} + .451 X_{14}
\]

The findings of these overall models are appealing for few reasons. The first one is that male gender is confirmed to be a strong risk factor for TB. This finding is heavily opinionated the idea that the observed male bias is true. This, of course, does not preclude the possibility that women either have not been diagnosed with TB, or that they hide their TB from the surveyor for one reason or another. Nonetheless, it justifies the division of male from female into categories for separate analysis.
Second, low body mass index is also found to be a highly important predictor of TB. This result is also unsurprising, since BMI is an effective measure of malnutrition. This, combined with age older than 35 as a risk factor, confirms many of the notions about age and malnutrition presented in previous studies.

Third, tobacco smoking and solid fuel use are found to have a relatively small, or even negative, effect. This casts doubt on the effect of solid fuel use and tobacco smoking on TB positivity, though the lack of male data on solid fuel use may distort its true effect. In addition, tobacco smoking is ruled out altogether in the reduced general model due to its low confidence level.

Fourth, it should be noted conditionally that the need to ask permission to seek healthcare is a highly negative predictor of TB. This likely emerges as an effect of the lack of need for men to ask permission, for whom every case is assumed to be negative. A look at the women’s models below may shed more light on its relevance for women as opposed to men.

Lastly, it is interesting that the mean p values for TB-positive (µp) and TB-negative cases (µt) are so small for both the inclusive and reduced general models. There is a considerable difference between the values for both models, suggesting that the p value is roughly eight times higher for TB-positive than TB-negative persons in the former case. The latter case cannot be accurately measured, since zero is an infinitely small value. In sum, these models are able to predict to a small degree whether a person will have TB or not based on the above fifteen variables, though in most cases the difference is quite small.

4.2 Individual Logistic Regression Models

The following formulae represent logistic regression modelling for men's and women's TB positivity for each of the categories: total, age over 35, and desire for discretion about TB. In each variable, the expected TB-predicting option is registered as value = 1, and the null option as value = 0, e.g. for anaemia 1 = anaemic and 0 = not anaemic. A maximum of 50 iterations are permitted to find an optimal solution; models are rejected if they exceeded this processing time. The described models are built upon the same criteria, and are presented in the same format, as that above. Again, µpandµtare provided to give a measure for power of prediction.

4.2.1 Men's Prevalence of TB (No Filter)

\[ y = \log(p/(1-p)) = -9.121 + .392 X_1 + .018 X_2 + .179 X_3 + .293 X_4 + .211 X_5 + .142 X_6 + .352 X_7 + .494 X_8 + .252 X_9 + .187 X_{10} - .093 X_{11} \]

Where \( X_1 = \) Use of public healthcare (.000), \( X_2 = \) Current age of respondent (.000), \( X_3 = \) Knowledge of TB transmission (.000), \( X_4 = \) Literacy: literate vs. illiterate (.000), \( X_5 = \) Wealth index: middle/high vs. low (.000), \( X_6 = \) Belief that TB is spread by touching a person with TB (.004), \( X_7 = \) Belief that TB can be cured: yes vs. no (.000), \( X_8 = \) Body mass index: normal vs. low (.000), \( X_9 = \) Membership in a scheduled tribe (.000), \( X_{10} = \) Anaemia level: normal vs. anaemic (.001), \( X_{11} = \) Smokes cigarettes (.022).

The most significant factors in shaping men’s TB prevalence are the choice of a public rather than private health facility, illiteracy, low wealth index, disbelief that TB can be cured, low body mass index, anaemia, and membership in a scheduled tribe. Less important factors included smoking cigarettes, correct knowledge of TB transmission, higher age, and belief that TB can be spread by touching. Each of these responses is found to increase the respondent’s chances of having TB, with the highly notable exception of cigarette smoking--this slightly lowered the respondent’s chances of tuberculosis.

This model supports the belief that anaemia and low body mass index weaken the immune system to TB infection in men. Since these are both measures of malnutrition, these findings are unsurprising, along with the finding that low wealth index is a determinant for TB. Third, the finding that belief that TB cannot be cured maps onto TB positivity suggests that men may feel fatalistic about their infection, which may in turn lead them to noncompliance with treatment. This is an issue that needs to be addressed at the level of public health education.

These pronouncements support some earlier assertions about the epidemiology of TB prevalence among men, but challenge others. The finding that TB sufferers are more likely to use public, rather than private, health sources indicates that, for men, the overwhelming use of private health resources is not an issue, though it is clear that some populations remain underserved. It is also a fact that illiteracy is a determinant of TB in men, though this may be a symptom of general poverty no direct causal connection can be made between TB positivity and illiteracy.
Two covariates in this analysis are somewhat unexpected: the negative effect of smoking on TB positivity, and the positive effect of knowledge about transmission. These results suggest that smoking is not a good indicator of TB prevalence indeed; smokers seem to catch TB less often. Instead, knowledge about TB transmission appears to correlate with TB positivity. This may be the case because of personal relevance, since having TB encourages one to learn about TB.

4.2.2 Men’s Prevalence of TB (Filter: Age> 35)

\[ y = \log\left(\frac{p}{1-p}\right) = -7.142 + .744 X_1 + .652 X_2 + .912 X_3 - .459 X_4 + .796 X_5 + .412 X_6 + 1.041 X_7 + 1.814 X_8 - .497 X_9 \]

Where \( X_1 = \) Literacy: literate vs. illiterate (.000), \( X_2 = \) Wealth index: Middle/high vs. low (.000), \( X_3 = \) Use of public healthcare (.000), \( X_4 = \) Smokes cigarettes (.001), \( X_5 = \) Knowledge of TB transmission (.000), \( X_6 = \) Belief that TB is spread by touching a person with TB (.010), \( X_7 = \) Belief that TB can be cured (.008), \( X_8 = \) Body mass index: normal vs. low (.000), \( X_9 = \) Desire for discretion if family member has TB (.015)

Many of the relevant variables from the Total Men’s Prevalence of TB model are also represented here, with one important exception: desire for discretion about TB infection. In men who said they would want discretion, the likelihood of having TB was significantly lower. This indicates that older men may be hiding TB positivity, since a desire for discretion maps onto a higher social cost for having TB, and heavily supports the notion that stigmatization of TB may drive inaccuracy in reported prevalence statistics.

Lastly, it is notable that this category presents the only outlier in terms of power of prediction, since \( \mu_T \) is actually smaller than \( \mu_p \).

4.2.3 Men’s Prevalence of TB (Filter: Desire for Discretion about TB)

\[ y = \log(\frac{p}{1-p}) = -6.902 + .683 X_1 + .441 X_2 + .970 X_3 + 1.022 X_4 + .659 X_5 \]

Where \( X_1 = \) Body mass index: normal vs. low (.001), \( X_2 = \) Trouble procuring funds for medical treatment (.049), \( X_3 = \) Use of public healthcare (.000), \( X_4 = \) Smokes cigarettes (.018), \( X_5 = \) Current age > 35 (.000)

This model proved to be highly similar to the women’s overall regression. It is notable that illiteracy is not found to be a significant cofactor, but cigarette smoking was found to be highly important (\( \beta = 1.022 \)). This is the only category in which cigarette smoking correlated with TB positivity, though the implications of this finding are not clear. Secondly, women who stated they were older than 35 were found to be significantly more likely to be positive for TB.

4.2.4 Women’s Prevalence of TB (Filter: Desire for Discretion about TB)

\[ y = \log(\frac{p}{1-p}) = -7.056 + .932 X_1 + .411 X_2 + .747 X_3 + .608 X_4 + .022 X_5 \]

Where \( X_1 = \) Body mass index (.000), \( X_2 = \) Trouble procuring funds for medical treatment (.000), \( X_3 = \) Use of public healthcare (.000), \( X_4 = \) Literacy: literate vs. illiterate (.000), \( X_5 = \) Current age of respondent (.000)

Many of the variables affecting men’s TB positivity are also found to affect women’s rates of TB positivity, including low body mass index, use of public health facilities, illiteracy, and advanced age. Body mass index is found to be twice as strong a factor for women than men, along with use of public facilities and illiteracy. The effect of age is similar, but remained small (<.02). However, the knowledge determinants—namely, belief in a cure and TB transmission knowledge—did not prove significant, along with scheduled tribe status or anaemia.

Women’s positivity is predicted by an inability to seek healthcare due to financial restraints (this variable...
is not available for men). This supports the idea that certain populations remain underserved by the universal TB treatment system (DOTS), which is in theory free of charge. The bulk of the burden, then, either derives from the cost of transportation or cost of treatment at private facilities-invites research in this direction.

### 4.2.6 Women’s Prevalence of TB (Filter: Age > 35)

\[
y = \log(p/1-p) = -6.294 + 1.104 X_1 - .375 X_2 + .868 X_3 + .673 X_4
\]

Where \( X_1 \) = Body mass index: normal vs. low (.000), \( X_2 \) = Desire for discretion about TB infection (.032), \( X_3 \) = Literacy: literate vs. illiterate (.000), \( X_4 \) = Use of public healthcare (.000).

For women over the age of 35 the low body mass index, illiteracy, and use of public health facilities are all important determinants of TB positivity. Illiteracy and body mass index are stronger indicators for women over the age of 35 than in all age categories. In addition, desire for discretion about TB becomes an important factor here, in a manner similar to men over the age of 35 (\( \beta_{female} = -.375, \beta_{male} = -.497 \)). For now, it is also notable that women over the age of 35 do not have trouble seeking healthcare due to financial constraints-this may stem from a possible trend toward financial stability with age.

### 5. Conclusion

The findings support some assumptions about TB differential between genders, but challenge others. It is clear, most of all, that gender and tuberculosis interact in complex ways, and that the epidemiological process is loaded with difficulties that must be recognized and addressed before standard research can take place. In this way, the findings of this study may serve to highlight the structural issues within TB research. We are now discuss a few general themes in these results, and attempt to integrate them. First, it appears that the male bias presents itself in these results as a real and powerful risk factor, second only to a low body mass index. Assuming the data used here are accurate, this provides strong support for a biological explanation for the gender bias, having shown a reduced importance of cultural factors such as education, knowledge about how TB is transmitted and scheduled tribe membership. Interestingly, the methodological factor of a woman's lack of control over her health trajectory also proved to be relatively low-priority predictor of the presence of lack of TB. However, this does not exclude the existence of other procedural factors such as the desire for discretion about TB. However, there is only contingent indication for this claim. Second, it is highly important that the gender discrepancy in TB rates is concentrated almost entirely in male and female respondents over the age of 35. This appears to be a highly critical age interval for a study of gender and TB, though it is a domain in which little research has been conducted. It is also the area which requires more exploration; although there may be organic impenetrable factors that better explain the receptiveness of males in this period. Even so, it is for this reason that this group has been aside in the logistic analysis, so that those social or behavioural cofactors might come forward.

### 6. References

11. Nirmala DM, Balamurugan AA, Reshma M. Developing a Modified Logistic Regression Model for Diabetes Mellitus and Identifying the Important Factors of Type II Dm.

