ISSN: 09763031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 4(I), pp. 26159-26169, April, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

A STUDY ON THE INCIDENCE OF TUBERCULOSIS USING BINARY LOGISTIC REGRESSION

Senthilkumar V[1] and Sachithanantham S[2]

[1]Department of Statistics, Manonmaniam Sundaranar University, Thirunelvali 627012
[2]Department of Statistics, Arignar Anna Government Arts College, Villupuram 605602

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.2004

ARTICLE INFO

Article History:
Received 5th January, 2018
Received in revised form 20th February, 2018
Accepted 8th March, 2018
Published online 28th April, 2018

ABSTRACT

In many cases research focuses on models where the dependent variable is categorical. Instead we would carry out a logistic regression analysis. Hence, logistic regression may be thought of as an approach that is similar to that of multiple linear regressions, but takes into account the fact that the dependent variable is categorical. Binary responses are commonly studied in many fields. Examples include the presence or absence of a particular disease, death during surgery, or a consumer purchasing a product. However, in many situations, there are multiple descriptors, or one or more of the descriptors are continuous. Without a statistical model, studying patterns such as the relationship between age and occurrence of a disease, for example, would require the creation of arbitrary age groups to allow estimation of disease prevalence as a function of age. . In biomedical research it is common to observe multivariate time series data where the outcomes are binary. the purpose of analysis include assessing the association among variable at one time, identifying lead lag relationships among variables, and regressing, one outcome on others as well as on fixed covariates. In this paper the in cadence of tuberculosis using binary Logistic regression has been studied.

INTRODUCTION

The statistical properties of linear regression models are invariant to the (unconditional) mean of the dependent variable, the same is not true for binary dependent variable models. The mean of a binary variable is the relative frequency of events in the data, which, in addition to the number of observations, constitutes the information content For example, that logit coefficients are biased in small samples (under about 200) is well documented in the statistical literature, but not as widely understood is that in rare events data the biases in probabilities can be substantively meaningful with sample sizes in the thousands and are in a predictable direction: estimated event probabilities are too small. A separate, and also overlooked, problem is that the almostuniversally used method of computing probabilities of events in logit analysis is suboptimal in finite samples of rare events data, leading to errors in the same direction as biases in the coefficients. Applied researchers virtually never correct for the underestimation of event probabilities.

A second source of the difficulties in analyzing rare events lies in data collection. Given fixed resources, a tradeoff always exists between gathering more observations and including better or additional variables. In rare events data, fear of collecting data sets with no events (and thus without variation on *Y*) has led researchers to choose very large numbers of observations with few, and in most cases poorly measured, explanatory variables. This is a reasonable choice, given the perceived constraints, but it turns out that far more efficient data collection strategies exist. For one example, researchers can collect all (or all available) ones and a small random sample of zeros and not lose consistency or even much efficiency relative to the full sample. This result drastically changes the optimal tradeoff between more observations and better variables, enabling scholars to focus data collection efforts where they matter most. A detailed study an is given in section 2.

Latent Mycobacterium Tuberculosis Infection

Pathogenesis

After inhalation of *M. tuberculosis*, innate immune responses involving alveolar macrophages and granulocytes begin to combat the infection; in some persons, the bacilli are cleared, whereas in others, infection is established. Replication of bacilli in macrophages and regional lymph nodes leads to both

Corresponding author:* **Senthilkumar V
Department of Statistics, Manonmaniam Sundaranar University, Thirunelvali - 627012

lymphatic and hematogenous dissemination, with seeding of multiple organs, which may eventually give rise to extrapulmonary disease. Containment of bacilli within macrophages and extracellularly within granulomas limits further replication and controls tissue destruction, resulting in a dynamic balance between pathogen and host. The classic interpretation of this as a binary process with either truly latent *M. tuberculosis* infection or active tuberculosis disease has recently been challenged as an oversimplification. Instead, a spectrum of immunologic responses that are both protective and pathogenic and correlate with a range of bacterial activation has been suggested. This continuum encompasses a variety of host–microbe interactions, which are characterized by clinical latency when host responses predominate and by disease when bacterial replication exceeds the threshold required to cause symptoms. Recent evidence suggests that host inflammatory responses, particularly with interleukin1β, may actually enhance mycobacterial replication, which illustrates that the doubleedged sword of immune responses seen in tuberculosis disease may also be present in latent infection. In addition, persisting extracellular bacilli may remain active in a biofilmtype of environment and thus evade host defenses; in such cases, the term persistent (rather than latent) infection has been suggested to explain the complexity of the phenomenon. Animal models such as mice, guinea pigs, rabbits, macaques, and zebrafish have been used to study the pathogenesis and treatment of latent tuberculosis. A shortcoming of all models, however, is the lack of pathological, clinical, and therapeutic conformity with human infection and disease. Thus, each model may be used to elucidate some aspects of the human situation mice, for example, recapitulate human treatment experiences, whereas rabbits display histopathological features that are similar to those in humans but no model can capture the full spectrum of infection, disease, and treatment.

Epidemiology and Risk Groups

Current tools are insufficient to measure the global prevalence of latent tuberculosis infection, but modeling carried out a decade ago estimated that approximately one third of the world population (>2 billion people) is latently infected with *M. tuberculosis*.9 Currently, annual rates of infection range from 4.2% in South Africa10 and 1.7% in Vietnam11 to 0.03% in the United States. As tuberculosis treatment has expanded in the past 15 years and living conditions have improved worldwide, the annual risk of infection may have declined in many places; the current global burden of latent infection is therefore uncertain and needs to be reassessed. Persons with untreated tuberculosis of the respiratory tract are the source of transmission in essentially all new cases of tuberculosis infection, and up to one third of their household contacts become infected. Factors associated with an increased risk of infection in a household contact include severe disease in the index patient, long periods of exposure to the index patient, and poor ventilation and poor exposure to ultraviolet light during proximity to the index patient. Reactivation of latent tuberculosis infection accounts for the majority of new tuberculosis cases, especially in countries in which the incidence of tuberculosis is low.

The likelihood of progression of latent infection to active clinical tuberculosis disease is determined by bacterial, host, and environmental factors. It has been postulated that there are differences in the ability of various strains of *M. tuberculosis* to cause disease, but little clinical or epidemiologic data support this theory. The initial bacterial load, inferred by the severity of disease in an index case and the closeness of the contact, is directly associated with the risk of development of the disease. Disease develops at a higher rate among infants and very young children who have latent infection than among older children with latent infection; after a child reaches approximately 5 years of age, age appears to have little correlation with the risk of disease.

Suppression of cellular immunity by human immunodeficiency virus (HIV) infection,14 tumor necrosis factor α inhibitors, glucocorticoids, and organ or hematologic transplantation increases the risk of progression of latent infection substantially. Endstage renal disease confers an increased likelihood of progression to active tuberculosis. Silicosis and exposure to silica dust are also associated with increased rates of progression, and the combination of HIV and silicosis in South African miners has contributed to an explosive epidemic of tuberculosis in this population. Other risk groups that should be considered for management of latent tuberculosis infection on the basis of a high prevalence or an increased risk of active tuberculosis disease include prisoners, illicitdrug users, homeless adults, recent immigrants from countries that have a high tuberculosis burden, the elderly, health care workers and medical students, patients with diabetes, and persons with recent conversion of a negative tuberculin skin test to a positive test. Table 1 presents the range of published data on the risk of active tuberculosis and the prevalence of latent infection in selected highrisk groups.

Diagnosis

There are no perfect methods for the diagnosis of latent tuberculosis infection. The tuberculin skin test and the IGRAs indirectly measure tuberculosis infection by detecting memory Tcell response, which reveals only the presence of host sensitization to M. tuberculosis antigens.8 The tests are generally considered to be acceptable but imperfect. The tuberculin skin test is widely used and inexpensive, but it has poor specificity in populations vaccinated with bacille Calmette–Guérin (BCG), is subject to crossreactivity with environmental nontuberculosis mycobacteria, and has poor sensitivity in immunocompromised persons. There are also logistic drawbacks, including the need for a return visit in 2 to 5 days to read the amount of induration, since selfreading is associated with a high error rate. Furthermore, there is a worldwide shortage of tuberculin, attributed to market forces. IGRAs (the QuantiFERONTB Gold InTube assay [Cellestis] and the TSPOT.TB assay [Oxford Immunotec]) measure in vitro responses of T cells or peripheralblood mononuclear cells to M. tuberculosis antigens that are not found in BCG and most nontuberculous mycobacteria, and thus specificity for M. tuberculosis is higher than with the tuberculin skin test. However, recent studies involving serially tested health care workers in the United States have shown that false conversions (from a negative to a false positive result) and reversions (from a positive to a false negative result) are more common with IGRAs than with tuberculin skin tests. In addition, IGRAs are more costly and require more work in the laboratory. The ability of tuberculin skin tests and IGRAs to identify persons at

the highest risk of progressing to active tuberculosis (i.e., the positive and negative predictive values) is poor. Neither test can reliably predict future disease among persons with positive tests, and strong positive tests do not suggest a higher risk. In one metaanalysis, the pooled positive predictive value for progression to active tuberculosis was 2.7% (95% confidence interval [CI], 2.3 to 3.2) for IGRAs and 1.5% (95% CI, 1.2 to 1.7) for the tuberculin skin test. A metaanalysis of only longitudinal studies of IGRAs, with a median followup of 4 years, showed a moderate association between positive tests and subsequent tuberculosis (pooled, unadjusted incidence ratio, 2.10 [95% CI, 1.42 to 3.08]).30 In a 2year prospective study in the India involving adult contacts of persons with active tuberculosis, a positive IGRA was associated with a significantly higher risk of the development of tuberculosis except among contacts older than 35 years of age. The comparative performance of the tuberculin skin test and IGRAs varies between highincidence countries and lowincidence countries, possibly because of the effects of BCG vaccination and reinfection. Computed tomography might prove to be a promising complementary imaging method to chest radiography in distinguishing latent tuberculosis infection from active disease.32 Although currently no standard immunodiagnostic biomarkers have been identified to measure latent tuberculosis infection, there is a growing landscape of chemokines, tumor necrosis factor, interleukins, growth factors, and soluble receptors under development that could improve diagnostic capacity.

compared with placebo, the risk of progression to active tuberculosis at 6 months (relative risk, 0.44; 95% CI, 0.27 to 0.73) is similar to that at 12 months (relative risk, 0.38; 95% CI, 0.28 to 0.50). Isoniazid was associated with a reduction in the incidence of tuberculosis among persons with HIV who were receiving antiretroviral therapy, and one study showed the benefit of isoniazid in patients with negative tuberculin skin tests or IGRAs who were also receiving antiretroviral therapy. A recent study from Uganda showed a high rate of conversion from a negative tuberculin skin test to a positive tuberculin skin test (30 cases per 100 personyears) among persons with HIV during the first 6 months of antiretroviral therapy. In geographic areas known for a high rate of transmission of tuberculosis, the protective effect of isoniazid against tuberculosis among people with HIV wanes over time, and continuous protection is maintained through a lifetime duration of treatment for tuberculosis. The World Health Organization recommends that HIVinfected persons in countries with high rates of transmission of tuberculosis receive at least 36 months of isoniazid as a proxy for lifelong treatment. In Brazil, a country with low rates of transmission of tuberculosis, isoniazid therapy for 6 months has been shown to have longterm protective benefits in HIVinfected adults. Other effective regimens are daily rifampin for 3 or 4 months, daily isoniazid and rifampin for 3 months, and isoniazid (900 mg) and rifapentine (900 mg) once weekly for 12 weeks. A regimen of rifampin and pyrazinamide that was initially shown to be effective in people with HIV infection was found to cause

Table 1 Incidence of Active Tuberculosis and Prevalence of Latent Tuberculosis Infection in Selected High-Risk Groups, According to Published Studies

High-Risk Group	Incidence of Active Tuberculosis	Prevalence of Latent Tuberculosis Infection†		
		QuantiFERON-TB Gold In-Tube	T-SPOT.TB	Tuberculin
	median rate per 1000 population (range)	*median percentage (range)*		
Persons with HIV infection	16.2 (12.4–28.0)	14.5 (2.7–21.5)	11.3 (4.3–67.6)	19.2 (2.1–54.8)
Adult contacts of persons with tuberculosis	0.6‡	21.1 (6.6–55.1)	48.0 (29.6–59.6)	26.3 (1.8–82.7)
Patients receiving tumor necrosis factor blockers	1.4‡§	11.8 (4.0–22.3)	20.0 (12.9–25.0)	18.6 (11.3–68.2)
Patients undergoing hemodialysis	26.6 (1.3–52.0)	33.4 (17.4–44.2)	43.6 (23.3–58.2)	21.9 (2.6–42.1)
Patients undergoing organ transplantation	5.1‡	21.9 (16.4–23.5)	29.5 (20.5–38.5)	7.7 (4.4–21.9)
Patients with silicosis	32.1‡	46.6‡	61.0‡	-
Prisoners	2.6 (0.03–9.8)	-	-	45.5 (23.1–87.6)
Health care workers	1.3 (0.4–4.1)	14.1 (0.9–76.7)	5.2 (3.5–28.7)	29.5 (1.4–97.6)
Immigrants from countries with a high tuberculosis burden	3.6 (1.3–41.2)	30.2 (9.8–53.8)	17.0 (9.0–24.9)	39.7 (17.8–55.4)
Homeless persons	2.2 (0.1–4.3)	53.8 (18.6–75.9)	-	45.6 (20.5–79.8)
Illicit-drug users	6.0‡	63.0 (1.4–66.4)	45.8 (34.1–57.5)	85.0 (0.3–86.7)
Elderly persons	-	16.3‡	-	31.7‡

Treatment

The aim of the treatment of latent tuberculosis infection is the prevention of progression to active clinical disease. Isoniazid administered daily for 6 to 12 months has been the mainstay of treatment, with efficacy ranging from 60 to 90%. Reanalysis and modeling of the U.S. Public Health Service isoniazid trials of the 1950s and 1960s showed that the benefit of isoniazid increases progressively when it is administered for up to 9 or 10 months and stabilizes thereafter. As a consequence, in the absence of controlled, clinical trials comparing isoniazid with placebo, the 9month isoniazid regimen has been recommended as adequate treatment. However, a metaanalysis of 11 isoniazid trials involving 73,375 HIVuninfected persons showed that, as

severe liver injury in HIV uninfected people; thus, it is no longer recommended. In a multicenter, randomized clinical trial, a regimen of daily rifampin for 4 months was associated with fewer serious adverse events and better adherence and was more costeffective than a 9month regimen of isoniazid. Regimens containing rifampin should be considered for persons who are likely to have been exposed to an isoniazid resistant strain of *M. tuberculosis*. In one study, the efficacy of a once weekly, directly observed isoniazid–rifapentine regimen for 3 months was similar to that of a 9month, selfadministered regimen of isoniazid alone and was associated with higher treatment completion rates (82.1% vs. 69.0%) and less hepatotoxicity (0.4% vs. 2.7%), although permanent

discontinuation of the regimen due to side effects was more frequent with the isoniazid–rifapentine regimen (4.9% vs. 3.7%). Similar results were observed in a study involving 1058 children 2 to 17 years of age; however, hepatotoxic effects attributed to treatment were not observed in either study group. A followup study involving HIV infected persons showed that the 3month isoniazidrifapentine regimen was as effective as the 9month isoniazid regimen and was associated with a higher treatmentcompletion rate (89% vs. 64%).

combination of rifapentine (300 mg) and isoniazid (300 mg) is expected to be marketed soon in tablet form, which will facilitate treatment. The 3month isoniazid–rifapentine regimen may be a costeffective alternative to the 9month isoniazid regimen, particularly if the cost of rifapentine decreases and the treatment is selfadministered.

Table 2 Regimens for Latent Tuberculosis Treatment, According to Pooled Efficacy, Risk of Hepatotoxicity, Adverse Events, and Drug Interactions

Drug Regimen	Dosage	Efficacy vs. Placebo*	Efficacy vs. 6 Mo of Isoniazid*	Hepatotoxicity vs. 6 Mo of Isoniazid*	Adverse Events
		odds ratio (95% confidence interval)			
Isoniazid alone for 6 mo or 9 mo	Adults, 5 mg/kg; children, 10 mg/kg (maximum, 300 mg)	6-mo regimen, 0.61 (0.48–0.77); 9-mo regimen, 0.39 (0.19–0.83)	Not applicable for 6-mo regimen, and not available for 9-mo	Not applicable for 6-mo regimen, and not available for 9-mo regimen	Drug-induced liver injury, regimen nausea, vomiting, abdominal pain, rash, peripheral neuropathy, dizziness, drowsiness, and seizure
Rifampin alone for 3 to 4 mo	Adults, 10 mg/kg; children, 10 mg/kg (maximum if <45 kg, 450 mg; maximum if ≥45 kg, 600 mg)	0.48 (0.26–0.87)	0.78 (0.41–1.46)	0.03 (0.00–0.48)	Influenza-like syndrome, rash, drug-induced liver injury, anorexia, nausea, abdominal pain, neutropenia, thrombocytopenia, and renal reactions (e.g., acute tubular necrosis and interstitial nephritis)
Isoniazid plus rifampin for 3 to 4 mo	Adults, 10 mg/kg; children, 10 mg/kg (maximum if <45 kg, 450 mg; maximum if ≥45 kg, 600 mg)	0.52 (0.33–0.84)	0.89 (0.65–1.23)	0.89 (0.52–1.55)	Influenza-like syndrome, rash, drug-induced liver injury, anorexia, nausea, abdominal pain, neutropenia, thrombocytopenia, and renal reactions (e.g., acute tubular necrosis and interstitial nephritis)
Weekly rifapentine plus isoniazid for 3 mo	Adults and children: rifapentine, 15–30 mg/kg (maximum, 900 mg)‡; isoniazid, 15 mg/kg (maximum, 900 mg)	Not available	0.44 (0.18–1.07)§	0.16 (0.10–0.27)§	Hypersensitivity reactions, petechial rash, drug-induced liver injury, anorexia, nausea, abdominal pain, and hypotensive reactions

Table 3 Common Drugs That Could Interact with the Regimen

Antiretroviral Agents	Opioids and Immunosuppressants	Other
Efavirenz (efavirenz levels may increase in slow metabolizers of both drugs)	None	Carbamazepine, benzodiazepines metabolized by oxidation (e.g., triazolam), acetaminophen, valproate, serotonergic antidepressants, disulfiram, warfarin, and theophylline
Efavirenz†; dolutegravir (dolutegravir dose should be increased to 50 mg every 12 hr); rifampin should not be administered with any protease inhibitor (regardless of ritonavir boosting), rilpivirine, elvitegravir, or maraviroc	Methadone (methadone dosage may need to be increased 50%); cyclosporine; glucocorticoids	Mefloquine, azole antifungal agents, clarithromycin, erythromycin, doxycycline, atovaquone, chloramphenicol, hormone-replacement therapy, warfarin, cyclosporine, glucocorticoids, anticonvulsant drugs, cardiovascular agents (e.g., digoxin), theophylline, sulfonylurea hypoglycemic agents, hypolipidemic agents, nortriptyline, haloperidol, quetiapine, benzodiazepines, zolpidem, and buspirone
Efavirenz†; dolutegravir (dolutegravir dose should be increased to 50 mg every 12 hr); isoniazid plus rifampin should not be administered with any protease inhibitor (regardless of ritonavir boosting), rilpivirine, elvitegravir, or maraviroc	Methadone (methadone dosage may need to be increased 50%); cyclosporine; glucocorticoids	Mefloquine, azole antifungal agents, clarithromycin, erythromycin, doxycycline, atovaquone, chloramphenicol, hormone-replacement therapy, warfarin, cyclosporine, glucocorticoids, anticonvulsant drugs, cardiovascular agents (e.g., digoxin), theophylline, sulfonylurea hypoglycemic agents, hypolipidemic agents, nortriptyline, haloperidol, quetiapine, benzodiazepines, zolpidem, and buspirone
Rifapentine plus isoniazid should not be administered with any protease inhibitors, any integrase inhibitors, or maraviroc; early studies show nonsignificant interactions with dolutegravir, emtricitabine, and tenofovir	Methadone (methadone dosage may need to be increased 50%); cyclosporine; glucocorticoids	Mefloquine, azole antifungal agents, clarithromycin, erythromycin, doxycycline, atovaquone, chloramphenicol, hormone-replacement therapy, warfarin, cyclosporine, glucocorticoids, anticonvulsant drugs, cardiovascular agents (e.g., digoxin), theophylline, sulfonylurea hypoglycemic agents, hypolipidemic agents, nortriptyline, haloperidol, quetiapine, benzodiazepines, zolpidem, and buspirone

The weekly isoniazid rifapentine regimen was also evaluated in 1148 South African adults who had HIV infection and a positive tuberculin skin test and were not receiving antiretroviral therapy; the efficacy of that regimen was shown to be similar to a 6month isoniazid regimen. Recent studies of interactions between rifapentine, with or without isoniazid, and efavirenz showed that coadministration of efavirenz for the treatment of HIV infection did not result in reduced efavirenz exposure that could jeopardize antiviral activity. A fixeddose

Currently, the 3month isoniazid– rifapentine regimen is not recommended for children younger than 2 years of age, persons with HIV infection who are receiving antiretroviral therapy, and women who are pregnant. A few small studies have explored treatment of latent tuberculosis infection in contacts (both children and adults) of persons with multidrugresistant tuberculosis on the basis of the results of drugsusceptibility testing of the source patient. However, evidence is lacking on the best treatment approach. Rather, strict observation and

monitoring for at least 2 years for the development of active tuberculosis disease are the preferred clinical measures.

Clinical Evaluation and Monitoring

The clinical management of latent tuberculosis infection starts with tuberculin skin testing, IGRAs, or both and careful clinical and radiologic evaluation to rule out active tuberculosis disease. Persons receiving treatment should be educated about the potential toxic effects of the medications and counseled to stop treatment and seek attention if signs or symptoms such as jaundice, abdominal pain, severe nausea, or fever develop. Hepatotoxicity and clinical hepatitis are serious adverse events associated with drugs that are currently used for the treatment of tuberculosis (Table 2). Unfortunately, there is a paucity of data on the role of baseline tests and the reasonable frequency of visits to monitor adverse events. The role of the tests and the frequency of visits should be defined on the basis of the clinical indications and social profile of the person being treated, as well as the capacity of clinical services. Initial screening with liverfunction tests and regular measurement of liver function afterward could facilitate clinical management. Persons with underlying liver disease, those receiving antiretroviral therapy, women who are pregnant or post partum, alcohol abusers, or persons who are receiving longterm treatment with potentially hepatotoxic medications should be given priority for regular liverenzyme monitoring. Clinical management of latent tuberculosis infection should also address such concomitant risk factors as illicitdrug use, alcohol abuse, and smoking through opioidsubstitution treatment and counseling about alcohol and smoking cessation, respectively. Table 2 summarizes the common drug interactions associated with latent tuberculosis infection treatment that warrant attention. Acceptance of and adherence to the full course of latent tuberculosis treatment must be encouraged. In a study conducted in the United States and Canada, 17% of persons who were offered treatment for latent infection refused it. Treatment completion varies widely (from 19% to 96%), and the reasons for noncompletion need to be fully assessed. The use of various incentives to promote treatment initiation and adherence, depending on the specific need of the person being treated, should be considered. Peer education, counseling, peoplefriendly services, and properly trained service providers boost confidence and may improve adherence to treatment.

Logistic Regression: Model and Notation

In logistic regression, a single outcome variable Yi ($i = 1, \ldots, n$) follows a Bernoulli probability function that takes on the value 1 with probability πi and 0 with probability $1 - \pi i$. Then πi varies over the observations as an inverse logistic function of a vector $\mathbf{x}i$, which includes a constant and $k - 1$ explanatory variables:

$Yi \sim \text{Bernoulli}\ (Y_i\ /\pi_i)$

$$\pi_i = \frac{1}{1+e^{-x_i\beta}} \quad (1)$$

The Bernoulli has probability function $P(Yi \mid \pi i\ ) = \pi_i^{Yi}(1\pi_i)^{1Yi}$. The unknown parameter $\beta = (\beta_0, \beta_1')'$ is a $k \times 1$ vector, where β_0 is a scalar constant term and $\boldsymbol{\beta}_1$ is a vector with elements corresponding to the explanatory variables.

An alternative way to define the same model is by imagining an unobserved continuous variable Y_i^* (e.g., health of an individual or propensity of a country to go to war) distributed according to a logistic density with mean μ_i . Then μ_i varies over the observations as a linear function of $\mathbf{x}i$. The model would be very close to a linear regression if Y_i^* were observed:

$$Y_i^* \sim logistic\ (Y_i\ /\pi_i\ )$$

$$\mu i\ = x_i\beta \quad (2)$$

where Logistic($Y_i^*\ |\mu_i\ $) is the oneparameter logistic probability density,

$$p\ (Y_i^*\ ) = \frac{e^{-(Y_i^*-\mu_i)}}{(1+e^{-(Y_i^*-\mu_i)})^2} \quad (3)$$

Unfortunately, instead of observing Y_i^*, we see only its dichotomous realization, Yi, where $Yi = 1$ if $Y_i^* > 0$ and $Yi = 0$ if $Y_i^* \leq 0$. For example, if Y_i^* measures health, Yi might be dead (1) or alive (0). If Y_i^* were the propensity to go to war, Yi could be at war (1) or at peace (0). The model remains the same because

$$\Pr(Yi = 1 \mid \boldsymbol{\beta}) = \pi i = \Pr(Y_i^* i > 0 \mid \boldsymbol{\beta})$$
$$= \int_0^\infty \text{Logistic}\ (Y_i^* / \pi i\ )\ d\ (Y_i^*\ ) = \frac{1}{1+e^{-x_i\beta}} \quad (4)$$

which is exactly as in Eq. (1). We also know that the observation mechanism, which turns the continuous Y^* into the dichotomous Yi , generates most of the mischief. That is, we ran simulations trying to estimate $\boldsymbol{\beta}$ from an observed Y^* and model 2 and found that maximumlikelihood estimation of $\boldsymbol{\beta}$ is approximately unbiased in small samples. The parameters are estimated by maximum likelihood, with the likelihood function formed by assuming independence over the observations: $L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^n \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$

By taking logs and using Eq. (1), the loglikelihood simplifies to

$$\ln L(\boldsymbol{\beta} \mid \mathbf{y}) = \sum_{\{y_i=1\}} \ln(\pi_i) + \sum_{\{y_i=0\}} \ln(1-\pi_i)$$
$$= \sum_{i=1}^n \ln(1 + e^{(1-2y_i)X_i\beta}) \quad (5)$$

Maximumlikelihood logit analysis then works by finding the value of $\boldsymbol{\beta}$ that gives the maximum value of this function, which we label $\widehat{\boldsymbol{\beta}}$. The asymptotic variance matrix, $\boldsymbol{V}(\widehat{\boldsymbol{\beta}})$, is also retained to compute standard errors. When observations are selected randomly, or randomly within strata defined by some or all of the explanatory variables, $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically efficient (except in degenerate cases of perfect collinearity among the columns in $\mathbf{X}$ or perfect discrimination between zeros and ones). That in rare events data ones are more statistically informative than zeros can be seen by studying the variance matrix,

$$V(\hat{\beta}) = [\sum_{i=1}^n \pi_i(1-\pi_i)x_i' x_i]^{-1} \quad (6)$$

The part of this matrix affected by rare events is the factor $\pi_i(1 - \pi_i)$. Most rare events applications yield small estimates of $\Pr(Y_i = 1 \mid \mathbf{x}_i) = \pi_i$ for all observations. However, if the logit model has some explanatory power, the estimate of πi among observations for which rare events are observed (i.e., for which $Y_i = 1$) will usually be larger [and closer to 0.5, because probabilities in rare event studies are normally very small refer to Beck *et al*. 2000)] than among observations for which $Yi = 0$. The result is that $\pi i\ (1-\pi i)$ will usually be larger for ones than zeros, and so the variance (its inverse) will be smaller. In this situation, additional ones will cause the variance to drop more and hence are more informative than additional zeros (refer to

Imbens (1992), Cosslett (1981a); Lancaster and Imbens (1996b). Finally, we note that the quantity of interest in logistic regression is rarely the raw $\widehat{\boldsymbol{\beta}}$ output by most computer programs. Instead, scholars are normally interested in more direct functions of the probabilities. For example, *absolute risk* is the probability that an event occurs given chosen values of the explanatory variables, $\Pr(Y = 1 \mid X = x)$. The *relative risk* is the same probability relative to the probability of an event given some baseline values of X, e.g., $\Pr(Y=1 \mid X=1)/\Pr(Y=1 \mid X=0)$, the fractional increase in the risk. This quantity is frequently reported in the popular media (e.g., the probability of getting some forms of cancer increase by 50% if one stops exercising) and is common in many scholarly literatures. In political science, the term is not often used, but the measure is usually computed directly or studied implicitly. Also of considerable interest is the *first difference* (or attributable risk), the change in probability as a function of a change in a covariate, such as $\Pr(Y=1 \mid X=1) - \Pr(Y=1 \mid X=0)$. The first difference is usually most informative when measuring effects, whereas relative risk is dimensionless and so tends to be easier to compare across applications or time periods.

Selection of dependent variable and Data collection strategies

When one of the values of Y is rare in the population, considerable resources in data collection can be saved by randomly selecting within categories of Y . This is known in econometrics as *choicebased* or *endogenous stratified* sampling and in epidemiology as a *casecontrol* design (Breslow 1996); The casecohort study is especially appropriate when adding an expensive variable to an existing collection, such as the dyadic data discussed above and analyzed below, or Verba and coworkers' (1995) detailed study of activists, each of which was culled from a larger random sample, with very few variables, of the entire U.S. population. In this paper, the author use information on the population fraction of ones when it is available, and so the same models we describe apply to both casecontrol and casecohort studies.

Selecting on the dependent variable in the way we suggest has several pitfalls that should be carefully avoided. First, the sampling design for which the prior correction and weighting methods are appropriate requires independent random (or complete) selection of observations for which $Y = 1$ and $Y = 0$. This encompasses the casecontrol and casecohort studies, but other endogenous designs such as sampling in several stages, with nonrandom selection, or via hybrid approaches require different statistical methods. Second, when selecting on Y , we must be careful not to select on X differently for the two samples. The classic example is selecting all people in the local hospital with liver cancer ($Y = 1$) and a random selection of the U.S. population without liver cancer ($Y = 0$). The problem is that the sample of cancer patients selects on $Y = 1$ *and* implicitly on the inclination to seek health care, find the right medical specialist, have the right tests, etc. Not recognizing the implicit selection on X is the problem here. Since the $Y = 0$ sample does not similarly select on the same explanatory variables, these data would induce selection bias. One solution in this example might be to select the $Y = 0$ sample from those who received the same liver cancer test but turned out not to have the disease. This design would yield valid inferences, albeit only for the health conscious population with liver cancerlike symptoms. Another solution would be to measure and control for the omitted variables. This type of inadvertent selection on X can be a serious problem in endogenous designs, just as selection on Y can bias inferences in exogenous designs. Moreover, although in the social sciences random (or experimenter control over) assignment of the values of the explanatory variables for each unit is occasionally possible in exogenous or random sampling (and with a large n is generally desirable since it rules out omitted variable bias), random assignment on X is impossible in endogenous sampling. Fortunately, bias due to selection on X is much easier to avoid in applications such as international conflict and related fields, since a clearly designated census of cases is normally available from which to draw a sample. Instead of relying on the decisions of subjects about whether to come to a hospital and take a test, the selection into the data set in our field can often be entirely determined by the investigator. Refer to Holland and Rubin (1988). Third, another problem with intentional selection on Y is that valid exploratory data analysis can be more hazardous. In particular, one cannot use an explanatory variable as a dependent variable in an auxiliary analysis without special precautions (see Nagelkerke *et al*. 1995). Finally, the optimal tradeoff between collecting more observations versus better or more explanatory variables is application specific, and so decisions will necessarily involve judgment calls and qualitative assessments. Fortunately, to help guide these decisions in fields like international relations we have large bodies of work on methods of quantitative measurement and, also, many qualitative studies that measure hardtocollect variables for a small number of cases.

Prior Correction

Prior correction involves computing the usual logistic regression MLE and correcting the estimates based on prior information about the fraction of ones in the population, τ, and the observed fraction of ones in the sample (or sampling probability), $\bar{y}$. Knowledge of τ can come from census data, a random sample from the population measuring Y only, a casecohort sample, or other sources.

Prior correction requires knowledge of the fraction of ones in the population, τ. Fortunately, τ is straightforward to determine in international conflict data since the number of conflicts is the subject of the study and the denominator, the population of countries or dyads, is easy to count even if not entirely in the analysis.4 A key advantage of prior correction is ease of use. Any statistical software that can estimate logit coefficients can be used, and Eq. (7) is easy to apply to the intercept. If the functional form and explanatory variables are correct, estimates are consistent and asymptotically efficient. The chief disadvantage of prior correction is that if the model is misspecified, estimates of both $\beta 0$ and $\boldsymbol{\beta}1$ are slightly less robust than weighting refer to Xie and Manski (1989), a method to which we now turn.

Weighting

An alternative procedure is to weight the data to compensate for differences in the sample ($\bar{y}$) and population (τ) fractions of ones induced by choice based sampling. The resulting *weighted exogenous sampling maximum likelihood estimator* (due to Manski and Lerman (1977) is relatively simple. Instead of maximizing the loglikelihood in Eq. (5), we maximize the weighted loglikelihood:

$$\ln L_w(\boldsymbol{\beta} \mid \mathbf{y}) = \omega_1 \sum_{\{y_i=1\}} \ln(\pi_i) + \omega_0 \sum_{\{y_i=0\}} \ln(1-\pi_i)$$
$$= -\sum_{i=1}^{n} \omega_i \ln(1 + e^{(1-2y_i)X_i\beta}) \quad (8)$$

where the weights are $\omega_i = \tau/\bar{y}$ and $\omega_0 = (1-\tau)/(1-\bar{y})$, and where

$$= \omega_1 Y_i + \omega_0 (1 - Y_i) \quad (9)$$

One perceived disadvantage of this model has been that it seemed to require specialized software for estimation. However, the alternative expression in the second line of Eq. (8) enables researchers to use any log it package, since the weight, ω_1, appears in one term. All researchers need to do is to calculate ω_1 in Eq. (8),

Rare Event and Finite Sample Corrections

Let **x**0 be a 1 × k vector of chosen values of the explanatory variables. The nearly universal method used for computing the probability, given $\mathbf{x}_0$, is a function of the maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}$,

$$\Pr(Y_0 = 1 \mid \widehat{\boldsymbol{\beta}}) = \hat{\pi}_0 = \frac{1}{1+e^{-X_0\widehat{\beta}}} \quad (10)$$

and is thus statistically consistent.

Unfortunately, the method of computing probabilities given in Eq. (10) is affected by two distinct problems in finite samples of rare events data: First, $\widehat{\boldsymbol{\beta}}$ is a biased estimate of **β**. Second, even if $\widehat{\boldsymbol{\beta}}$ were unbiased, Pr (Y0 = 1 | $\widehat{\boldsymbol{\beta}}$) would still be, as we show below, an inferior estimator of $\Pr(Y_0 = 1 \mid \boldsymbol{\beta})$.

Estimation

The bias in $\hat{\beta}$ can be estimated by the following weighted least squares expression:

$$\text{bias}(\hat{\beta}) = (X' WX)^{-1} X' W\xi \quad (11)$$

where $\xi_i = 0.5Q_{ii}\,[(1+w_1)\hat{\pi}_i - w_1]$, Q_{ii} are the diagonal elements of Q = X(X_WX)−1X_, and W = diag{$\hat{\pi}_i$ $(1-\hat{\pi}_i)w_i$ }. This expression is easy to estimate, as it involves running a weighted least squares regression with X as the "explanatory variables," ξ as the "dependent variable," and Was the weight. The biascorrected estimate is then $\tilde{\beta}$ = $\hat{\beta}$-bias ($\hat{\beta}$).

The special case with a constant term and one explanatory variable, and with β_0 estimated and $\beta_1 = 1$ fixed: $\Pr(Y_i = 1) = 1/(1 + e^{-(\beta 0 + Xi)})$. For this case,

bias in $\hat{\beta}_0$, where $\bar{\pi} = (1/n)\sum_{i=1}^{n} \pi_i$, as

$$E(\hat{\beta}_0 - \beta_0) \approx \frac{\bar{\pi} - 0.5}{n\bar{\pi}(1-\bar{\pi})} \quad (12)$$

Since $\bar{\pi} < 0.5$ in rare events data, the numerator, and thus the entire bias term, is negative.

This means that $\hat{\beta}_0$ is too small and, as a result, Pr(Y = 1) is underestimated, which is consistent with what we argued intuitively above and show via Monte Carlo experiments below.

Probability Calculations

This section concerns estimating the probability π in Eq. (1). Since $\tilde{\beta}$ is less biased and has smaller variance, and hence has a smaller mean square error, than $\hat{\beta}$,

$$\tilde{\pi}_0 = \Pr(Y_0 = 1 \mid \tilde{\beta}) = \frac{1}{1+e^{x_0\tilde{\beta}}} \quad (13)$$

is usually preferable to $\hat{\pi}$ [from Eq. (10)]. However, $\tilde{\pi}$ is still not optimal because it ignores the uncertainty in $\tilde{\beta}$ (e.g., Geisser 1993; King *et al*. 2000). This uncertainty can be thought of as sampling error or the fact that $\tilde{\beta}$ is estimated rather than known, and it is reflected in standard errors greater than zero. In many cases, ignoring estimation uncertainty leaves the point estimate unaffected and changes only its standard error. However, because of the nature of π as a quantity to be estimated, ignoring uncertainty affects the point estimate too. Thus, instead of conditioning on an uncertain point estimate with $\tilde{\pi}$, we should be conditioning only on known facts and averaging over the uncertainty in ˜β as follows:

$$\Pr(Yi = 1) = \int \Pr(Yi = 1 \mid \beta^*)p(\beta^*)P(\beta^*)\, d\beta^* \quad (14)$$

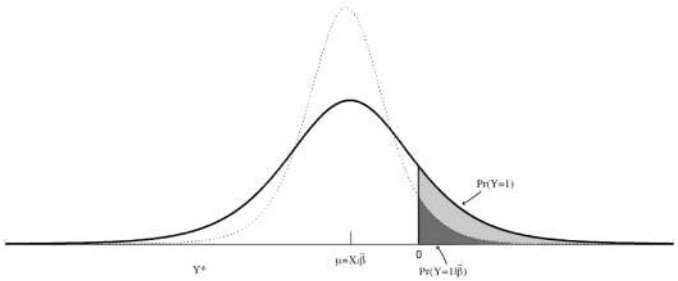


Fig 1 The effect of uncertainty on probabilities. Although the dotted density (which does not reflect uncertainty in β) has a smaller variance than the one drawn with a solid line (which has the uncertainty about β added in), the mean μ stays the same in both. However, the probability, the shaded area to the right of the zero threshold in the two curves, differs.

where β^* is the integration dummy, and to summarize estimation uncertainty P(·) we take the Bayesian viewpoint and use the posterior density of β Normal [β |$\tilde{\beta}$, V($\tilde{\beta}$)] (although it will turn out that we will not need this normality assumption). The estimation uncertainty P(·) can also be thought of from a frequentist perspective as the sampling distribution of ˜β so that Eq. (14) is the expected value E $\tilde{\beta}$ [Pr(Yi = 1 | $\tilde{\beta}$)], which is an estimate of $\pi_i = \Pr(Yi = 1 \mid \beta) = 1/(1 + e^{-xi\beta})$.

Equation (14) can be computed in two ways. First, we could use simulation (see Tanner 1996; King *et al*. 2000): take a random draw of β from P(β), insert it into $[1 + e^{-xi\beta}]^{-1}$, repeat, and average over the simulations. Increasing the number of simulations enables us to approximate Pr (Yi = 1) to any desired degree of accuracy.

A second method of computing Eq. (14) is through an analytical approximation we have derived. It is more computationally efficient than the simulation approach, is easy to use, and helps illuminate the nature of the correction. This result, proven in Appendix E, shows that Eq. (14) may be approximated without simulation as

$$\Pr(Yi = 1) \approx \tilde{\pi}i + C_i \quad (15)$$

where the correction factor is

$$Ci = (0.5 - \tilde{\pi}i)\,\tilde{\pi}i\,(1 - \tilde{\pi}i)x_0V(\tilde{\beta})x_0' \quad (16)$$

Standard errors or confidence intervals can easily be computed as part of the simulation in the first approach or by simulating each component of Ci in the second.

Analyses

We first generated n observations from a logistic regression model with a constant and one explanatory variable drawn from a standard normal density, for fixed parameters $\beta 0$ and $\beta 1 = 1$. For each i, we drew a random uniform number u and assigned $Yi = 1$ if $\pi i < u$ and $Yi = 0$ otherwise. We set the sample size to

$n = \{150, 250, 550, 1100, 2100, 3100, 4100, 5100, 12{,}000, 22{,}000\}$

and intercept to

$\beta_0 = \{-7, -6, -5, -4, -3, -2, -1, 1\}$

These values of $\boldsymbol{\beta}$ generate **y** vectors with the percentages of ones equaling $(100 \times \bar{y})\% = \{0.15, 0.4, 1.1, 2.8, 6.9, 15.6, 30.4, 50\}$ respectively. We excluded experiments with both very small percentages of ones and small sample sizes so as to avoid generating **y** vectors that are all zeros. This mirrors the common practice of studying rarer events in larger data sets. For each of these experiments, we computed the maximum difference in absolute risk by first taking the difference in estimates of $\Pr(Y = 1 \mid X = x)$ between the traditional logit model and our preferred approximate Bayesian method, for each of 31 values of x, equally spaced between −5 and 5, and then selecting the maximum. We also computed one relative risk, where we changed X from −1 to 1: $\Pr(Y = 1 \mid X = 1)/\Pr(Y = 1 \mid X = -1)$. The pair of X values, $\{-1, 1\}$, defines a typical relative risk that might be computed in examples like this, since it is at plus and minus one standard deviation of the mean of X, but it is of course neither the maximum nor the minimum difference in relative risk that could be computed between the two methods. Finally, for each Monte Carlo experiment, we computed the maximum absolute risk and the relative risk averaged over 1000 simulated data sets. We have repeated this design with numerous other values of n, $\beta 0$, and $\beta 1$, and explanatory variables in different numbers and drawn from different (including asymmetric and partially discrete) densities. We also computed different absolute and relative risks. These other experiments led to similar conclusions as those presented here.

The properties of the coefficients and standard errors of logistic regression with and without our corrections, and for both cohort and casecontrol designs. With $\beta 0 = -4$ (i.e., about 2.8% ones) and $n = 1000$, and then successively drop {0, 0.225, 0.45, 0.675, 0.9} fractions of observations with zeros. Since it has been well studied by Xie and Manski (1989), and are shown in fig.2.

The correction bias in standard errors, RMSE in probability estimates full sample, Bias in probability estimates full sample and RMSE of probability estimates Subsampled data and RMSE of relative risk estimates: subsampled data. For the simulated data are show in Fig. 3, Fig.4, Fig.4, Fig.5, Fig.6.

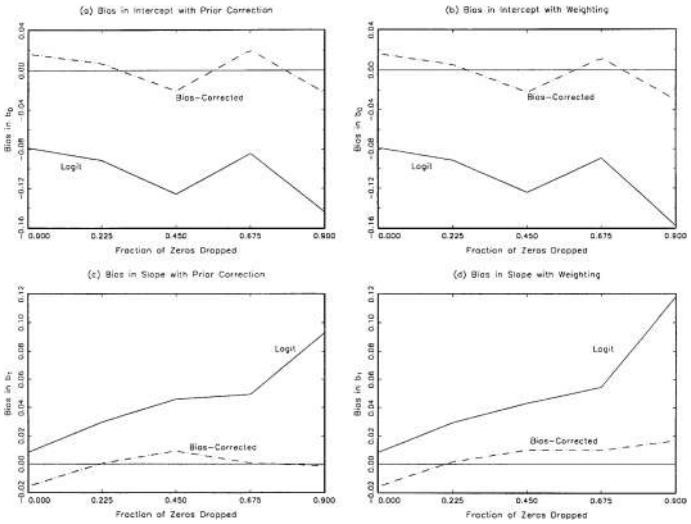


Fig 2 Correcting bias in logit coefficients.

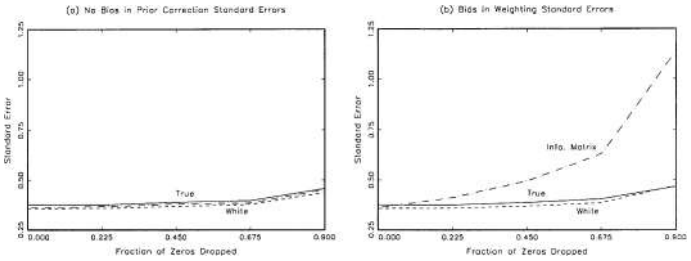
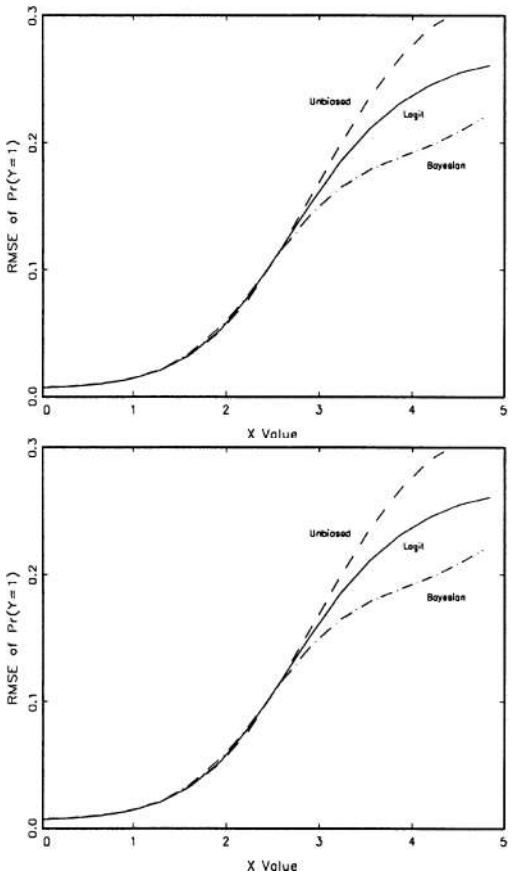


Fig 3 Correcting bias in standard errors



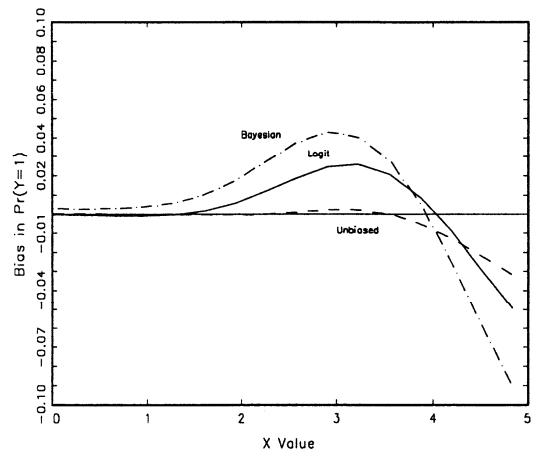


Fig 4 RMSE in probability estimates: full sample, Bias in probability estimates: full sample

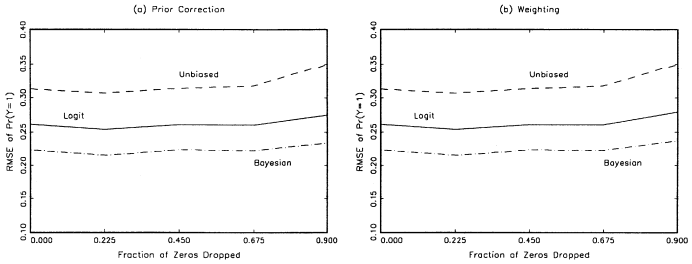


Fig 5 RMSE of probability estimates: subsampled data.

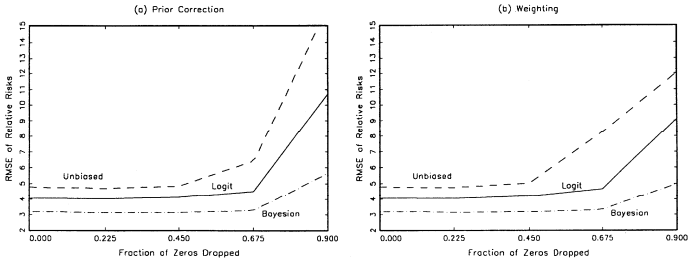


Fig 6 RMSE of relative risk estimates: subsampled data.

Binary Logistic Regression Approach

Logistic regression is an alternative to discriminant analysis due to several reasons, whenever the dependent variable has only two categories. The first reason is that logistic regression is less affected by the variance/ covariance inequalities across the groups. Secondly logistic regression can handle categorical independent variables easily, whereas in discrimnant analysis the use of dummy variables can create problems. Finally logistic regression results are parallel to those of multiple regression in terms of their interpretation. In logistic regression there are no assumptions such as multivariate normality, equal variance covariance matrices. In discriminant analysis the non metric character of a dichotomous dependent variable is accommodate by making predictions of group membership based on discriminant Z scores. This requires the calculations of cutting scores and the assignment of observation to groups. Logistic regression approaches this task in a manner more similar to that found in multiple regression. It differs from multiple regression in the sense that it directly predicts the probability of an event occurring. To define a relationship bounded by zero and one. Logistic regression uses an assumed relationship between the independent and dependent variables that resembles a Sshaped curve.

Let y be a cluster n binary observations y_j (j = 1, ..., n) with x_j being a p × 1 covariate. Denote by y. the sum of the y_j's. Rosner (1984) proposed a polychotomous logistic regression model

$$pr(y/x_1,...,x_n) = c(\theta,\beta)\exp\left[\sum_{k=0}^{y-1}\log\left\{\frac{\theta_1+k\theta_2}{1-\theta_1+(n-1-k)\theta_2}\right\}+\sum_{j=1}^{n}\beta' x_j y_j\right], \quad \ldots (1)$$

where c is the normalizing constant which involves a sum of 2^n exponential terms. It follows from equ. (1) that the logit conditional probability of $y_j = 1$ given $y_j = (y_1,...,y_{j1}, y_{j+1},...,y_n)$ and x_j is such that

$$\text{logit}\,\text{pr}(y_j = 1/y_{-j}x_j) = \log\left\{\frac{\theta_1+w_j\theta_2}{1-\theta_1+(n-1-w_j)\theta_2}\right\}+\beta'' x_j, \quad \ldots (2)$$

where $w_j = y.y_j$ is the sum of the y's excluding y_j. In other words, the conditional probability for $y_j = 1$ in eqn. (1.2) depends on y_j only through the sum. The equ. (1.2) due to Qu *et al*. (1987) who point out that $\theta_2/(1+\theta_2)$, the common intraclass correlation when all the x_j's = 0, may be negative.

In logistic regression procedure, the logistic coefficient compares the probability of an event occurring with the probability of its not occurring. This odds ratio is given as

$$\text{Odds Ratio} = \frac{\text{Probability of event occuring}}{\text{Probability of the event not occuring}} = e^{B_0+B_1X_1+\cdots+B_nX_n}$$

Here X_1, X_2, …, X_n are in the independent or influencing variables and β_0 is the constant intercept β_1, β_2, … , β_n are the estimated regression coefficients and they are measures of the changes in the ratio of the probabilities.

Numerical Illustration

The data were collected from 60 TB infected patients during Jan 2016 to Dec 2016 from the list of Revised National Tuberculosis Control Programme (RNTCP), Dharmapuri under the areas namely Harur, Paperiratipatti and Pennagram. Another 60 noninfected persons were enrolled for this study from the same region through the outpatient register of Government Medical College, Dharmapuri. Using this data set we fit binary logistic regression. In this present study the probability or chance of an individual to getting TB infection is computed as a function of a number of independent or influencing variables. The variables which are taken up under this model are follows

Y = the dependent variable representing the desire or preparedness to leave the organization (0-yes, 1-no)
X_1 = age of the patient (0-≤ 30, 1->30)
X_2 = gender (0-male, 1-female)
X_3 = marital status (0-unmarried, 1-evermarried)
X_4 = type of family (0-nuclear, 1-joint)
X_5 = education status (0-illiterate, 1-literate)
X_6 = occupation of the patient (0-Sedentary, 1-nonsedentary)
X_7 = type of house (0-kutcha, 1-pucca)
X_8 = family members (0-≤ 4, 1-> 4)
X_9 = percapita income (0-≤ 4500, 1-> 4500)
X_{10} = body mass index (0-normal, 1-abnormal)
X_{11} = smoker (0-yes, 1-no)

X_{12} = alcoholic (0-yes, 1-no)

The variables which have an operative influence over the infection should have a significant partial regression coefficient. Such of these variables are alone retained in the model and all those variables which have significant regression coefficients may be classified as i) the demographic variables like age, sex, family type. etc, ii) the other variables which are based on psychological, environmental and other considerations are relevant from the view point of our study. The demographical variables are called the non control variables in the sense that their influence over the decision to leave the organization cannot be changed by intervention and strategies of retention. But in the case of certain variables like smoking habit, consumption of alcohol etc. some suitable remedial measures can be incorporated by the personnel. So that their risk over the infection of TB can be reduced and mended to a favorable status with the result that the intensity of the feeling to get infection can be changed and brought down. Hence such variables which are amenable to adjustment are called control variables.

An advantage of the logistic regression approach is that, it helps as a prediction equation. In the present model the dependent variable namely Y = 0, if a particular individual is not prepared to leave the organization, where as if Y = 1, it indicates the person's willingness to leave the organization. Also the independent variables which exercise significant influence over the Y value can be identified and isolated. Once the logistic regression equation is obtained on the basis of the data collected from the sample of respondents chosen for study, the mathematical form of the regression equation can be formulated. Then the same questionnaire can be given to a set of individuals or for all the personnel working in that organization. Collecting the relevant data from the individuals and using the same in the regression equation it is possible to identify, the chance of getting infection of TB.

Table 1 Omnibus Tests of Model Coefficients

		Chisquare	df	Sig.
Step 1	Step	3.297	12	0.306
	Block	3.297	12	0.306
	Model	3.297	12	0.306

Table 2 Model Summary

Step	2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	378.572[a]	.0.724	0.767

Table 3 Hosmer and Lemeshow Test

step	ChiSquare	df	Sig.
1	3.193	12	0.436

Table 4 Variables in the Equation

		β	S.E.	Wald	df	Sig.	Exp(β)
Step 1[a]	X1	0.004	0.011	3.304	1	0.032	1.005
	X2	0.156	0.521	0.059	1	0.593	1.167
	X3	0.202	0.501	3.031	1	0.031	0.705
	X4	0.156	0.786	3.045	1	0.023	0.745
	X5	0.014	0.518	4.002	1	0.065	1.021
	X6	0.323	0.411	3.000	1	0.019	0.290
	X7	0.007	0.511	0.000	1	0.886	0.981
	X8	0.002	0.610	0.000	1	0.898	1.000
	X9	0.146	1.257	0.011	1	0.802	1.123
	X10	0.075	0.299	0.035	1	0.731	1.059
	X11	0.262	0.699	5.012	1	0.001	1.815
	X12	0.298	0.529	5.001	1	0.001	1.881
	Constant	2.130	2.452	4.559	1	0.011	9.130

It is seen that hosmer and lemshow test for finding the goodness of fit of the model for the data, in other words this model fit shows the computed value of chisquare statistic as 3.193 with a corresponding significance value p=0.436. Since p value is greater than 0.05, it suggests that the logistic regression model fitted to the data is a good fit. A good model fit indicated by a non significant chisquare value.

The results on the variables in the equation suggested that the independent variables namely

X_1 = age of the patient
X_3 = marital status
X_4 = type of family
X_6 = occupation of the patient
X_{11} = smoker
X_{12} = alcoholic

have p values less that 0.05 and they are all significant there by implying that they have significant regression coefficients. Hence they all have influence over the chance to getting infection of TB. β_0 is also significant. The regression coefficients for X_3 ($\beta_3 = 0.202$), X_4 ($\beta_4 = 0.156$) and X_6 ($\beta_6 = 0.323$) are negative. Hence it implies these variables are having risk in manner or reversal order. Some of the variables X_2, X_5, X_7, X_8, X_9 and X_{10} are make insignificant contribution with this regression equation and p values of these regression coefficients are more than 0.05.

It may observe that the regression coefficient of the independent variables namely age, marital status, family type, occupational status, smoking habit and consumption of alcohol have influence over the incidence of TB. The other variables do not have significant regression coefficients. It may also be observed that among the influencing variables there are some variables which can be under control. For example smoking habit and consumption of alcohol can be discontinuous which in term will help the avoidance of TB. It is also possible to have the occupational status. The age, family type, marital status cannot be brought under control and hence the preventives strategies can be advocated in this study.

CONCLUSION

Based on the results discussed in the previous session, it is g generally observed that Better understanding of the pathogenesis of latent tuberculosis infection is a critical research priority, as is the development of biomarkers and diagnostic tests with improved performance and predictive values. The availability of new drugs and regimens that can be administered for a shorter duration and with fewer adverse events is imperative to allow largerscale implementation. Trials should be performed to define the benefits and harms of treatment for latent tuberculosis infection in patients with diabetes, in alcohol abusers and tobacco smokers, and in contacts of persons with multidrug resistant tuberculosis. Innovative research synergies between public and private

funders are required to overcome market shortcomings. The development of better diagnostic tests, preventive therapies, and vaccines for tuberculosis will confer enormous public benefit.

Reference

1. Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):1–15. (Preprint at http://GKing.Harvard.Edu.)
2. Cosslett, Stephen R. 1981a. "Maximum Likelihood Estimator for ChoiceBased Samples." *Econometrica* 49(5):12891316.
3. Imbens, Guido. 1992. "An Efficient Method of Moments Estimator for Discrete Choice Models with ChoiceBased Sampling." *Econometrica* 60(5):11871214.
4. Breslow, Norman E. 1996. "Statistics in Epidemiology: The CaseControl Study." *Journal of the American Statistical Association* 91:1428.
5. Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.
6. Nagelkerke, Nico J. D., Stephen Moses, Francis A. Plummer, Robert C. Brunham, and David Fish. 1995. "Logisti Regression in CaseControl Studies: The Effect of Using Independent as Dependent Variables." *Statistics in Medicine* 14:769775.
7. Holland, Paul W., and Donald B. Rubin. 1988. "Causal Inference in Retrospective Studies," *Evaluation Review* 12(3):203231.
8. Xie, Yu, and Charles F. Manski. 1989. "The Logit Model and ResponseBased Samples." *Sociological Methods and Research* 17(3):283302.
9. King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355. (Preprint at http://Gking.harvard.edu.)
10. Greene, William H. 1993. *Econometric Analysis*, 2nd ed. New York: Macmillan.
11. Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. New York: Springer Verlag.
12. Rosner, B. (2000). Fundamentals of Biostatistics. Duxbury Thomson Learning, Australia.

How to cite this article:

Senthilkumar V and Sachithanantham S.2018, A Study on the Incidence of Tuberculosis Using Binary Logistic Regression. *Int J Recent Sci Res.* 9(4), pp. 26159-26169. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.2004

****