Chapter 3

Preliminary Data Analysis

In this chapter we present the preliminary data analysis of pulmonary tuberculosis patients and HIV infection in Zonal TB Center 11, Nakhon Si Thammarat. This analysis includes the demographic, health status and behaviour, and disease characteristics of the subjects selected for the study. The results are presented in three parts. Section 1 summarizes all the variables of interest. In Section 2 the crude associations between the outcome variables and the determinants of interest are presented. In Section 3, these associations are investigated after taking into account the possibility of confounding, using the Mantel-Haenszel method of adjustment for confounding.

Summaries of All Variables of Interest

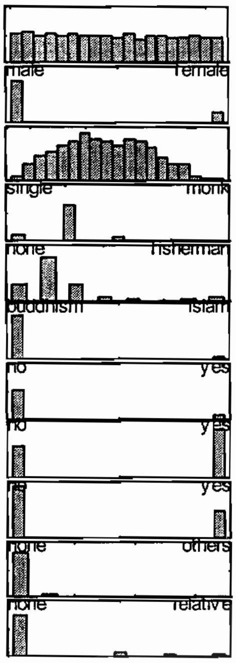
As explained in Chapter 2, the data for this study comprise 1,080 subjects, including 124 cases (11.48%) with HIV and 956 cases (88.52%) without HIV. These consist of 81.02% males and 18.98% females with the mean age of 46 years. Most of them were married (76.52%) while the rest was single (13.87%) and the others (9.61%). The common occupations were agriculture (47.32%) and wage earner (19.91%). The Buddhism was found to be 94.03% and 5.97% in Islam.

For each subject, 22 variables were recorded, including ID (an integer ranging from 1 to 1,080) and HIV status. The list of variables and their values is given in Table 3.1, shown on the next page. For convenience of data analysis using the student version of Matlab, which has limited storage capability, the data are stored in two files. Each file is indexed by ID, includes HIV status, and contains 10 of the 20 remaining variables. Histograms and simple descriptive statistics for the distributions of the raw data, separated according to HIV status into the comparison group (HIV negative) and the study group (HIV positive) are depicted in Figure 3.1 and Figure 3.2, respectively. Each figure has two panels corresponding to the two sets of variables in the two files. The list of variables and data codings is shown in Table 3.1.

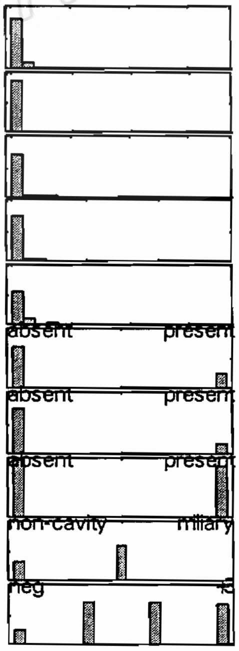
Table 3.1 Variables and data codings

ID	Identification number (1-1080)
Gender	1 = male, 2 = female
Age	years
Marital status	1 = single, 2 = married, 3 = widowed,
Occupation	1 = none, 2 = agriculture, 3 = wage earner,
Religion	1 = Buddhism, 2 = Islam
Receiving BCG	0 = no, 1 = yes
Smoking	0 = no, 1 = yes
Drinking	0 = no, 1 = yes
Other diseases	0 = none, 1 = DM, 2 = peptic ulcer, 3 = VDRL+,
Family history with TB	0 = none, 1 = husband/wife, 2 = parent/daughter,
Cough	The number of days
Haemoptysis	The number of days
Chest pain	The number of days
Dyspnea	The number of days
Fever	The number of days
Weakness	0 = absent, 1 = present
Weight loss	0 = absent, 1 = present
Other symptoms	0 = absent, 1 = present
Chest x-ray finding	0 = non-cavity, 1 = cavity, 2 = miliary
Degree of sputum	0 = none, 1 = positive 1, 2 = positive 2, 3 = positive 3

Figure 3.1 Summaries of all variables in comparison group



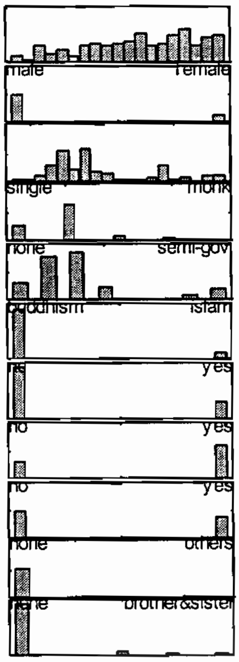
col	variable	size	mean	st dev	min	max
1	ID	956	525.024	312.776	1	1080
2	gender	956	1.191	0.394	1	2
3	age	956	46.95	15.798	13	90
4	marital status	836	2.007	0.597	1	5
5	occupation	834	2.576	1.69	1	8
6	religion	837	1.055	0.228	1	2
7	receiving BCG	622	0.153	0.36	0	1
8	smoking	741	0.603	0.49	0	1
9	drinking	739	0.341	0.474	0	1
10	other diseases	907	0.217	0.939	0	7
11	family hx	895	0.35	0.918	0	4



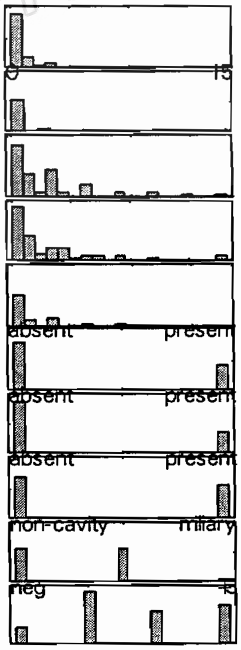
col	variable	size	mean	st dev	min	max
1	cough	955	120.364	209.965	0	3650
2	haemoptysis	932	1.314	9.244	0	180
3	chest pain	872	28.455	63.422	0	730
4	dyspnea	859	72.731	177.162	0	2920
5	fever	729	14.686	36.502	0	365
6	weakness	956	0.251	0.434	0	1
7	weight loss	956	0.186	0.389	0	1
8	other symptoms	954	0.478	0.5	0	1
9	chest x-ray	945	0.666	0.505	0	2
10	degree of sputum	947	1.781	0.975	0	3

TB patients without HIV infection (comparison group)

Figure 3.2 Summaries of all variables in study group



col	variable	size	mean	st dev	min	max
1	ID	124	659.815	278.716	18	1079
2	gender	124	1.177	0.384	1	2
3	age	124	39.339	14.234	15	74
4	marital status	101	1.871	0.702	1	5
5	occupation	100	3.06	1.847	1	8
6	religion	101	1.099	0.3	1	2
7	receiving BCG	64	0.234	0.427	0	1
8	smoking	90	0.667	0.474	0	1
9	drinking	91	0.44	0.499	0	1
10	other diseases	117	0.402	1.445	0	7
11	family hx.	111	0.306	0.882	0	4



col	variable	size	mean	st dev	min	max
1	cough	124	106.008	189.656	0	1825
2	haemoptysis	122	0.32	1.506	0	15
3	chest pain	111	27	36.833	0	180
4	dyspnea	108	50.185	72.046	0	365
5	fever	90	20.111	34.27	0	180
6	weakness	124	0.347	0.478	0	1
7	weight loss	124	0.29	0.456	0	1
8	other symptoms	124	0.452	0.5	0	1
9	chest x-ray	120	0.525	0.534	0	2
10	degree of sputum	123	1.659	1.015	0	3

TB patients with HIV infection (study group)

For some of the variables of interest, including ID and the categorical variables (marital status, occupation etc.), the means and standard deviations are not particularly meaningful. However, these statistics are included for completeness, and are useful for data checking purposes.

The Figures 3.1 and 3.2, the histograms show the distributions of each variable in both groups. It can be seen that the histograms in each column have slightly skewed distributions. The descriptive statistics give the sample size for each variable, mean, standard deviation, and minimum and maximum values. As seen in the upper panels, the mean age of TB patients without HIV infection (the comparison group) was 46.95 years, while that for TB patients with HIV infection (the study group) was 39.34 years. This indicates that TB patients with HIV infection may be younger than other TB patients.

In addition, when comparing the distributions of disease characteristics between the comparison group and the study group, as shown in the bottom panels, the mean values of duration (number of days) of cough, haemoptysis, chest pain and dyspnea in the study group are lower than those in the comparison group, except for fever symptoms. Also, the prevalence of weakness and weight loss for the study group (35% and 29%, respectively) are each 6% higher than those for the comparison group (25% and 19%, respectively).

Associations between the Outcome and the Determinants of Interest

Since the outcome variable (HIV status) is binary and many of the determinants of interest are categorical variables, odds ratios are appropriate for assessing the association between the outcome and these determinants. Moreover, the continuous determinants (age and the duration of the symptoms cough, haemoptysis, chest pain, dyspnea and fever) can be grouped without substantial loss of information, so odds ratios may also be used for their analysis.

Due to small cell counts for some categories of marital status, “widowed”, “divorced/separated” and “monk” were combined with the “others” category. For the same reason, the occupations “none”, “government”, “semi-government” and “fisherman” were combined with the “others” category. Similarly, the presence or absence of other diseases and family history with TB, the presence of all diseases were grouped into the “yes” category. The family history with TB, “husband/wife”, “parent”, “brother/sister” and “relative” were also united into a single category labeled “present”.

The crude odds ratios with 95% confidence intervals and corresponding p-values for testing the hypothesis of no association between HIV status and each demographic determinant are shown in Table 3.2. As explained in Chapter 2, these odds ratios are calculated by comparing each specified category with all other categories combined.

Table 3.2 Crude odds ratios and p-values for associations between HIV status and demographic variables

Determinant	Odds ratio	95% CI	p-value
Gender			
Male	1.10	0.56-1.48	0.708
Age (years)			0
< 30	1.76	1.12-2.75	0.0126
30-39	3.11	2.11-4.60	0
40-49	0.31	0.16-0.61	0.00032
50-59	0.68	0.40-1.16	0.157
> 60	0.39	0.22-0.70	0.00102
Marital status			0.00033
Single	2.60	1.60-4.22	0
Married	0.48	0.31-0.75	0.00096
Others	1.04	0.52-2.08	0.915
Occupation			0.00005
Agriculture	0.49	0.31-0.76	0.00116
Wage earner	2.44	1.55-3.81	0
Merchant	2.04	0.99-4.21	0.0479
Others	0.78	0.48-1.28	0.33
Religion			
Islam	1.89	0.92-3.87	0.0777
Family history with TB			
Present	0.79	0.43-1.45	0.441

Table 3.2 shows that age, marital status and occupation exhibit statistically significant associations with HIV status (p-values < 0.05). The associations between the disease characteristics and HIV status are shown in Table 3.3.

Table 3.3 Crude odds ratios and p-values for associations between disease characteristics and HIV status

Determinant	Odds ratio	95% CI	p-value
Cough (days)			0.075
0-60	1.47	1.0-2.17	0.0513
61-90	0.53	0.28-1.01	0.0502
> 90	0.88	0.58-1.34	0.548
Haemoptysis			
Present	0.39	0.21-0.73	0.00225
Chest pain (days)			0.153
0-30	0.81	0.49-1.32	0.393
31-90	1.67	0.95-2.94	0.07
> 90	0.69	0.29-1.62	0.389
Dyspnea (days)			0.38
0-30	0.97	0.64-1.47	0.884
31-90	1.31	0.82-2.11	0.256
> 90	0.73	0.41-1.32	0.301
Fever (days)			0.0444
0-60	0.46	0.20-1.02	0.0513
61-90	3.51	1.21-10.22	0.0141
> 90	1.29	0.37-4.44	0.687
Weakness			
Present	1.58	1.06-2.36	0.0226
Weight loss			
Present	1.79	1.17-2.72	0.00623
Other symptoms			
present	0.90	0.62-1.31	0.58
Chest x-rays			0.00959
Non-cavity	1.79	1.22-2.63	0.00246
Cavity	0.56	0.38-0.82	0.00253
Miliary	1.05	0.24-4.65	0.948
Degree of sputum			0.216
Negative	1.23	0.69-2.20	0.481
+ 1	1.37	0.93-2.02	0.116
+ 2	0.67	0.43-1.04	0.0728
+ 3	0.94	0.62-1.43	0.784

Among TB patients, cough, chest pain, dyspnea, other symptoms and degree of sputum are not significantly associated with HIV infection status. However, haemoptysis, fever, weakness, weight loss and chest x-ray finding highly correlate with the outcome (p-values < 0.05).

The crude odds ratios for the associations between HIV status and intervening variables such as receiving BCG, smoking, drinking and other diseases are listed in Table 3.4. These factors are not significantly associated with HIV infection status since the corresponding p-values are all greater than 0.05.

Table 3.4 Crude odds ratios and p-values for associations between HIV status and intervening variables

Determinants	Odds ratio	95% CI	p-value
Receiving BCG			
yes	1.70	0.92-3.15	0.0903
Smoking			
yes	1.32	0.83-2.09	0.244
Drinking			
yes	1.52	0.98-2.36	0.0634
Other diseases			
yes	1.07	0.54-2.13	0.853

So far, we have obtained the following results.

(a) The demographic variables including age, marital status and occupation are significant risk factors for HIV status, but there is no evidence to implicate gender, religion or family history of TB.

(b) Among disease characteristics, haemoptysis, fever duration, weakness, weight loss, and a chest x-ray finding, are significant risk factors for HIV status, but not duration of cough, duration of chest pain, duration of dyspnea, other symptoms, or degree of sputum.

Confounding

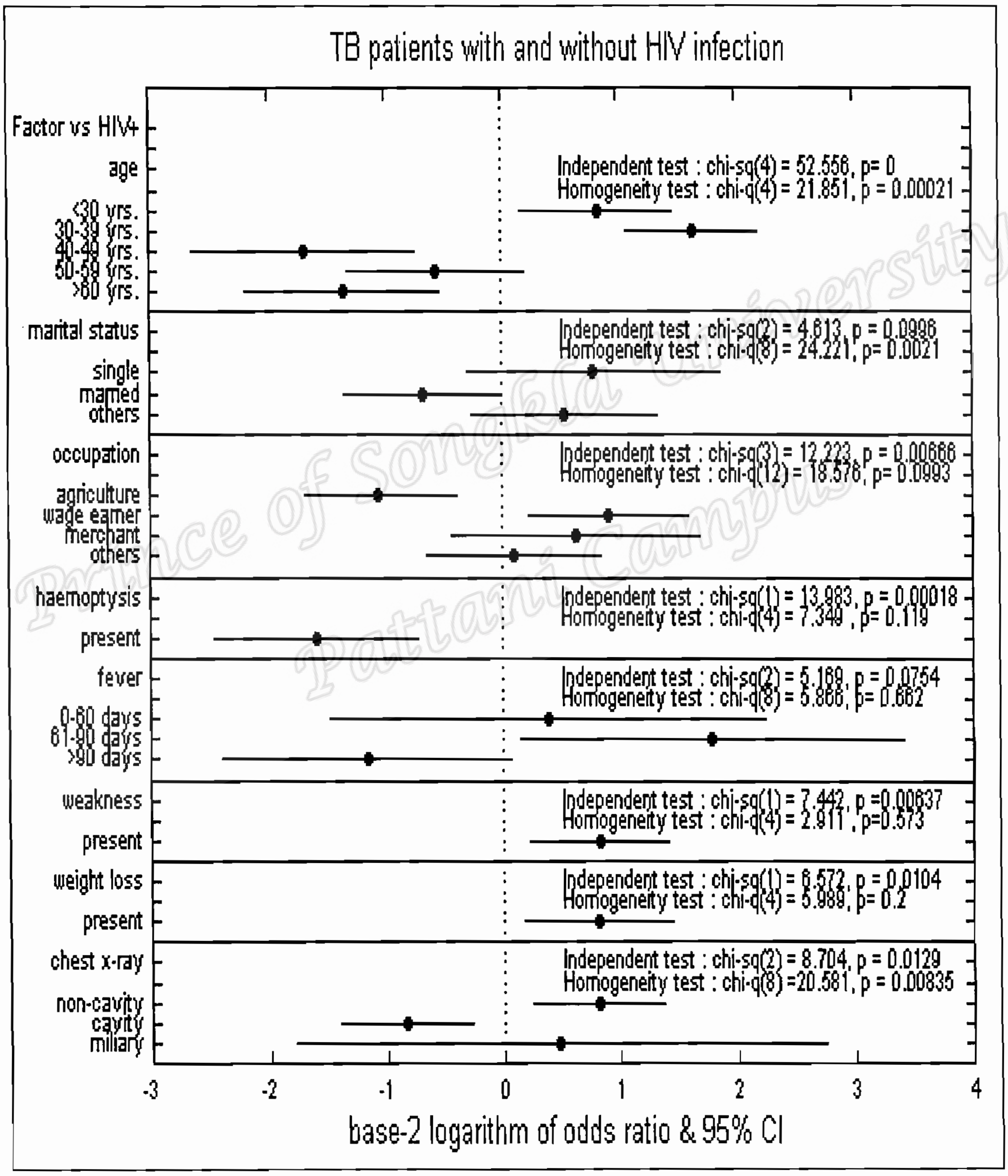
In this section we investigate the possibility of bias, due to confounding by age and gender, in the results obtained in the preceding section. As described in Chapter 2, the Mantel-Haenszel method of adjustment of odds ratios is used in this investigation. Table 3.5 shows these adjusted odds ratios.

Table 3.5 Adjusted odds ratios and p-values for all significant variables

Determinant	Odds ratio	95% CI	p-value
Age (years)			0
< 30	1.77	1.13-2.76	0.011
30-39	3.12	2.11-4.61	0
40-49	0.31	0.16-0.60	0.00031
50-59	0.68	0.40-1.16	0.154
> 60	0.39	0.22-0.70	0.00106
Marital status			0.0996
Single	0.46	0.84-2.52	0.134
Married	0.63	0.39-1.0	0.0335
Others	1.73	0.82-3.65	0.146
Occupation			0.00666
Agriculture	0.48	0.31-0.77	0.00136
Wage earner	1.88	1.17-3.01	0.00722
Merchant	1.55	0.74-3.24	0.236
Others	1.08	0.64-1.81	0.774
Haemoptysis			
present	0.33	0.18-0.61	0.00018
Fever (days)			0.0754
0-60	0.45	0.19-1.06	0.0662
61-90	3.44	1.11-10.70	0.0256
> 90	1.31	0.36-4.74	0.686
Weakness			
present	1.77	1.17-2.67	0.00637
Weight loss			
present	1.76	1.14-2.73	0.0104
Chest x-rays			0.0129
Non-cavity	1.76	1.19-2.60	0.00398
Cavity	0.56	0.38-0.83	0.0031
Miliary	1.39	0.29-6.66	0.682

All significant variables in crude odds ratio are adjusted for gender or age. Age is only adjusted for gender while the others are all adjusted for age. We illustrate them with graphical methods as shown in Figure 3.3.

Figure 3.3 Factors associated with HIV infection adjusted for age or gender



The top panel of Figure 3.3 shows the association between HIV status and age, after adjusting for gender. We see that this association is statistically significant.

Pulmonary TB patients aged below 40 years have a higher risk of HIV infection than the others. Additionally, age group 40-49 years and over 60 years have lower risk than the others corresponding to small p-values (less than 0.05). However, the association between age group 50-59 years and HIV infection is not statistically significant, as the 95% confidence interval of the odds ratio covers the null value. Since the adjusted odds ratios shown in Table 3.5 are not perceptibly different from the crude odds ratios given in Table 3.2, gender is not a confounder in this case. The chi-squared statistic for the homogeneity test is significant, indicating that the odds ratios in the various age groups are not the same. Further investigation reveals that there are proportionately fewer males than females with positive HIV status in the age group over 60 years.

The second panel of Figure 3.3 shows the association between marital status and HIV infection after adjusting for age. In agreement with Table 3.2, this shows that the crude association among both married and single persons is statistically significant (p-value = 0.00033). The single group has a higher risk of HIV infection than the others. On the other hand, the married group has lower risk. The 95% confidence intervals of the odds ratio in the others covers the null value, so the result is inconclusive. After adjusting for age as shown in the second panel of Figure 3.3, the association is no longer statistically significant (p-value = 0.0996). Thus, age is a confounder in the association between HIV infection and marital status. However, the homogeneity chi-squared test gives a small p-value. Further investigation reveals that this is because the subjects aged below 30 do not have the same risk pattern as the older subjects.

The association between occupation and HIV infection after adjusting for age is presented in the third panel of Figure 3.3. Wage earners are found to have higher risk of HIV infection. On the other hand, agriculture workers have lower risk than the others. Due to small cell counts in the merchant group, the corresponding confidence interval is very large and covers the null value. The adjusted odds ratios are essentially the same as the crude odds ratios shown in Table 3.2. Moreover, the homogeneity test is not significant in this case. Thus age is not a confounder for this association.

The fourth panel of Figure 3.3 displays odds ratios (with 95% confidence intervals) for the association between haemoptysis and HIV infection. These are shown with the Mantel-Haenszel adjusted odds ratio based on combining the results in the different age groups. Since the chi-squared test for homogeneity of haemoptysis symptom is not significant (p-value = 0.119), it is reasonable to combine these results. The adjusted odds ratio (0.33) given in Table 3.5 is similar to the crude odds ratio for this association (0.39) given in Table 3.3, and the corresponding p-value for testing the null hypothesis (0.00018) is comparable with that given in Table 3.3 (0.00225). Therefore, age is not a confounder in this case.

The fifth panel of Figure 3.3 depicts the association between fever duration and HIV infection after adjusting for age. In contrast to the result based on the crude analysis shown in Table 3.3, the association is not quite statistically significant, so age appears to be a confounder in this case. However, the odds ratios and their confidence intervals shown in Table 3.5 are very similar to those given in Table 3.3, so the extent of confounding is extremely slight.

Weakness and weight loss are both statistically significantly risk factors for HIV infection in TB patients after adjusting for age. The 95% confidence interval of the odds ratio is graphed on the right hand side of the dotted line and thus excludes the null value corresponding to a small p-value. This indicates that weakness and weight loss are associated with HIV infection.

The association between chest x-ray findings and HIV infection after adjusting for age (the bottom panel) shows a statistically significant association in the non-cavity and cavity groups. Due to a very small sample size in the miliary group, the 95% confidence interval is very large and covers the null value. Thus this association is inconclusive. The conclusion is that TB patients without HIV infection have a higher risk of cavity infiltration in the lung. On the other hand, TB patients with HIV infection have a higher risk of non-cavity infiltration.

After adjusting the odds ratio for age, we conclude that occupation, haemoptysis, weakness, weight loss and chest x-ray findings are all associated with HIV infection.