A statistical investigation of the risk factors for tuberculosis

A thesis submitted in partial fulfilment of the requirements for the Degree of
Master of Science in Statistics

by Irene van Woerden

under the supervision of
Dr. Raazesh Sainudiin
and
Dr. Arindam Basu

Department of Mathematics and Statistics
University of Canterbury

2013

[This page intentionally left blank]

[This page intentionally left blank]

Contents

[This page intentionally left blank]

List of Figures

[This page intentionally left blank]

List of Tables

[This page intentionally left blank]

Acknowledgements

I became interested in Tuberculosis after becoming infected myself. In 2011 I found myself about to start 9 months of Isoniazid treatment for a sickness I knew little about, and associated only with HIV/AIDS, starvation, and death. I now know so much more about Tuberculosis, for instance, did you know that each year more women die from TB than from maternal deaths worldwide [1].

During the course of this thesis I have learnt much more than statistical skills. My spelling and grammar have improved dramatically, as have my latex, R and Matlab abilities. I have discovered folder layouts and naming conventions which did, and did not work, and the all important difference between including and excluding a negative sign in a formula. I have (finally) learnt the difference between the affect and the effect, and discovered that the words greater, more, and less all need to be followed by a than.

This thesis would have been of a much lower standard and less enjoyable had it not been for the following people. Special thanks go to:

- My supervisors, Raazesh Sainudiin and Arindam Basu. These guys have provided the technical support for this thesis and kept me on track. They have made me re-write pages and delete entire sections but in doing so they have made this thesis better than it would have been.
- The IT guys, Steve and Paul. They were always cheerful as I explained what process I had been doing when I overloaded and crashed a uni server. They helped solve some more technical issues and provided computer expertise.
- The Learning Skills centre. They patiently taught me about colons and semi-colons and ignored the fact I still called them 'dot dot' and 'dot comma'. They tore my early sample drafts to pieces and gave me several crash courses on how to write for a professional audience.
- The Postgraduate students in the department. These are the people who have kept me company and sane. They have had lunch with me most days and are the few people who understand what latex not compiling, or Matlab and R code not working, actually means.
- My family and friends. They try in vain to be interested in my findings and have promised to read this thesis once it is finalised. I love them for their continual support and their ability to tempt me to take time off.
- The Demographic and Health Surveys and the National Family Health Survey who provided me with the dataset used in this thesis free of charge.

[This page intentionally left blank]

Abstract

Tuberculosis (*TB*) is called a disease of poverty and is the main cause of death from infectious diseases among adults. In 1993 the World Health Organisation *(WHO)* declared TB to be a global emergency; however there were still approximately 1.4 million deaths due to TB in 2011. This thesis contains a detailed study of the existing literature regarding the global risk factors of TB. The risk factors identified from the literature review search which were also available from the NFHS-3 survey were then analysed to determine how well we could identify respondents who are at high risk of TB.

We looked at the stigma and misconceptions people have regarding TB and include detailed reports from the existing literature of how a persons wealth, health, education, nutrition, and HIV status affect how likely the person is to have TB. The difference in the risk factor distribution for the TB and non-TB populations were examined and classification trees, nearest neighbours, and logistic regression models were trialled to determine if it was possible for respondents who were at high risk of TB to be identified. Finally gender-specific statistically likely directed acyclic graphs were created to visualise the most likely associations between the variables.

[This page intentionally left blank]

Chapter 1

Background and Literature Review

Tuberculosis (*TB*) is known as a disease of poverty and was declared a global emergency by the World Health Organization *(WHO)* in 1993 [2]. TB is one of the world's most deadly infectious diseases, second only to HIV with over one million people dying of TB in 2011 [2]. The purpose of this thesis was to report the association between the variables found from the existing literature which were thought to influence the TB incidence (wealth, health, education, household crowding). Using the Indian National Family Health Survey *(NFHS)* from 2005-2006 an exploratory data analysis was conducted. This provided an overview of the variables determined from the literature review to be of potential significance. With a selection of these variables generalized linear modelling was carried out to find the model which best predicted the TB status of respondents. The most statistically likely directed acyclic graph was also created to determine how the variables thought to be significant interacted with each other. These help obtain a wider understanding of TB and its associated variables.

This thesis is arranged as follows: In the next sections we introduce TB and discuss its history (Section 1.1), biology (Section 1.2), prevalence and treatment (Sections 1.3 and 1.4). We include a summary of the public health problem of TB and reasons for selecting India and the NFHS dataset (Section 1.6). We then introduce a summary of the global risk factors found (Section 1.7) and a summary of the misconceptions and stigma surrounding TB (Section 1.8). We conclude with a summary for the India-specific analysis investigating the association between wealth, health, education, nutrition, HIV, and TB in detail (Sections 1.10 to 1.12). In the remaining chapters of this thesis we explain the methods used in the analysis (Chapter 2), show a summary of the results (Chapter 3), and discuss out findings (Chapter 4).

This thesis was completed with the following structure: Initially the existing literature was searched for articles relating to TB and India. From the literature an overview of TB was obtained and a summary of the risk factors for TB compiled. The variables in the NFHS dataset relating to the risk factors found from the literature review were investigated in detail. Using an exploratory data analysis the null hypothesis that the TB and non-TB distribution functions for each variable were the same were tested. Techniques

such as nearest neighbours and classification trees were used to determine if any simple combination of variables could accurately predict a respondents TB status. After the initial analysis the variables which were the most accurate at predicting a respondents TB status were selected for the secondary analysis. The most parsimonious generalized linear model which predicted TB cases was found. The most statistically likely directed acyclic graph and moral graph was calculated to determine how the variables were affecting each other. The results were then examined and discussed.

1.1 Tuberculosis history

The tuberculosis bacilli were initially identified and described by Robert Koch on March 24th, 1882 (now celebrated as 'World TB Day'). There was no known cure for TB until after the bacilli were discovered. Before the cause of TB was known the treatments for TB were diverse, consistent only in their failure to treat TB. Potential treatments for early TB sufferers were ranging from rest or exercise; indulging in food or starving; going high into the mountains or underground; eating wolf livers or drinking elephant blood; bathing in human urine; or touching a member of the royal family.

In the 17th and 18th centuries an estimated 25% of European deaths were caused by TB [3]. Naturally mummified human remains from 1731-1838 AD were found in Hungary in 1994; 93 of the 168 examined showed evidence of having had TB (55%) [3]. The buried skeletal remains of a mother and baby off the coast of Israel were found in 2008. Both the mother and baby had evidence of TB and were carbon dated to have died around 6000BC [4]. TB has also been identified in the spines of Egyptian mummies (3500-2500 BC). A report by Zink states that of 85 mummies tested, 25 had TB [5]. TB has also been found in Neolithic Sweden (3200 BC to 2300 BC) and pre-dynastic Egypt (3500 BC to 2650 BC) [4].

Throughout history TB has been called by a variety of names such as:

- Consumption and the Wasting Disease - as TB slowly consumes the person, starting with significant reduction in energy level and weight which leads to loss of life
- White plague – as becoming pale is a symptom of individuals who are in their final stages of TB
- Phthisis – named by the Greek philosopher Hippocrates, which stands for consumption in Greek.
- King's Evil – as a touch from royalty was believed to cure TB
- Vampirism – after the death of a family member due to TB the others were likely to also succumb, causing people to think the dead person was sucking the life out of his/her living relatives
- Kochs – named after Robert Koch who isolated the Tuberculosis bacilli in 1882
- Scrofula – the medical term for TB in the neck lymph nodes

- Potts disease – named after Sir Percval Pott, the medical term for extrapulmonary TB of the spine.

1.2 Biology of tuberculosis

Human TB is caused by *Mycobacterium tuberculosis*. It is passed from person to person via the airways by small, 1 - 5 μm in diameter, bacilli [6]. When a person with active TB coughs, sneezes, or spits, the bacilli are suspended in the air [6]. If a person breathes in the tuberculosis bacilli they are at increased risk of developing TB [6]. Most people will not develop TB even if exposed due to their immune system counter-acting the infection [6]. Sunlight destroys the bacilli which makes transmission far more likely to occur in indoor settings that are also poorly ventilated, dark and damp [7]. In the right conditions the bacilli can remain airborne for around 2 hours [6]. It is also possible for people to catch TB from the cattle strain of TB, *Mycobacterium bovis (M. bovis)*. However it is rare for a person to pass *M. bovis* on to another person. *M. bovis* is spread from a warm blooded animal such as a cow, deer, pig or possum to a person by the ingestion of bacteria. This generally occurs by someone ingesting an unpasteurized milk product from an infected animal - for instance milk or cheese.

TB can infect any of the body's organs but it is generally found in people's lungs. People affected with TB can experience all, or none (asymptomatic TB) of the possible symptoms. The symptoms of TB are:

- Persistent cough with blood in sputum
- Scarred lungs and permanent lung damage
- Fever
- Night sweat
- Loss of appetite and weight loss
- Fatigue

Two form of TB are defined: active and latent. Active TB is contagious and spreads by aerosol transmission (sneeze, spit, cough). It is not possible to catch TB from skin to skin contact or from sharing food or drink with someone who has active TB. Active TB generally destroys the person's lung tissue however it can affect anywhere in the body and attack kidneys, skin, bones, reproductive system, spine and brain. Latent TB is a clinical disorder: the person's immune system is strong enough to keep the replication of the TB bacterium under control. Latent TB is not contagious and there are no symptoms or tissue damage. Around 10% of latent TB cases will develop into active TB. At any point latent TB can become active but people with a weakened immune system are much more likely to develop active TB than those with a healthy immune system [2]. This puts the young and elderly at a higher risk, along with substance

abusers, people with low body weight or HIV, and people with health issues such as diabetes, cancer, and kidney disease.

1.3 Prevalence and incidence of tuberculosis

The WHO provides global estimates for the prevalence, incidence, and mortality rate due to TB. The information provided by the WHO presents that, generally, more males are reported as having TB than females [8]. Of the 24 country profiles available only one country (Afghanistan) has a male:female infection ratio less than 1. This is consistent with other reports which state that gender may play a significant role in how likely someone is to be infected [9, 10, 11]. However, it is also possible that less female TB cases than those in male population are reported or realised [7, 10, 11, 12, 13, 14]. Countries bordering each other are did not always have a similar female:male infection ratio. For example Afghanistan had far more female than male TB cases reported. In the same region Pakistan had similar numbers of female and male TB cases reported. In China and Russia there have been more male than female TB cases reported. [2]

The WHO state that the global number of TB cases has steadily been dropping since the 1990s with significant reduction during the period 2000 - 2010. The findings show there were 8.7 million new cases of TB reported globally in 2011 of whom 1.4 million died [2]. Five countries (India, China, South Africa, Indonesia, Pakistan) made up over half of the incidence. South-East Asia accounts for 40%, Africa 26%, and the Western Pacific 19%, of the global TB incidence [2]. The mortality table shows a constant decline from 1990 to 2010 with the TB death rate reducing by 40% in this period [2] . According to the report, slightly less than one million people died from TB in 2011 and an additional 0.4 million died who were also HIV positive [2]. Refer to Table 1.1 for a detailed summary of the mortality, prevalence, and incidence due to TB worldwide.

There were 1.5 million TB cases notified in India in 2011 [8]. From which 80% (1.2 million) were new case notifications and 20% (0.3 million) were patients who were re-treated [8]. Of the 45% (0.7 million) TB cases of patients whose HIV status was known 6% (45,000) were HIV positive [8]. The highest TB incidence was shown to be in the sub-Saharan African region, followed by Asia, Eastern Europe, and Africa respectively. The lowest incidences were recorded in North/South America, Western Europe, and Oceania. Refer to Figure 1.1 for the estimated incidence of TB worldwide.

Table 1.1: Estimates of Mortality, Prevalence, and Incidence of TB worldwide.

Numbers in 1000's. Source courtesy of the WHO Global Tuberculosis Report 2012 [2].

		Mortality			Prevalence			Incidence			TB & HIV+ve Incidence		
Country	Population	Best	Low	High	Best	Low	High	Best	Low	High	Best	Low	High
Afghanistan	32,358	13	5	23	110	55	190	61	51	73	0	0	0
Bangladesh	150,494	68	29	120	620	300	1,100	340	280	400	1	0	1
Brazil	196,655	6	5	7	91	36	170	83	69	97	16	13	19
Cambodia	14,305	9	4	16	120	99	140	61	52	70	3	3	4
China	1,347,565	47	45	49	1,400	1,200	1,600	1,000	890	1,100	13	9	17
DR Congo	67,758	36	16	65	350	180	570	220	190	250	34	27	41
Ethiopia	84,734	15	11	20	200	160	240	220	160	280	38	28	49
India	1,241,492	300	190	430	3,100	2,100	4,300	2,200	2,000	2,500	94	72	120
Indonesia	242,326	65	29	120	680	310	1,200	450	380	540	15	11	20
Kenya	41,610	9	5	15	120	63	200	120	110	120	47	45	49
Mozambique	23,930	11	4	22	120	56	200	130	91	180	83	58	110
Myanmar	48,337	23	11	40	240	190	310	180	160	210	18	15	22
Nigeria	162,471	27	6	64	280	71	620	190	90	330	50	23	86
Pakistan	176,745	59	26	110	620	280	1,100	410	340	490	2	1	2
Philippines	94,852	28	25	31	460	400	520	260	210	310	1	1	2
Russian Federation	142,836	22	22	23	180	72	330	140	120	160	9	7	11
South Africa	50,460	25	11	44	390	200	630	500	410	600	330	270	390
Thailand	69,519	10	4	18	110	51	200	86	71	100	13	10	15
Uganda	34,509	5	2	9	63	33	100	67	54	81	35	28	42
UR Tanzania	46,218	6	3	11	82	43	130	78	73	83	30	28	32
Viet Nam	88,792	30	12	55	290	130	500	180	140	220	14	11	18
Zimbabwe	12,754	6	2	11	70	37	110	77	59	96	46	36	58
High-burden countries	4,370,719	820	680	980	9,700	8,300	11,000	7,100	6,800	7,500	890	810	970
Africa	857,382	220	180	270	2,500	2,100	3,000	2,300	2,100	2,400	870	800	950
Americas	943,019	21	18	24	330	250	420	260	240	280	37	34	40
Eastern Mediterranean	608,628	99	61	150	1,000	660	1,500	660	590	740	9	8	10
Europe	899,500	45	44	46	500	370	650	380	350	400	23	20	25
South-East Asia	1,830,361	480	350	630	5,000	3,800	6,300	3,500	3,200	3,700	140	120	170
Western Pacific	1,808,797	130	100	150	2,500	2,200	2,800	1,700	1,500	1,800	36	31	42
Global	6,947,687	990	840	1,100	12,000	10,000	13,000	8,700	8,300	9,000	1,100	1,000	1,200

Notes:
Best, Low, High are the best estimate, the lowest estimate, and the highest estimate
Mortality excludes deaths from those with HIV co-infection. These are reported as HIV deaths
Lower and upper limits are from the 95% uncertainty intervals
India estimates are not yet officially approved and should be taken as provisional only.

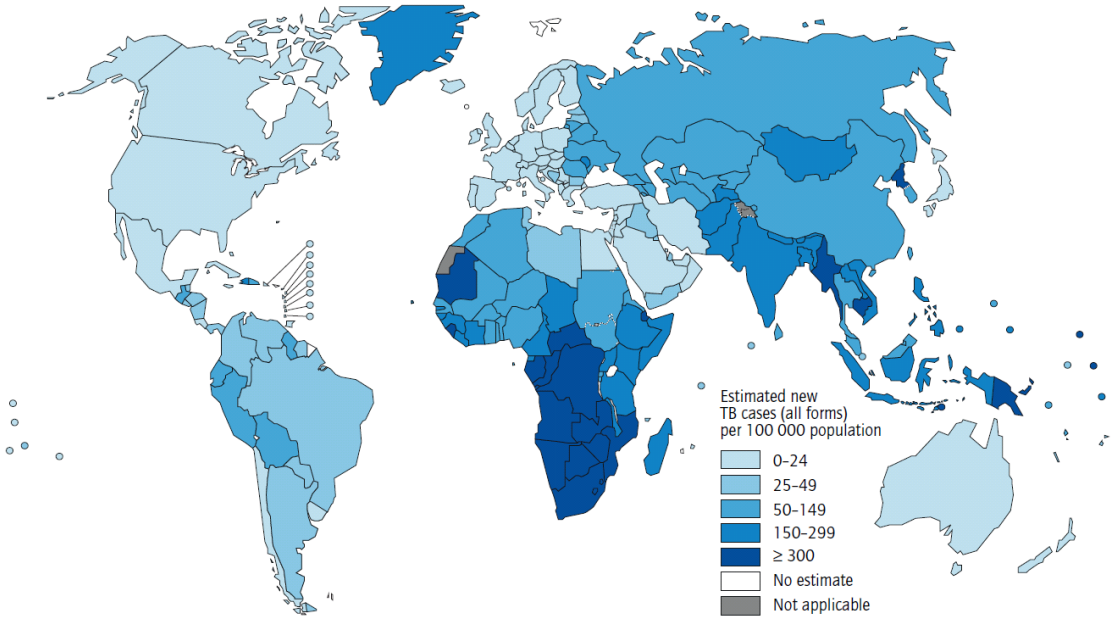

Figure 1.1: Estimated global active TB rates.
Source courtesy of the WHO Global Tuberculosis Report 2012. [2]

1.4 Tuberculosis prevention and treatment

Approximately 33% of the world's population is estimated to be infected with TB [15]. However, most of these are in terms of latent TB, not active. The infected populations are not uniformly spread globally. Currently around 95% of TB deaths are from the developing world. Generally TB can be fully treatable with chemotherapy but there are new strains of TB emerging which are resistant to the standard treatment. Drug-resistant TB has occurred for multiple reasons such as medical staff having poor TB training and not providing appropriate treatment. Other examples are patients not complying with the treatment regime and the drugs being of low quality or not consistently available [16, 17, 18, 19, 20]. It is not possible to guarantee full protection against TB but being vaccinated and having limited contact with the infected persons can diminish the risk of infection.

The current vaccination against TB is called the Bacille Calmette Guerin *(BCG)* vaccine which works by a process called 'active immunity'. An injected small amount of weak *M. bovis* causes the person's immune system to create antibodies against the bacteria. The created antibodies stay in the body and provide immunity against future exposure to *M. bovis*. Despite providing immunity there are also challenges with the BCG vaccination. One of these challenges is the effectiveness of the vaccine as it decreases over the time, having the vaccine as a child is not a fail-safe method of avoiding TB. The WHO recommends a single BCG vaccine as multiple injections and boosters have not yet been proven to be more effective. Some countries including Turkey, Hungary, and Russia recommend multiple BCG injections [21] while in the United States due to the low risk of TB there is generally no administration of the BCG vaccination.

In New Zealand the BCG vaccination is only recommended for high-risk children. These children include the ones who live with someone who has TB, those with parents from a country with high prevalence of TB, and children who will spend at least three months of their first five years of life in a country with high TB prevalence.

Early diagnosis and immediate treatment are vital to stop the spread of TB because the longer the delay in diagnosis and treatment the more chances of passing the infection to others. Studies have shown that people delayed seeking treatment due to the travel time and cost to get to a clinic [22, 23]. People who tried to hide their TB and those who were unable to travel to a clinic discreetly delayed their treatment. In general TB is fully treatable with chemotherapy, commonly using isoniazid and rifampicin. A person is no longer contagious once he/she has been on treatment for 2 weeks. The length of treatment varies but it generally takes between 6-18 months. Differences were observed between the medication doctors' reported their patients as taking, and what medication patients were actually taking. For instance, while patients were reported to be on three or four drugs to treat TB, it was found that the patients were actually only taking one or two drugs [24].

Drug-resistant TB is a large and relatively new problem for the medical professionals in the eradication of the disease. TB has a long treatment period, often 6-18 months, and the rate of patients defaulting from the treatment regime is high. Surprisingly, there was no significant difference between the rate of drop out from treatment between urban and rural populations [24]. As with any antimicrobial treatment, stopping before the full course of the medication has been taken is potentially dangerous. With the treatment stopped the remaining bacteria can begin to multiply. These remaining bacteria are potentially now resistant to the treatment. It is often not possible to resume taking the drugs initially given after defaulting from treatment.

Some of the main reasons for defaulting were due to the cost of treatment [14, 25], the side effects from the treatment [25, 26, 27], because they felt better [14, 20, 26, 27], having limited knowledge of TB and TB treatment [25, 27, 28], not having the drugs available to take [27], the distance needed to travel to get to a medical centre [20], and the stigma attached to admitting to having TB [28]. The most likely groups to drop out of treatment were elderly patients, workers, the poor, alcoholics, lower educated groups, people with a language barrier, and people with limited knowledge of TB and TB treatment [29]. The locations and hours TB clinics were open meant patients sometimes had to decide between continuing working or getting treatment; seeking treatment and working was not possible [14].

Some strains of TB are now resistant to the drugs which have successfully been used to treat TB in the past. There are currently three levels of TB which people contract: standard TB, Multi Drug Resistant TB *MDR-TB)*, and Extensively Drug Resistant TB *(XDR-TB)*. A fourth level of Totally Drug Resistant

TB *(TDR-TB)* has been coined but is not recognised by the WHO [15]. Some features of these levels of TB are listed below:

- TB
  - Treatable by isoniazid and rifampicin
  - Resistant to none
- Multi Drug Resistant TB (MDR-TB)
  - Treatable by an aminoglycoside, a fluoroquinolone, Ethionamide or Prothionamide, one of Cycloserine, PAS, Clofazimine or Moxifloxacin
  - Resistant to isoniazid and rifampicin
- Extensively Drug Resistant TB (XDR-TB)
  - Treatable by specialised combination of second line drugs (not isoniazid or rifampicin)
  - Resistant to isoniazid, rifampicin, one of ciprofloxacin, levofloxacin or moxifloxacin, and one of capreomycin, kanamycin or amikacin
- Totally Drug Resistant TB (TDR-TB)
  - Not treatable
  - Resistant to all drugs currently used to treat TB.

Standard TB is relatively easy to treat, it is harder to treat MDR-TB and XDR-TB but treatment is still possible. The term TDR-TB was only coined in late 2011 after the discovery of a TB strain which was untreatable. The more resistant the TB strain is, the more difficult it is to treat and consequently the lower the survival rate. In South Africa, the mortality rate of TB (excluding TB and HIV co-infection cases) in 2012 was 49 per 100,000 [8]. A study from 2006 in Tugela Ferry, South Africa analysed the sputum of 1,540 patients with suspected TB from a provincial government hospital. Of the 1,540 patients, 53 were infected with XDR-TB and 52 of these patients died (98%). The median time until death for 52 patients who died was 16 days [30]. The increased death rates for the XDR-TB example are further complicated by other factors. For instance 44 of the XDR-TB patients were tested for HIV and all were HIV positive.

1.5 Public health and TB

In 1920 Winslow defined public health as the prevention of disease, the prolonging of life, and the raising of peoples health levels [31]. This occurs when all levels of a community are working together, making

informed health decisions, and implementing effective health policies. The WHO is working closely with the 'STOP TB partnership' to eradicate the public health problem of TB. The STOP TB partnership operates in over 100 countries and works with all levels of the society, from government programs to community groups [32]. It aims to prevent the transmission of TB; to have every TB patient effectively diagnosed, treated and cured; and to have the global incidence of TB in 2050 less than 1 per million [32].

The strategies for prevention and control of TB need to be multi-level if they are to be effective. The Directly Observed Treatment System *(DOTS)* is recommended by the WHO. DOTS is based on five key factors thought necessary to control TB:

- Commitment from the national government. New legislation where required, sufficient funding for health and TB care
- Laboratories and trained staff which enables the detection of new TB cases through bacteriology
- Accessible treatment services which use effective drugs with short courses. Supervision of patients to ensure they take the full course of medication
- High quality, consistently available, and free drugs to be used for the treatment of TB
- Monitoring and analysis of TB at the national, state, province and district level via compilations of every TB patients data [33].

This thesis is focused on one aspect of the DOTS fifth key factor – the analysis of TB as the national level in India. National Health Surveys allow the monitoring of a nations health; they see the progress, or otherwise, of the nation over time. They can also be used to determine if interventions are improving the communities health or not.

1.6 Reason for selecting India's NFHS as dataset

By area India is the seventh largest country in the world. India has a population of 1.2 billion people and is the second most populous country. Of the countries with a population over 10 million, India the sixth highest population density. It is possible that over half of India's population are infected with TB [34]; however the majority of these will have latent TB which is not contagious and has no symptoms. The WHO estimates the incidence of TB in India to be the highest globally; between a quarter and a third of the world's TB cases are thought to be in India [2, 35]. India has over double the TB incidence of the next highest incidence country (China) and more than four times larger than the third (South Africa) [2, 35]. Examining the TB population in India is an interesting challenge due to the size of the country, its large

population, and its high TB incidence. I was also personally interested in India as I had spent five months exploring the country.

India has completed three National Family Health surveys *(NFHS)*, the first in 1992-1993, the second in 1998-1999, and the third in 2005-2006. These surveys were conducted under the Ministry of Health and Family Welfare (MOHFW), Government of India and co-ordinated with the Demographic and Health Surveys *(DHS)*. DHS are experts in surveying developing countries and specialise in health and population studies; they have provided technical support for over 260 surveys throughout 90 countries. Several large international organizations were also involved, including the United States Agency for International Development *(USAID)* and the United Kingdom Department for International Development *(DFID)*.

NFHS-3 was the version used in this analysis. NFHS-3 had data collected by 18 different research organizations across the 29 states. The interviewers were trained in interview techniques and conducted the survey in either English or the native language of the state they were in, depending on the respondents preference. The survey instrument was also translated into 18 languages. For more details about the NFHS survey instruments please see the Section 5.2 in the Appendix. The response rate for NFHS-3 is very high. In total, of the of the 116,652 sampled, 109,041 households had completed interviews. There were 124,385 women aged 15-49 interviewed (of 131,596 eligible, 94.5% response rate), as well as 74,369 men aged 15-54 (of 85,373 eligible, 87.1% response rate) [36].

The NFHS-3 provided information on the wealth, health, household information, education, gender, and TB status of Indian households. These variables are also thought to influence the prevalence of TB, they were vital to be included in the analysis. Questions ranged from general population questions such as the altitude of the area, to health questions such as the person's height, weight and haemoglobin levels. Information about education levels, wealth, smoking habits, TB knowledge, household size and marital status were also obtained. See Table 3.10 for the amount of TB per state.

For meaningful results the data used needed to be from a well planned and extensive survey, preferably from throughout all of India. The data also had to be accessible, reliable, relevant, and recent. The NFHS dataset fulfilled these expectations which is why it was chosen. A summary of the number of prevalence of TB from the NFHS dataset by state is shown in Table 3.10 in the Appendix.

While the NFHS data was the best available at the time, there have been data integrity concerns. For instance, the nutritional status of similarly aged children of the same gender in a household would be expected to be similar. The indicators of malnutrition for both the NFHS-2 and NFHS-3 surveys were if a person was below three standard deviations from expected in stunting, wasting, underweight, and having less than 10 g/dl haemoglobin for anaemia. The percentage of siblings with the same level of malnutrition was lower in NFHS-3 compared to NFHS-2 [37]. This possibly shows a lessoning of data quality from

NFHS-2 to NFHS-3. It has been discussed that children in the same household would be expected to have similar nutritional levels and the percentage of similarities in the NFHS is suspiciously low [37]. The wealth index has also been criticized as being biased towards urban respondents and not distinguishing between respondents who were very poor and respondents who were poor [38, 39]. For instance it was found that 80% of respondents who were identified as living in slums from the NFHS-3 survey belonged to the two highest quartiles of the wealth index [38]. Wealth is a calculated field as while it would be possible to identify each respondents economic wealth exactly this would take a large amount of time [39]. It is possible that a slightly different calculation for the wealth index would identify the respondents economic wealth more accurately.

1.7 Global risk factors of TB

TB is called a disease of poverty due to people living in poverty being at increased risk of TB infection. People living in poverty have very few possessions, ways of supporting themselves, or money, by definition. Some of the expectations of people living in poverty and the associated risks of TB are listed below:

- To live in low quality housing; which is directly related to an increase in infection risks [40, 41]
- To be malnourished and lack nutrients such as iron, iodine, and vitamin A; which increases the risk of illness [42, 43]
- To have limited access to health care and to be of poor health; which exacerbates a low financial status due to the cost of absences from work for medical care and medical costs [40, 41]
- To use biomass fuels for cooking instead of cleaner fuels such as LPG; air pollution from biomass fuels causes negative health effects [44, 45]
- To work long hours and to have extremely limited time for leisure, child raising, family, and developmental activities; which also impacts on the time available to see medical staff [42]
- To have little or no education and to be illiterate; which impacts on the type of work available for them, their income, and their opportunities to seek medical help [42]
- to lack freedom and opportunities.

The relationship between poverty and TB is complicated due to people who are poor being more likely to become infected with TB, but once infected with TB, a person is less able to earn and more likely to become poor [46]. It is clear that poverty increases the risk of contracting TB; however, once it is contracted, poverty worsens the disease. People in poverty are often in less favorable circumstances for early diagnosis of TB, are less likely to be able to keep medical appointments, are more likely to have other

medical problems along with TB, and are less likely to complete appropriate treatment [41]. Along with the diminished probability of successful treatment of people in poorer communities, the risk of drug-resistant TB occurring is increased in these groups. A summary of the risk factors found for TB from the existing literature are presented in Table 1.2.

Table 1.2: Summary of risk factors for TB as found from the existing literature.

Population studied, Risk factors	OddsRatio, 95% CI
Samara, Moscow, Russia [47]	
Diabetes	2.66 (1.10 - 6.46)
Relative with TB	2.94 (1.79 - 4.85)
Drinking raw milk	3.58 (2.58 - 4.97)
Having few assetts	16.70 (8.87 - 31.43)
Low living space per person	2.99 (1.92 - 4.68)
Not employed	6.10 (4.32 - 8.61)
Food shortages	2.72 (1.56 - 4.74)
Low financial security	5.67 (3.29 - 9.76)
Smoking 1-2 cigarettes per day (cf none)	3.76 (1.15 - 12.30)
Heavy drinker	2.89 (1.50 - 5.56)
Used illicit drugs	8.74 (3.06 - 25.01)
Been in pretrial detention centre	5.70 (2.63 - 12.36)
Been in prison	12.50 (3.80 - 41.13)
Tamil Nadu, Southern India [48]	
Smoking $\geq$ 10 cigarettes per day (cf none)	2.6 (2.2 - 3.1)
Smoking more than 15 bidis per day (cf none)	1.5 (3.7 - 5.5)
Catalonia, Spain [49]	
Male	1.7 (1.5-2.1)
Aged $\leq$15 (cf aged 44)	0.1 (0.1-0.2)
Aged 15-24 (cf aged 44)	1.5 (1.2-2.0)
Aged aged 25-44 (cf aged 44)	(1.4; 1.1-1.7)
HIV infection *1	0.7 (0.5-0.8)
Significant alcohol abuse	2.2 (1.8-2.8)
Drug use	0.7 (0.6 - 1.0)
Varying populations [50]	
At least 10% underweight (U.S. navy recruits)	nearly 4 times higher

Lowest BMI category (Norway)	more than 5 times higher
BMI less than 18.5 (India)	11 times higher
Mid arm circumference ≤24cm (cf greater) (India)	7 times higher
Lacto-vegetarian (cf eat meat/fish daily) (London)	8.5 times higher
Vitamin D, E, C deficiency	significant
Selenium deficiency *2	significant
Sub-Saharan Africa [51]	
Male	2.58 (1.85 - 3.60)
Aged 55+ (cf aged 15-24)	4.08 (2.64 -6.31)
Household crowding	no difference
House has ceiling	no difference
Only male adults in household	2.21 (1.57 - 3.12)
Only female adults in household	2.11 (1.10 - 4.04)
Crowding caused by adults	1.68 (1.18 - 2.39)
Crowding caused by children	0.78 (0.58 - 1.07)
Balanta ethnic group cf Pepel *3	2.13 (1.36 - 3.32)
Poor quality of house foundations	1.66 (1.24 - 2.22)
South Africa [52]	
Female gender	0.82 (0.49 - 1.39)
Increasing age cf 15 - 29 yr population.	1.23 (0.65 - 2.30)
Caucasian cf African	0.12 (0.02 - 0.84)
Coloured cf African	1.00 (0.43 - 2.33)
One additional year of education	0.90 (0.86 - 0.94)
Worked for payment in last 12 months	0.59 (0.34 - 1.04)
Ever worked in a mine	1.55 (0.61 - 3.92)
Ever worked in a goldmine	2.40 (0.94 - 6.10)
Urban residence	0.61 (0.34 - 1.09)
Ever smoked 100 cigarettes or more	2.28 (1.30 - 4.00)
Ever drunk alcohol	1.72 (0.99 - 2.97)
CAGE score greater than one	3.09 (1.74 - 5.48)
BMI less than 18.5	4.71 (2.63 - 8.43)
One additional adult per bedroom	1.27 (1.03 - 1.55)
Meal missed due to lack of funds	2.44 (1.31 - 4.54)
Highest asset score quantile (cf average)	0.15 (0.03 - 0.69)

Birmingham [53]	
Markers of deprivation *5	significant
India [54]	
Biomass fuels (cf cleaner fuels)	2.58 (1.98 - 3.37)
Has separate kitchen	0.71 (0.63 - 0.81)
Pucca or semi pucca house (cf Kaccha) *4	0.89 (0.78 - 1.02)
More than 2 people per room	0.96 (0.85 - 1.09)
Aged 60-69 (cf 20-29)	4.44 (3.58 - 5.49)
Female	0.56 (0.50 - 0.63)
Urban residence (cf rural)	1.12 (0.92 - 1.36)
High school or above education (cf illiterate)	0.46 (0.36 - 0.60)
Hindu (cf muslim)	0.81 (0.65 - 1.01)
"Other" religion (cf muslim)	OR 0.64 (0.45 - 0.91)
North and North East region (cf South)	1.45 (1.14 - 1.86)
Central and East (cf South)	1.19 (0.99 - 1.43)
West (cf South)	1.13 (0.91 - 1.40)
Bangalore, India [35]	
Unmarried/widowed/separated	1.05 (0.58 - 1.88)
Religion other than hindu	0.80 (0.51 - 1.27)
Over 10 years schooling (cf none)	0.24 (0.11 - 0.51)
Not employed	1.28 (1.78 - 2.07)
Unskilled labor (cf business)	1.59 (0.87 - 2.93)
Skilled labor (cf business)	1.20 (0.59 - 2.44)
More than 4 people per house	1.24 (0.81 - 1.90)
Household income $\geq$ 5000 Rs (cf less 1000)	0.36 (0.20 - 0.67)
Household income 1000-5000 Rs (cf less 1000)	0.77 (0.46 - 1.29)
Basic household posessions (cf modcons)	1.62 (1.02 - 2.69)
Multi roomed house (cf single room)	0.97 (0.60 - 1.58)
More than 2 people per room (cf 2 or less)	0.79 (0.53 - 1.19)
No separate kitchen	6.00 (2.53 - 14.24)
Uses biomass fuels (cf gas/electic)	1.80 (1.10 - 2.90)
Past smoker (cf never smoked)	2.31 (1.12 - 4.79)
Current smoker (cf never smoked)	1.17 (0.59 - 2.33)
Current drinker (cf non drinker)	2.13 (1.02 - 4.44)

	Past drinker (cf non drinker)	1.06 (0.54 - 2.08)
	Has chronic disease	1.80 (1.10 - 2.93)
	Has no TB contact	1.24 (0.73 - 2.10)
	BMI less than 18.5	11.11 (5.62 - 21.98)

notes

*1	Low transmission of TB found in HIV positive respondents
*2	Selenium helps maintain the immune processess
*3	Only half of the ethnic groups had significantly different OR
*4	See Table 5.2 for Pucca and Kaccha definitions
*5	Only for Caucasian population, not for Asian population

While the transmission method of TB is well understood the social factors, such as smoking and household overcrowding, associated with its spread are understood less well. How much of an affect each of the risk factors has on TB levels is debatable with articles having conflicting results. The commonly cited risk factors of overcrowding, low education, and poverty, were not significant in all studies. One study from Zambia even reported that no socio-demographic variable, such as gender, literacy, employment status and smoking/indoor pollution were significant factors of TB [55]. Understanding which of the variables are significant, and which are highly correlated due to confounding factors is a difficult task. For instance it is difficult to distinguish if the higher rates of TB are due to ethnicity or some behaviour linked to ethnicity - such as the level of household crowding. While income, education and occupation do not always end up as significant factors they are often mentioned as risk factors or investigated as potential risk factors in past studies.

Overcrowded living conditions are commonly cited as a main cause of TB spreading, and the scientific view of TB shows that people living in close proximity to each other give each person a higher chance of catching TB.

- A study from 1999 investigated 1,516 notified TB cases from 39 electoral wards in Birmingham (England). It was found that for single variable analysis, the TB rate was significantly associated with the proportion of households with more than 1.5 people per room (P-values 0.0036) in the Caucasian population. However, it was also found that the Asian population had no single variable significantly associated with TB; the proportion of households with more than 1.5 people per room was not significant under single variable analysis (P-value 0.18).

- A study from 2004 compared 247 TB patients from Guinea-Bissau (Africa) to the non-TB population. The adjusted odds ratio for 3-4 adults in a household (compared to 1-2 adults) increased to 1.67 (95% CI 1.15 - 2.42). The odds ratio for 1-2 children in a household (compared to 1 child) decreased to 0.72 (95% OR 0.50 - 1.04) and further decreased with additional children to 0.51 (95% OR 0.32 - 0.80) for 5 or more children [51].
- In 2006 there were 3,000 TB cases aged less than 15 years from California analysed. The results of the multilevel analysis showed households with over crowding had an incidence rate ratio of 0.87 (95% CI 0.77 - 0.98) [56].
- Another report from 2006 looked at 189 TB patients from the South Indian population. It was reported that household crowding did not increase the risk of TB [35].
- A study from 2008 reported that it was currently not known how important household crowding was due to inconsistent findings [57].
- This contrasts articles from Russian [47] and summarizing articles which state that household crowding is clearly a factor for TB [58, 59].

TB is an Acquired Immune Deficiency Syndrome *(AIDS)* defining criteria for people with Human Immunodeficiency Virus (HIV) infection. TB is the leading cause of death for HIV patients and in 1993 the WHO declared HIV to be a global emergency. Active TB is much more likely to develop when the immune system is weakened. HIV, by definition, weakens the immune system; people with HIV are much more likely to develop active TB than people without HIV. Because of this relationship TB has become the leading cause of death among HIV patients. The TB and HIV/AIDS synergy has been aptly called "the synergy from hell" due to its deadly implications. There are approximately 34 million people globally who are HIV positive and 12 million people infected with TB, while 1.1 million people are infected with both HIV and TB. Of the 1.8 million deaths due to AIDS, and the 1.1 million deaths due to TB, approximately 0.35 million were due to the TB-HIV combination.

1.8 Tuberculosis and stigma

TB is highly stigmatized in some populations, including India, and is not always well understood [59, 60]. Misconceptions about how TB was acquired were varied and included:

- physical contact [23]
- dirty hands [23]
- contaminated food and drink [19, 23]

- poor diet [61, 62]
- smoking [61, 63]
- drinking alcohol [61]
- injections [23]
- genetic disposition [19]
- blood transfusions [19, 23]
- witchcraft [29, 62]
- disreputable behaviour [29]
- germs and unhygienic living conditions [61]
- worry [19, 61]
- being punished for sins [62].

People associated TB with sexual promiscuity [19, 64], drinking alcohol [64], smoking [64] , sins [64], hard work [22, 62], cold air [22], and exposure to dust [22].

The amount of stigma associated with TB in some regions has changed over time. There were reports of three phases of stigma associated with TB: the negative stigma when TB was untreatable, a lessening of the stigma when treatment became available, and now, a large negative stigma due to the association of TB with AIDS [64]. Examples of TB stigma are: being rejected and excluded from events and isolated from their families and communities [19, 22, 23, 29, 61, 62, 64, 65], being blamed [61], being thought less of [29], no longer being able to find a partner to marry [19, 23] and at risk of divorce [22, 29]. In Rhinii, a town in the Makana Municipality of the East Cape of South Africa, a survey from 2007 found that 71.2% of the population agreed that people who became infected with TB due to drinking and smoking 'got what they deserve' and 87.3% agreed that people who drank and smoked would never be able to be treated for TB [64].

Respondents who had TB worried about their future marriage prospects and thought that having TB made them, and their family, less marriable [66], [23]. Surprisingly, while nearly all of the respondents knew that TB was fully treatable, nearly 40% said they themselves would not marry someone who had been infected by TB [66]. Of the 20% of respondents who would keep their TB status secret, not wanting to be excluded or lose friends comprised 69% [23]. Evil spirits, sorcery, witchcraft, and sexual intercourse were thought of as causes for a person to contract TB; people with TB are not only stigmatized but socially rejected [29, 66].

As TB is stigmatized, infected patients are likely to try and conceal that they have TB. If people are not presenting to medical clinics to be tested the medical clinics do not accurately know the level of TB in their area. If people are not reporting that they have TB when questioned, the level of TB reported in an area can be much lower than the reality. This results in difficulties in obtaining accurate information about who is infected, which in turn makes it harder to treat the infected population. It can also be difficult to obtain treatment secretly so patients may either not get treated or drop out of treatment when keeping the medical trips and medication secret is too difficult. People concealing that they have TB and not seeking treatment compounds their problem as while they are not being treated they are still contagious and have the possibility of infecting many more people than they would have otherwise [48]. The data from the NFHS indicated that around 16% of people who had heard of TB would want to keep their TB a secret from family, friends and neighbours. Women who tried to conceal that they had TB dropped out of treatment to prevent people finding out and marginalizing them [61]. A study from Pakistan found that nearly 40% of patients did not tell their family and friends that they had TB [19]. 11% of doctors reported that they would not always tell a patient that they (the patient) had TB [24], most likely due to the stigma of TB. Of the patients who were put on treatment, the ones who stayed with the treatment also tended to eat healthily and visit a health centre while the ones who dropped out of treatment were more likely to pray for healing [62].

1.9 TB and wealth in India

The NFHS-3 provides a detailed picture of the wealth distribution in India. The wealth index field was provided in the NFHS-3 data. It was calculated from variables such as the respondent's drinking water source, house type, and cooking fuel to determine the respondent's wealth and is explained further in Section 5.4. As high rates of TB are constantly associated with the people in the lower wealth brackets, this was an essential variable to include in the analysis.

India is a diverse country with highly differing regions. For example, the life expectancy in India was 66 years for females and 63 years for males in 2011 [67]. However, this varies dramatically by state: a life expectancy was 56 years in Madhya Pradesh and 74 years in Kerala [67]. In India, 33% of the population lives in urban areas and 67% live in rural areas. The wealth distribution was not even between the rural and urban population: a significantly larger percentage of the rural population (28%) was in the lowest wealth bracket compared with the urban population (3%) [68]. Conversely, 48% of the urban population was in the highest wealth bracket while only 7% of the rural population was in this bracket [68]. Of the rural population, 56% had electricity, this increased to 93% for the urban population [68]. See Figure 1.2 for how the wealth is distributed between the urban and rural populations.

People with limited funds are more likely to delay travelling to a TB clinic than their wealthier counterparts [22]. They are also more likely to be unable to keep medical appointments, have other medical problems, and not complete treatment for TB [41]. This is especially seen in people with severely limited funds in the rural community [22]. People with limited free time are also at increased risk of TB due to their inability to easily travel to a TB clinic. The WHO stated that 43% of married women and 99% of married men were employed [69]. This percentage changed depending on the definition of employment. When unpaid work, such as collecting fuel, fodder, fruit, and water were included, the time spent working for rural and urban men and women showed that women worked longer hours than men [70].

Table 1.3: The percentage of respondents in each NFHS-3 wealth index level who would keep a TB infection secret.

Wealth Index Level	Keep TB secret	
	Men	Women
Lowest	18.6	14.2
Second	17.7	16.7
Middle	18.7	18.1
Fourth	16.4	17.5
Highest	13.0	16.3

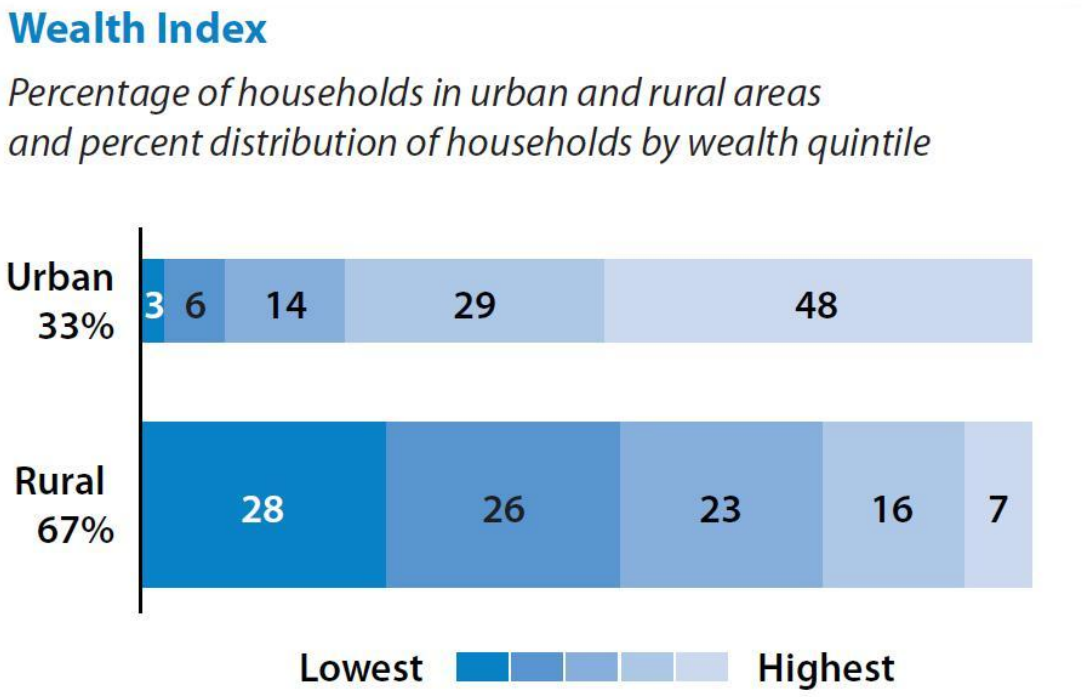

Figure 1.2: Wealth index from NFHS-3 split by rural and urban households.
Source thanks to USAID/MEASURE DHS [68]

A person's educational level affects their employment prospects. People with high education levels have a higher chance of finding secure employment that pays well. Well-paying employment enhances people's status and gives them more life opportunities such as the ability to seek medical help. It is generally people

with low education levels who work as casual labour. This is often in the agricultural field, does not pay well and is often seasonal, intermittent and uncertain [71]. The likelihood of women being involved in work outside of the lowly paid agriculture field increased dramatically with education; unmarried and divorced women were also less likely to work in the agricultural field [71]. Compared with an illiterate woman, the likelihood of not working in agriculture increased 4.8 times for a woman with secondary education, and over 30 times for a woman with graduate qualifications [71].

1.10 TB and health in India

Smoking is reported to increase the risk of TB. A recent report found a doubling of the death rate from TB in smokers [48], while another report found a 4-fold increase in risk of TB infection in smokers [34]. Of the 1.1 billion smokers globally, 16.6% of these are in India [63, 72, 73]. In India, 57% of men and 11% of women use some type of tobacco product [68]. In the 15-49 year age group 1.4% of females and 33% of males were reported in the NFHS-3 survey as smoking bidis (which have higher levels of nicotine, tar and carbon monoxide than 'regular' cigarettes) [72, 73]. The smoking proportion varies by state (as low as 14% of males in Goa and as high as 74% of males in Mizoram) and by education and rural/urban area (lower education and rural areas have higher rates of smokers) [68]. Of the smoking population, 40% reported smoking more than 10 bidis/cigarettes in the previous 24 hours.

While smokers are at a higher risk of TB, cooking smoke from biomass fuels is also suggested to increase the risk of TB [34]. There is an association between people using biomass fuels and being of poor health. For instance there were significantly higher levels of blindness (both partial and complete) in women and men from households which cooked with biomass fuels [54]. There are reports that high levels of air pollution are associated with health problems [54], including TB [34, 54, 68], but these reports are not definite in their association between cooking smoke and TB [54]. An association between biomass fuels and TB could be explained by the inhaled smoke interfering with lungs' natural processes. Over half of the worlds' population uses biomass fuels for cooking and heating [34]. From the 1992-93 NFHS survey in India it was found that biomass fuels such as wood, dung cakes and crop residues were the main cooking fuels for 75% of Indian households [34, 54].

Immunizations were associated with having educated parents over 20 years of age and having received antenatal care [74]. The vaccine against TB (BCG vaccine) was the most commonly administered vaccination in India with a 57% immunization rate. International guidelines specify that children should be fully immunized by 12 months. The overall immunization rates from the 1992-93 NFHS survey for children aged between 12 and 24 months were the following:

- 34% had received no immunizations [74]

- 37% had received some of the available immunizations [74]
- 29% were fully immunized against TB, diphtheria, pertussis, tetanus, polio and measles [74].

Children who were immunized, male, and born first were more likely to survive childhood than their counterparts. The probability of death for children who were not immunized was 6.4%; for immunized children this dropped to 1.8% [74].

Child marriage was linked to higher rates of pregnancy complications and death for both the mother and child during pregnancy and childbirth [75]. Females who have a child marriage also have an increased risk of becoming infected with HIV. The risk of any TB infection becoming active in people who are HIV positive is much higher than in the non-HIV population [76]. The demographic group who have child marriages are also at high risks of TB. Child marriages are more common in rural, central/eastern India, in the lower wealth bracket and in the lower educated population [75]. For instance the female slum population surveyed in the NFHS-3 were more likely than the non-slum population to marry before 18 years of age [38]. The slum population were also less likely than the non-slum population to use contraceptives and to have fewer children [38]. India has had laws prohibiting child marriage since 1929 when it specified that the legal age of marriage was 12 [75]. The legal age of marriage was increased to 18 in 1978 however this seems to be largely ignored as the median age for marriage in women is 17.2 years while for men it is 23.4 years.

1.11 TB and education in India

Respondents with low/no education are far more likely to have TB than respondents with high levels of education. India has 22% of the world's total population and 46% of the world's illiterate population [77]. The literacy rate is increasing. In 1951 the literacy rate was 9% for females and 27% for males [77]. The 2001 census in India showed the literacy rate was 54% for females and 76% for males [78]. The literacy rate had increased to 65% for females and 82% for males for the 2011 census [78]. These values do not show the increasing difference between the percentage of boys and girls attending school with age group. See Figure 1.3 for the percentage of male and female Indian children attending school by age group. The gender gap is small (max 2%) in urban schools even at the 15-17 age group, but is more pronounced in rural areas. In 2005 a survey was conducted to determine the school attendance rates in rural India. The results from this survey show a lower percentage of females, and older students being educated [77].

A child's grade was not always representative of the child's abilities. For instance in 2006 a study found that 47% of Indian students studying in grade 5 could not read a story from the grade 2 level [77]. This has implications for data analysis which uses a respondent's grades to estimate educational level. A survey

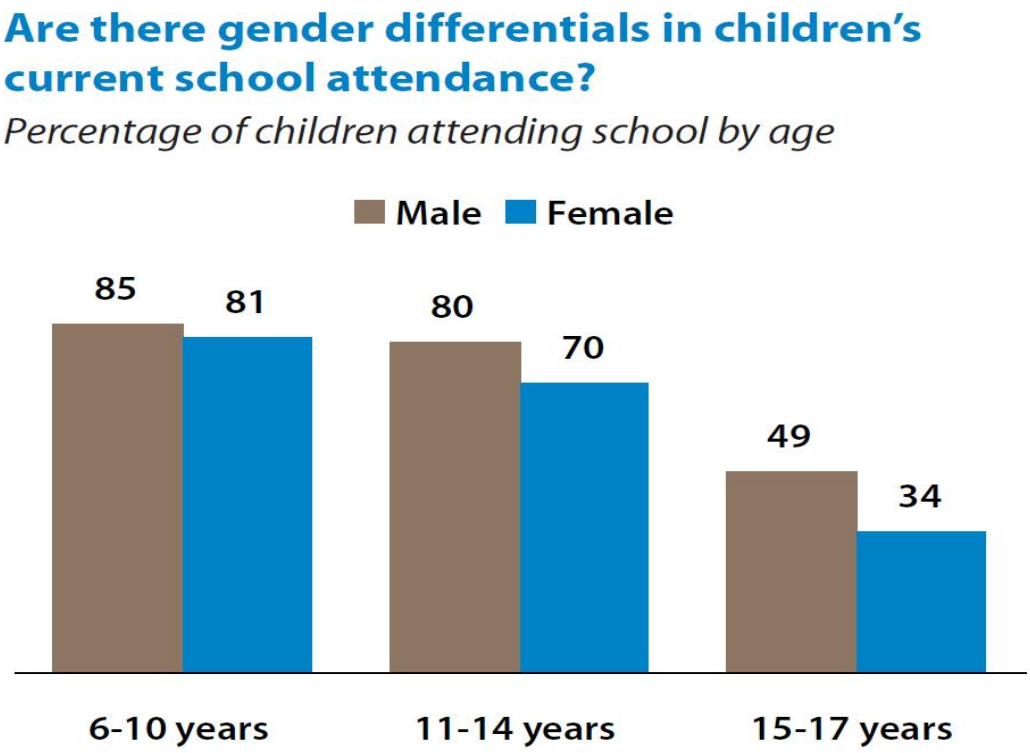

Figure 1.3: Percentage of children at school by age and gender from the NFHS-3 survey. Source thanks to USAID/MEASURE DHS [68]

was conducted by Pratham, India's largest non-government organization, in 2006. All of the 549 districts had 30 villages randomly selected, in each village 20 households were selected and all the children between 6 and 16 years interviewed. Of students who were in grade 5; 47% could not read a story from the grade 2 level and nearly 55% could not solve a simple division problem [77]. When police were called in to oversee exams in Utter Pradesh the pass rate dropped from 57% in 1991 to 14.7% in 1992 [77]. In 2005 a survey of teachers at rural Indian government schools was taken by visiting schools at random. It was found that on any given day approximately 25% of teachers were absent and of the teachers who were present only half were actually teaching [77].

As mentioned in Section 1.9, education and wealth are highly positively correlated. The higher someone is educated, the more prospects they have and the greater the expected rate of pay. This was shown clearly with the likelihood of an illiterate women working in the low-paying agricultural field 4.8 times higher than a women with secondary education and over 30 times higher than a women with graduate qualifications [71]. The number of children per women was also correlated with the mother's education: a higher education indicated a lower number of children [79]. In most states, women who had no education have two more children than women with 10 years or more education [79].

1.12 TB and nutrition in India

Malnutrition is the main risk factor for death, it is associated with a higher rate of infections, lower mental development, lower levels of achievement, lower activity levels, and lower levels of curiosity. Malnutrition leads to a weakened immune system (especially due to anaemia), higher risk of perinatal and prenatal death, being unable to breast feed, stunting and wasting [80]. Countries with high levels of malnourishment and child mortality significantly overlap countries with high levels of TB. This is expected as TB is associated with weakened immune systems, which in turn are associated with malnutrition.

India, Pakistan, and Bangladesh account for half of the world's malnourished children and also have high TB rates [81, 82]. India accounts for approximately 25% of all child deaths worldwide, more than in any other country. More than two million children aged less than five years died in India in 2006 [83]. A report from 2012 stated that the majority of the deaths of Indian children were preventable and were due to infectious diseases and malnutrition [84]. Another report from 2004 stated that 50% of Indian children may not have reached their physical and mental potential due to being malnourished, with another 20% being functionally impaired [85]. The proportion of undernourished children has increased from NFHS-2 to NFHS-3, as has the amount of anaemia and wasting [79]. The NFHS-3 survey found that of women 55% of women were anaemic and nearly 33% underweight [80]. The same survey found 24% of men were also anaemic. See Figure 1.4 for a the percentages of Indian women, men, and children with mild, moderate, and severe anaemia. In children, 70% were found to be anaemic, over half stunted, 42.7% wasted, 16.2% underweight (reported at over 40% elsewhere) and 11.9% with a low BMI for their age. There were 4.8% of children found to have severe anaemia, 33.6% had severe stunting, 17.3% had severe wasting, 5% were severely underweight and 4% had a severely low BMI.

The healthier and more nourished someone is the less likely they are to develop active TB. Low haemoglobin levels have been shown to increase the risk of TB recurrence [86] and someone who has TB and low haemoglobin levels is much more likely to be co-infected with HIV than someone with TB and higher haemoglobin levels [87]. There is a vicious malnourishment cycle which occurs between mothers and their children. The better nourished a mother is, the better nourished her child is expected to be; if a mother is malnourished her child is also expected to be malnourished [83]. If a child is malnourished, it is expected that as an adult, they will continue to be malnourished. To achieve a healthy birth, mothers are expected to gain around 10kg during the pregnancy. Most women is south Asia gain only 5kg which directly affects the child's birth weight [81].

The gender-specific discrimination in India motivated the gender-specific analysis in this thesis. Females children in India are discriminated against and are less likely to survive than their male counterparts [50]. One article stated that food was sometimes preferentially given to the males of the household, especially in

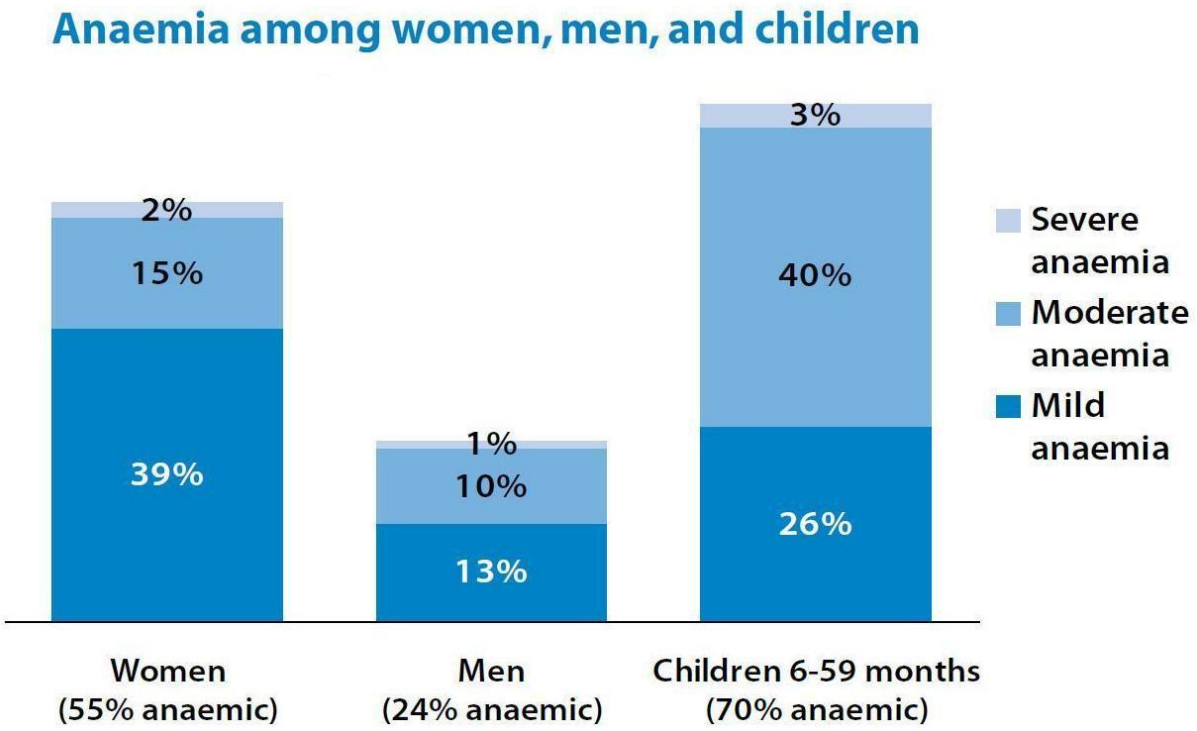

Figure 1.4: Respondents anaemia levels from NFHS-3.
Source thanks to USAID/MEASURE DHS [68]

poorer households [81]. Another article stated that spending on medical care was more than twice as much for male children than for female children, and that males were fed more milk and fats (nutritious but expensive) while females were fed more cereals (less nutritious but cheaper) [50] This article also looked at the child mortality rates near Delhi and reported the female mortality rate was higher than males in all age groups from one to five years [50]. It was discovered that a fourth born child was 1.5 more times likely to die if it was female instead of male [50]. It was also reported that of live first born children, 96 per 1,000 females will die before their fourth birthday compared to 127 per 1,000 males. For fourth and subsequent live born children 153 per 1,000 females born will die before their fourth birthday, compared to 99 males.

1.13 TB and HIV levels in India

Once someone is HIV positive their immune system is compromised and TB can easily infect them. The HIV prevalence in India according to the NFHS-3 survey is approximately 4% with an estimated 2.47 million (95% CI, 2.0 - 3.1 million) people with the disease [88]. Nagaland and Manipur had the highest HIV/AIDS prevalence, followed by Andhra Pradesh, then Maharashtra and Karnataka [89]. India accounts for approximately 7% of the global population of HIV (34 million) and is the biggest contributor after South

Africa and Nigeria [88]. Women accounted for roughly 40% of the HIV/AIDS population in India; the main vector of infection was their husbands [90]. Only 17% of women and 33% of men surveyed in the NFHS-3 survey had full understanding of HIV/AIDS [68]. Having full understanding meant if they knew that a healthy person could have HIV/AIDS, that mosquito bites and sharing food doesn't transmit HIV/AIDS and that condom use/fidelity help prevent it.

The size of an HIV epidemic is caused by two main populations; the high risk population that is likely to be infected, and the bridge population that transmits HIV to people who otherwise would not be affected. HIV is transmitted by unprotected sex, contaminated needles, and from mother to child. In India the high risk population for HIV/AIDS were sex workers and people who were injecting drugs [91]. Female sex workers had an elevated HIV prevalence; estimates of the rate of HIV in female sex workers varied widely with estimates as low as 2% in Tamil Nadu and higher than 30% in Maharashtra and Karnataka [88]. Injecting drugs was a significant HIV risk. An estimated 1.9% to 2.7% of adults from Manipur and Nagaland had injected drugs [91]. Of a sample of the drug users from Manipur and Nagaland, the majority were less than 19 years old, 75% were HIV positive, 66% were sexually active and only 3% reported using condoms [91]. The bridge population was clients of sex workers and long distance truckers. The rate of HIV increased from 0.3% to 0.7% for males who had had only one sexual partner, compared with males who had had more than one sexual partner [90]. Due to the interaction of these two groups a third group, the people who would be expected to be at a very low risk of HIV/AIDS from the high risk population are exposed due to the bridging population [92].

1.14 Questions addressed in this thesis

Using India as the country of interest and data from version three of India's National Family Health Survey *(NFHS)* this thesis seeks to answer the following questions:

- Which variables are the most appropriate to use in the exploratory data analysis as the 'risk factors of TB' to try and determine if someone has TB or not?
- Which of the risk factors of TB had the largest difference in distribution between respondents with and without TB, how significant were these differences, and in what sequence were the variables when ordered by decreasing difference in distribution?
- How well do nearest neighbour, classification tree, and logistic regression classify between respondents with and without TB? Which variables were used for the most parsimonious fit and is there a difference between the female and male population results?

- What is the most statistically likely visual representation of how the risk factors of TB are correlated with each other? What conclusions can we draw from the most statistically likely directed acyclic graph?

Prior to analysing the data, we expected the following:

- Wealth to be highly significant in all the analyses as TB has constantly been associated with poverty.
- Smoking, air pollution, and cooking smoke, to be significant due to TB affecting the lungs.
- Many of the other variables such as health, education and lifestyle to be directly related to wealth and to see higher TB rates in people with low BMI, low education, low quality housing and crowded living conditions.
- A significant difference between the female and male results due to the gender-specific effects on health and nutrition in India.

Chapter 2

Methods

There was a large amount of data available to analyse from the National Family Health Survey. The data was of high quality with few missing values. In general very little data cleaning and imputation was required. Section 2.1 provides an account of the data pre-processing applied to the NFHS dataset.

Two stages were completed for this thesis:

- Initial stage – Analysis of all the variables in the dataset which were potential risk factors according to previous studies. See Section 5.3 in the Appendix for a list of these variables
- Secondary stage – Detailed anaysis of a selection of the potential risk factors from the initial stage. See the start of Chapter 3 and Section 5.1 for a summary of these variables.

The initial stage involved an exploratory data analysis. All of the variables available which were determined from the literature review to be of potential significance and which were available in the NFHS dataset were included in the analysis. The purpose of this initial stage was to determine which variables to include in the detailed analysis in the second stage. Kolomogorov-Smirnov tests, permuted Kolomogorov-Smirnov tests, and chi-sqare tests were used. Using these tests the variables with significantly different distributions between the TB and non-TB data were shown both visually and numerically. Nearest neighbour and classification trees were also used to determine which variables, and combinations of variables, were the most reliable at separating the TB and non-TB data. Generalised linear models were also used to determine which variables were potentially useful in predicting TB cases. See Sections 2.2, 2.3, and 2.5 for details on these methods.

The second stage used only the risk factors which performed well in the initial analysis. In this stage logistic regression was used to find the most parsimonious model which predicted TB cases. The models were tested on their ability to predict TB cases, the number of incorrect TB cases predicted, and the ease

of obtaining the variables from a new respondent. The most statistically likely directed acyclic graphs were also found for both the female and male datasets. This allowed the most statistically likely combination of the correlations between the variables to be visualized. See Sections 2.5 and 2.6 for details on these methods.

2.1 Data preprocessing

The NFHS had several datasets available. The household dataset contained information on the household members and household characteristics. This was the main dataset used in the initial analysis. The female dataset contained information on the female household members, their health, and their knowledge of TB. The male dataset was similar to the female dataset and provided information on the male household members. The male and female datasets were used for the majority of the statistical analyses. Both the male and female dataset had information on everyone who had slept in the house the night before the survey. However, people who had slept in the house but were not household members were not questioned about their TB status, and subsequently did not have a TB status assigned to them. As this field was vital to the analysis it was not imputed and only household members were included in the analysis. This left 118,857 females and 72,607 males to be analysed. Not every variable had a value for every person. Depending on the number of values missing for each variable, some were imputed using fitted linear models, some were set to the mean of the group, some were kept as a factor of missing and non-applicable values, or calculated using known relationships.

For variables which were imputed using a fitted linear model the non-missing values for the variable were modelled. The following continuous variables had missing values and were imputed using a fitted linear model.

- Age of head of household
- Haemoglobin level.

The following variables had the largest proportion of missing values:

- Respondents who were not covered by an Anganwadi/ ICDS * centre for the variable 'year Anganwadi/ICDS centre started in respondents district'.
- No information on their HIV weight for the variable 'HIV weight for the respondents area'
- Respondents who were not married for the variable 'age of marriage'
- Respondents who never read a newspaper or magazine for the variable 'frequency of reading newspaper or magazine'.

These missing / Not Applicable values were kept as a separate group. There were generally very few missing and not applicable values for the other variables. * Anganwadi/ICDS centres provide health care, nutritional advise and supplements, basic medications, and vaccines for children. Their aim is to improve the health of their districts and they are focused on the poorer and less nourished groups - the same groups who are at high risk of TB

There were generally very few missing values from the NFHS-3 dataset. The majority of the variables had 99.9% of their data available. The following variables had more missing variables:

- Haemoglobin had many missing values since no respondent from Nagaland was measured. However, the total percentage of missing haemoglobin data was less than 10% for the female respondents and 13% for the male respondents.
- Respondent's weight and height were the next most likely to be missing. The total missing percentage for these variables was less than 5% for the female respondents just under 7% for the male respondents.
- The year /ICDS started had just under 2% of the female values missing and just under 3% of the male values missing.

2.2 TB and non-TB distribution functions

We tested the null hypothesis that the TB and non-TB distribution functions for a given variable were the same.

For a given variable, such as body mass index,
let $x_1, x_2, ...x_{m-1}, x_m \overset{iid}{\sim} F_{TB}$ and $x_{m+1}, x_{m+2}, ...x_{m+n-1}, x_{m+n} \overset{iid}{\sim} F_{nTB}$ be two sets of independent samples. We were interested in testing:

$$H0 : \text{distribution of TB respondents} = \text{distribution of non-TB respondents}$$
$$H1 : \text{distribution of TB respondents} \neq \text{distribution of non-TB respondents}$$

As multiple tests were used on the same dataset Bonferoni-corrected alpha levels and the Sidak correction were used to diminish the risk of finding an interaction to be significant by chance. The Bonferoni and Sidak corrected alpha levels are:

$$\text{Bonferoni } \alpha' = \left(\frac{\alpha}{k}\right) \tag{2.1}$$

$$\text{Sidak } \alpha' = 1-(1-\alpha)^{\frac{1}{k}} \tag{2.2}$$

where k is the number of tests performed.

The following three methods were used to test this null hypothesis: the Kolmogorov-Smirnov test, the Kolmogorov-Smirnov Permutation test, and Pearson's chi-squared test. We plotted the empirical distribution functions of F_{TB} and F_{nTB} using the point estimates to appreciate the differences and similarities in the distributions. We also used 95% confidence bands around the point estimates to visualise how different the distribution functions were.

2.2.1 Kolmogorov-Smirnov test

The Kolomogorov-Smirnov (KS) test is a standard non-parametric test used to find whether a sample is from the same distribution as another sample or distribution [93]. The KS test was used to obtain a quick overview of the differences in the TB and non-TB distributions for a given variable. The KS test should technically only be used for continuous variables. The categorical variables were tested with this method but the results have been marked with * to indicate the results need to be interpreted with caution. We explain the KS test in detail to help explain Sections 2.2.3 and 2.2.4.

The empirical cumulative distribution function is defined by:

$$\hat{F}(x) = \begin{cases} 0 & : \ x < x_{(1)} \\ \frac{r}{n} & : \ x_{(r)} \leq x < x_{(r+1)} \\ 1 & : \ x_{(n)} \leq x. \end{cases} \tag{2.3}$$

where $x_{(1)}$ is the smallest x value, $x_{(n)}$ the largest, and the values of $x_{(r)}$ are ordered by increasing size. $\hat{F}(x)$ is the empirical probability of a value equal to or smaller than any given x occurring from the data given. The KS test statistic is the maximal distance between the Empirical Cumulative distribution Function (ECDF) of $\hat{F}(TB)$ and $non\hat{-}TB(x)$.

The further the maximal distance between two ECDF were, the more likely they were from two different distributions. The distance between two ECDF were defined as:

$$t_{ks} = \max_{x} \mid \hat{F}_{TB}(x) - \hat{F}_{nTB}(x) \mid \tag{2.4}$$

The test was conducted using the KStest2 function in R.

The KS test is thought to be accurate as long as $\frac{n_1 * n_2}{n_1 + n_2} \geq 4$. The female data set had $\frac{472 * 118857}{472 + 118857}$, giving 470 and the male data set had $\frac{443 * 72164}{443 + 72164}$, giving 440. This meant this test was applicable to both the female and male NFHS datasets to determine if there was a significant difference in distribution between the TB and non-TB respondents for a continuous variable.

2.2.2 Pearsons chi-square test of independence

The Pearsons chi-square test is a standard non-parametric test of independence for categorical or discrete data. The chi-square test compares how similar the data is to the distribution expected under the hypothesis of independence. When two categorical variables are independent, the probability of each cell in a contingency table is the product of the sum of the cells column and row, divided by the total sum of the table, as in Equation 2.5

$$P_{ij} = P_{i.} * P_{.j} \tag{2.5}$$

The chi square statistic for the observed table is the sum of the squared difference between each cells observed and expected value, divided by the expected value.

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.6}$$

Where O_i is the observed probability, x_{ij}/n, E_i is the expected probability of a point occurring at the value P_{ij}, and n, m are the dimensions of the dataset.

2.2.3 Kolmogorov-Smirnov permutation tests

Permutation test is a standard non-parametric exact test. It can be used to determine if two groups of data came from the same distribution or not. This is usually done by comparing the difference between the means of both groups. Permutation tests were used to test whether the KS test statistic between the TB and non-TB distributions were significant. Since the KS test above is not appropriate for discrete data, we used the permutation test with the KS test statistic on all the variables (both continuous and discrete). This allowed an easy comparison of the KS test statistic for all the variables. We wanted to rank all of the variables by their KS test statistic, as they were all on the universal scale of [0,1] this was possible.

The maximal distance between the TB and non-TB CDF was calculated as in (2.4) for each variable. The TB and non-TB data was then permuted. The first nTB variables were assigned to the group $\text{TB}_{permuted}$ and the rest to the group $\text{nonTB}_{permuted}$. By setting m to the count of the TB variables, the size of the TB and non-TB groups was constant. The data was repeatedly permuted and the maximal distance between the $\text{TB}_{permuted}$ and $\text{nonTB}_{permuted}$ CDF was calculated and stored each time. After a large number of permutations the ECDF of the maximal distances was plotted. The p-value was estimated from the proportion of permuted t_{ks} values that were at least as high as that of the observed t_{ks} value.

2.2.4 Exploratory data analysis of ECDFs confidence band width

The ECDFs $\hat{F}_{nTB}$ and $\hat{F}_{TB}$ are non-parametric point estimates for each variable of interest (whether continuous or discrete). We visually explored the distributions between the TB and non-TB respondents for each variable by plotting the EDF and their confidence bands. This exploratory data analysis was applicable to both continuous and categorical data. It was mainly used as a method of visualising the point estimate and confidence bands of the distributions.

As the number of data points in a distribution function increases, the expectation nears the true expectation and the variance drops closer to zero. The confidence interval was found at all points for both the TB and non-TB variables to create a confidence band. The confidence interval for a large number of points will be narrower than a confidence interval for a small number of points. This was seen with the TB confidence intervals (with few data points) being wider than the non-TB (with many data points) confidence intervals.

The (1-α) confidence interval was found using the formula:

$$\epsilon(n) = \sqrt{\frac{1}{2 * n} * \log \frac{2}{\alpha}}. \tag{2.7}$$

where width was the vertical distance from the CDF point to the confidence interval, n the number of points in the sample, and alpha the level of testing.

2.3 Nearest neighbour classification

Nearest neighbour classifier was used to partition the data and classify the TB status of respondents. Both one and two dimensional cases were looked at. For the nearest neighbour analysis the household dataset was used, however the female and male data was still analysed separately. The one dimensional case looked at individual variables and classified the TB status of the respondent based on the single variable. The

two dimensional case looked at every combination of the variables and classified the TB status based on each pair of variables. Higher dimensional cases were only looked at briefly due to computational time.

The optimal nearest neighbour size was the size that correctly classified the most data without over-fitting the data. This was accomplished by re-substitution and cross-validation. Re-substitution error uses the same data in both the creating and the testing of model which can give a false, overly-optimistic, impression of the classifier. Re-substitution error gives a lower error rate than cross-validation and does not generalise to similar error levels when previously unseen data is classified. Cross-validation methods test how well the results generalise to previously unseen data. The data is split into k segments and each k-th segment of the data sequentially held back from the analysis. This analysis used 10-fold cross validation (where k = 10), the Euclidean distance to calculate distance between points, and the class of the nearest point in the case of a tie. How accurately the analysis predicted the class of the data points held back gave an estimate of how well the model was likely to accurately predict new data.

2.4 Classification trees

Classification trees, a data mining technique, were used to partition the dataset into TB and non-TB regions. Classification trees use the variables to partition the data until the areas in each partitioned region are of one class. By sequentially partitioning the data regions of high and low TB density were found. We set the cost of misclassifying a TB respondent much higher than the cost of misclassifying a non-TB respondent (0.99 to 0.01). This was so a higher emphasis was placed on correctly classifying respondents who had TB. When no cost was specified, the optimal decision tree classified all respondents as not having TB. We also looked at even higher costs, however this also led to all respondents being classified to one class.

Due to the complexity of the classification tree needing to have each partitioned region of one class we allow some impurity in the nodes. The re-substitution and cross-validation error as described in Section 2.3 were found as we increased the number of partitioned regions. This indicated how well the classification tree classified and generalised to unknown data. The number of partitions with the lowest cross-validation error was the favoured partition level.

2.5 Logistic regression

Logistic regression was used on the NFHS dataset to model the probability of respondents having TB. The most parsimonious model was found so the most relevant risk factors could be identified. A logistic

regression model with parameters which are easily identifiable could allow health workers to evaluate a new patients risk of TB quickly and at low cost.

Logistic regression can be used when the response X: $x_1, x_2, x_3, ...x_{n-2}, x_{n-1}, x_n$ to be modelled is a binary factor.

The logistic regression model finds the best B values for the formula:

$$P = \frac{\exp(A + \sum_{i=1}^{n} B_i x_i)}{1 + \exp(A + \sum_{i=1}^{n} B_i x_i)}, \tag{2.8}$$

where P was the probability of the event "having TB" occurring.

The link function:

$$g_{(x)} = ln\left(\frac{P(x)}{1 - P(x)}\right) = A + \sum_{i=1}^{n} B_i x_i \tag{2.9}$$

called the logit link, was also used.

The most parsimonious model was the model desired. Only one of highly correlated variables were included in the model to avoid collinearity problems. Variables which were factors were treated as such while continuous variables were manipulated to be as normally distributed as possible. Variables were initially removed from the full model using backwards selection and the Akaike information criterion *(AIC)*.

$$AIC = -2 * \log L + 2 * k \tag{2.10}$$

Where L was the likelihood and k the number of parameters in the model. The lower the AIC the better the model was: $-2 * logL$ gave a larger, negative number with larger likelihoods. $2 * k$ gave a smaller, positive number for models with few parameters The AIC formula found a model which was a compromise between fitting well to the data and having few parameters.

Due to stopping criterion, stepwise selection stopped before the most parsimonious model. Anova testing was used to test if dropping or adding a variable significantly improved the model. At each step only one variable was removed to prevent removing a significant variable accidentally. Variables were not removed if they had a higher-order interaction term in the model. For each model comparison, if the p-value was less than 0.01 the larger model was considered better, else there was no significant difference and the smaller model chosen.

The formula for the ANOVA tables is:

$$F = \frac{\left(\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{p_2 - p_1}\right)}{\left(\frac{\mathrm{RSS}_2}{n - p_2}\right)}, \tag{2.11}$$

where model 2 is nested in model 1, RSS is the residual sum of squares, p the number of parameters in the model. The p-value of F is calculated from the F-distribution; if it is not significant, there is no significant difference between the models and the nested model can be accepted.

Cross validation, as described previously in Section 2.3, was used to test how well the models predicted previously unseen data.

2.6 Statistically likely directed acyclic graphs

A directed acyclic graph is comprised of nodes and directed edges (arrows) between pairs of nodes. Each node represents a variable, and each edge represents an association between the variables. For example in Figure 2.1 the DAG has two nodes which represent the variables wealth and TB. The directed edge going from wealth to TB indicates that there is an expected association between someone's wealth and their TB status. If the direction of the edges are followed it is not possible to go from the TB node to the wealth node. If it was possible to return to a node which had previously been visited the graph would be cyclic. As we go from the wealth node to the TB node, but not from the TB node to the wealth node, we call the graph acyclic.

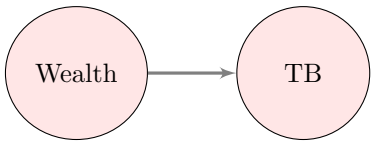

Figure 2.1: Example directed acyclic graph - Wealth and TB.

DAGs are used to visualise the conditional dependence and independence structure between the variables. Each node in a DAG represents a variable. Each directed edge indicates that the state of the variable in the first node affected the probability of the variable in the second node [94, 95]. Given a collection of DAGs for a set of nodes we can select the most likely and parsimonious DAG using Bayes Information Criteria (BIC). This gave the most statistically likely visual representation of the conditional dependence and independence structure among the variables.

The number of possible DAGs on n nodes is given by a_n in formula (2.12) [96]. The list below shows the number of possible DAGs , a_n, for n increasing from 1 to 30.

$$a_n = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}. \tag{2.12}$$

1. 1.00E+00
2. 3.00E+00
3. 2.50E+01
4. 5.43E+02
5. 2.93E+04
6. 3.78E+06
7. 1.14E+09
8. 7.84E+11
9. 1.21E+15
10. 4.18E+18
11. 3.16E+22
12. 5.22E+26
13. 1.87E+31
14. 1.44E+36
15. 2.38E+41
16. 8.38E+46
17. 6.27E+52
18. 9.94E+58
19. 3.33E+65
20. 2.34E+72
21. 3.47E+79
22. 1.08E+87
23. 6.97E+94
24. 9.44E+102
25. 2.66E+111
26. 1.56E+120
27. 1.90E+129
28. 4.79E+138
29. 2.51E+148
30. 2.71E+158

As the NFHS dataset used in the thesis had around 25 variables analysed, analysing every possible DAG was not feasible. The more correlated two variables are the more likely it is that they will share a directed edge. Using Pearson's correlation test and Pearson's chi-square test we ranked pairs of nodes on the basis of how correlated they were. Edges between pairs of variables with low correlation were excluded from the analysis. This allowed us to focus on highly correlated variables and reduced the number of DAGs investigated.

Bayes Information Criteria (BIC) was used to determine which DAG was the most likely. AIC as described in (2.10) and the Euclidean distances between the probability of the full model and the DAG occurring were also investigated.

The formula for the BIC from a given DAG is:

$$BIC = -2 * \log L + k \log(n) \tag{2.13}$$

where L was the maximum likelihood value of the DAG, k the degrees of freedom, and n the sample size.

The maximum likelihood value was calculated from the joint probability of the DAG and the DAGs degrees of freedom. Each node could take multiple values of the variable it represented. For instance the node 'Location' was comprised of the Northern, Eastern, Central, and Southern states and the node 'House Type' consisted of Kaccha, Semi-Pucca and Pucca houses. The joint probability of the variables were calculated from the conditional probability structure given by the directed edges of the DAG.

As an illustrative example we consider four DAGs over just three nodes X1, X2, and X3. We look at four different DAGs on these three nodes and calculate the BIC for each DAG. The DAG with the smallest DAG is chosen as the most likely. The four DAGs are shown in Figures 2.2 to 2.5. Table 2.1 presents some example data from these nodes and calculates the joint probability and maximum likelihood value

for them. Table 2.2 calculates the BIC and AIC value for the DAGs. DAG C is shown to have the smallest BIC and AIC which makes it the most statistically likely DAG.

The joint probability and degrees of freedom for each DAG were calculated as in Equations (2.14) and (2.14). The joint probability is:

$$\begin{aligned}
\text{DAG A: } f(X_1, X_2, X_3) &= f(X_1)f(X_2)f(X_3) \\
\text{DAG B: } f(X_1, X_2, X_3) &= f(X_1)f(X_3|X_1)f(X_2|X_3) \\
\text{DAG C: } f(X_1, X_2, X_3) &= f(X_1)f(X_2)f(X_3|X_1, X_2) \\
\text{DAG D: } f(X_1, X_2, X_3) &= f(X_1)f(X_3)f(X_2|X_1, X_3)
\end{aligned}$$

and the degrees of freedom:

$$\begin{aligned}
\text{DAG A: } DF(X_1, X_2, X_3) &= (N_{X_1} - 1) + (N_{X_2} - 1) + (N_{X_3} - 1) \\
\text{DAG B: } DF(X_1, X_2, X_3) &= (N_{X_1} - 1) + (N_{X_1}(N_{X_3} - 1)) + N_{X_3}(N_{X_2} - 1)) \\
\text{DAG C: } DF(X_1, X_2, X_3) &= (N_{X_1} - 1) + (N_{X_2} - 1) + N_{X_1}N_{X_2}(N_{X_3} - 1) \\
\text{DAG D: } DF(X_1, X_2, X_3) &= (N_{X_1} - 1) + (N_{X_3} - 1) + N_{X_1}N_{X_2}(N_{X_3} - 1)
\end{aligned}$$

where N_i is the number of categories for variable X_i.

				Full Model		DAG (A)		DAG (B)		—bf DAG (C)		DAG (D)	
X1	X2	X3	Count	Prob	ℓ	Prob	ℓ	Prob	ℓ	Prob	ℓ	Prob	ℓ
a	c	e	4	0.03	-5.94	0.11	-3.89	0.05	-5.32	0.03	-6.30	0.07	-4.54
a	d	e	8	0.07	-9.47	0.11	-7.56	0.05	-10.30	0.08	-8.61	0.15	-6.67
a	d	f	27	0.22	-17.68	0.10	-27.04	0.16	-21.40	0.18	-20.14	0.14	-23.13
a	d	f	13	0.11	-12.64	0.11	-12.65	0.17	-10.11	0.14	-11.25	0.07	-15.26
b	c	e	26	0.21	-17.46	0.14	-21.94	0.20	-18.23	0.26	-15.31	0.15	-21.34
b	d	e	25	0.20	-17.21	0.15	-20.38	0.22	-16.49	0.18	-18.84	0.15	-20.95
b	c	f	2	0.02	-3.57	0.13	-1.74	0.08	-2.23	0.02	-3.41	0.03	-3.07
b	d	f	17	0.14	-14.55	0.14	-14.34	0.08	-18.72	0.12	-15.66	0.25	-10.29
sum			122	1.00	-98.52	1.00	-109.54	1.00	-102.81	1.00	-99.52	1.00	-105.24

Notes:
Prob is the probability of the combination of X2, X3, X4 occurring given the DAG structure
ℓ is the maximum log likelihood of the DAG
DAG (A) is for the DAG where X1, X2, and X3 are independent
DAG (B) is for the DAG where X1 affects X3, and X4 affects X3
DAG (C) is for the DAG where X1 and X2 affect X3
DAG (D) is for the DAG where X1 and X3 affect X2

Table 2.1: Example calculations for nodes X1, X2, X3 to determine the maximum likelihood value for DAGs A, B, C, and D.

	MLV	DF	Penalty	BIC	AIC	Euclidean
Full Model	-98.52	7	7.30	106	106	
DAG A	-109.54	3	3.13	113	113	0.04
DAG B	-102.81	5	5.22	108	108	0.02
DAG C	-199.52	4	4.17	**104**	**104**	0.01
DAG D	-105.24	4	4.17	109	109	0.04

Table 2.2: Calculating the BIC, AIC and Euclidean distance for DAGs A, B, C, and D.

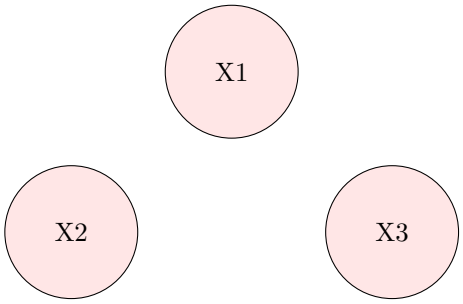

Figure 2.2: Example directed acyclic graph - A.

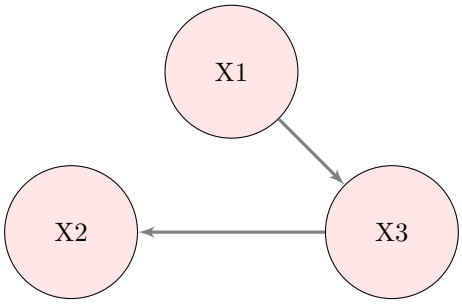

Figure 2.3: Example directed acyclic graph - B.

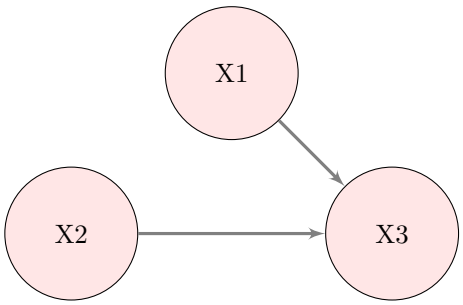

Figure 2.4: Example directed acyclic graph - C.

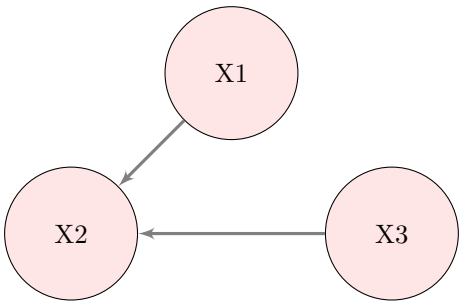

Figure 2.5: Example directed acyclic graph - D.

2.6.1 Moral graphs

Moral graphs can be found from directed acyclic graphs and are helpful in visualizing the data. Moral graphs are simpler than directed acyclic graphs in algorithm construction techniques. While the moral graphs have not been investigated in depth in this thesis there is potential for further investigation using them.

A DAG can be converted to a moral graph by adding undirected edge between all pairs or nodes a and b where at least one of the following hold:

- An edge from a to b exists
- An edge from b to a exists
- An edge from a to c, and b to c exists; where c is a third node common to both a and b.

For Figure 2.2 the moral graph would be the same as the DAG and for Figure 2.3 the moral graph would have the same structure as the DAG but the directed edges would be undirected. The moral graph would no longer have the same structure as the DAG for Figures 2.4 and 2.5. As Figure 2.4 has edges from X1 and X2 to X3, the moral graph would have undirected edges where the directed edges were, as well as an undirected edge between X1 and X2. Similarly, Figure 2.5 has edges from X1 and X3 to X2. The moral graph for Figure 2.5 would be identical to the moral graph from Figure 2.4. It would have undirected edges between X1 and X2, X3 and X2, and X1 and X3.

2.7 Variable analysis

The initial analysis included 86 variables for the female dataset and 103 variables for the male dataset. These variables, found from the literature review, were thought to be potentially significant in determining the TB status of a respondent. For a full list of the variables used in the initial analysis see Section 5.3. All of the initial analysis was conducted separately for both the female and male datasets; the results are displayed by gender.

Only a selection of the variables from the initial analysis were retained for the secondary analysis. Variables were discarded if other, highly correlated variables were included, and if the variable had not been shown to be a good indicator of TB. Some variables were retained as they were thought to both be significant variables to include in the model, others were retained out of interest. For a list of all the variables initially

investigated please see Sections 5.3 and 5.3 in the Appendix. For a summary of the variables used in the secondary analysis please see the Section 5.1 in the Appendix. The variables used in the secondary analysis were:

- **Wealth**
  - Wealth Index factor score
  - Standard of Living Index
  - House type
  - Cooking done under a chimney
  - Frequency of watching television
  - Food cooked on stove, chullah, open fire
  - City/Town/Countryside
  - If Anganwadi/ICDS in area, year Anganwadi/ICDS centre began
- **Health and Nutrition**
  - Body Mass Index
  - Haemoglobin level adjusted for altitude
  - National HIV weight
- **Education**
  - Keep secret when family member gets TB
  - Think tuberculosis can be cured
  - Education in single years
  - Partner's highest year of education
  - Frequency of reading newspaper or magazine
- **Household Information**
  - State (grouped)
  - Age of household head
  - Current age - respondent

- Age at first marriage
- Age at first intercourse
- Total number of sexual partners
- Number of household members
- Total children ever born
- Number of living children
- Sons at home
- Daughters at home

[This page intentionally left blank]

Chapter 3

Results

The initial analysis included distribution function tests as described in Section 2.2, nearest neighbour analysis as described in Section 2.3 and generalised linear modelling as described in Section 2.5. The results of these tests are in Sections 3.2, 3.3, and 5.6. The secondary analysis included generalised linear model analysis and directed acyclic graphs as described in Sections 2.5 and 2.6. The results of the generalised linear model analysis are in Section 3.6 and the results of the directed acyclic graphs are shown in Section 3.7. A brief summary of some of the variables included in the analysis are in Section 3.1 with a more detailed summary in Section 5.1.

3.1 Brief summary of TB distribution among variables

A summary of some of the variables used in this analysis are shown in Tables 3.1 to 3.7. For the categorical variables the count and percentage of the female and male respondents in each category is shown. For instance in the first table we show that there were 118,385 females who did not have TB, which was 99.6% of the female dataset. For the continuous variables the minimum, 1st quartile, median, mean, 3rd quartile, maximum, and standard deviation are shown. Tables 3.8 and 3.9 provide a numerical summary of the respondents TB status, age, gender, and education. While Table 3.10 provides a summary of the number of persons who reported having TB in each state. Figures 3.1 to 3.3 provide some visual summaries of the data. We analysed the results from 118,857 female respondents (62% of the dataset) and 72,607 male respondents (38% of the dataset). Of the 191,464 respondents analysed 915 had TB (0.5%).

	non-TB	TB
Female	118,385 (99.6%)	472 (0.4%)
Male	72,164(99.4%)	443 (0.6%)

Table 3.1: Count and percentages for the female and male respondents with and without TB

	northern	central	eastern	southern
Female	22,125 (19%)	28,916 (24%)	34,089 (29%)	33,727 (28%)
Male	8,251 (11%)	17,982 (25%)	17,337 (24%)	29,037 (40%)

Table 3.2: Count and percentages for the female and male respondents living in each region

	capital, large city	small city	town	countryside
Female	25,499 (21%)	8,878 (7%)	20,431 (17%)	64,049 (54%)
Male	19,997 (28%)	4,905 (7%)	12,565 (17%)	35,140 (48%)

Table 3.3: Count and percentages for the female and male respondents living in cities, town, or the countryside

	Min	1st Qu	Median	Mean	3rd Qu	Max	Std.dev
Female	15	21	28	29.37	37	49	0.10
Male	15	22	30	34.01	40	54	0.11

Table 3.4: Summary statistics for the female and male respondent's age

	Min	1st Qu	Median	Mean	3rd Qu	Max	Std.dev
Female	-1.75	-0.76	0.02	0.08	0.89	2.37	0.99
Male	-1.74	-0.70	0.04	0.09	0.85	2.40	0.96

Table 3.5: Summary statistics for the female and male respondent's wealth index factor score

	Min	1st Qu	Median	Mean	3rd Qu	Max	Std.dev
Female	5.40	18.24	20.32	21.06	23.07	68.03	4.02
Male	6.06	18.32	20.24	20.77	22.67	74.77	3.50

Table 3.6: Summary statistics for the female and male respondent's body mass index (BMI)

	Min	1st Qu	Median	Mean	3rd Qu	Max	Std.dev
Female	20	108	118	116.4	127	229	0.17
Male	22	128	141	139.2	152	238	0.19

Table 3.7: Summary statistics for the female and male respondent's haemoglobin levels

Age	Gender	Education							TOTAL
		none	1–5 yrs	6-8 yrs	9 yrs	10 – 11 years	12 –14 yrs	≥ 15 yrs	
15 – 24	Male	16	9	18	12	7	4	-	66
25 – 34	Male	25	18	28	15	10	3	5	104
35 – 44	Male	52	39	28	10	10	10	3	152
45 – 54	Male	58	21	18	9	6	3	6	121
15 – 24	Female	41	25	17	11	12	10	5	121
25 – 34	Female	93	27	23	14	8	5	6	176
35 – 44	Female	94	12	7	4	5	1	2	125
45 – 54	Female	32	9	7	1	-	-	1	50
	TOTAL	411	160	146	76	58	36	28	915

Table 3.8: Summary of the count of TB respondents by education level, TB status, age, and gender

Age	Gender	Education							TOTAL
		none	1–5 yrs	6-8 yrs	9 yrs	10 – 11 years	12 –14 yrs	≥ 15 yrs	
15 – 24	Male	1,893	2,964	5,297	4,355	5,050	3,775	1,367	24,701
25 – 34	Male	2,801	2,824	3,409	2,480	2,926	2,458	3,225	20,123
34 –44	Male	3,491	3,024	2,507	1,682	2,039	1,477	2,293	16,513
45 – 54	Male	2,319	2,189	1,758	868	1,501	803	1,389	10,827
15 – 24	Female	8,196	5,681	8,529	6,006	7,238	5,546	2,215	43,411
25 – 34	Female	12,546	4,960	5,297	2,894	3,935	2,838	4,177	36,647
33 –44	Female	12,721	4,541	3,560	1,552	2,640	1,437	2,240	28,691
45 – 54	Female	4,719	1,608	1,108	417	776	352	656	9,636
	TOTAL	48,686	27,791	31,465	20,254	26,105	18,686	17,562	190,549

Table 3.9: Summary of the count of non-TB respondents by education level, TB status, age, and gender

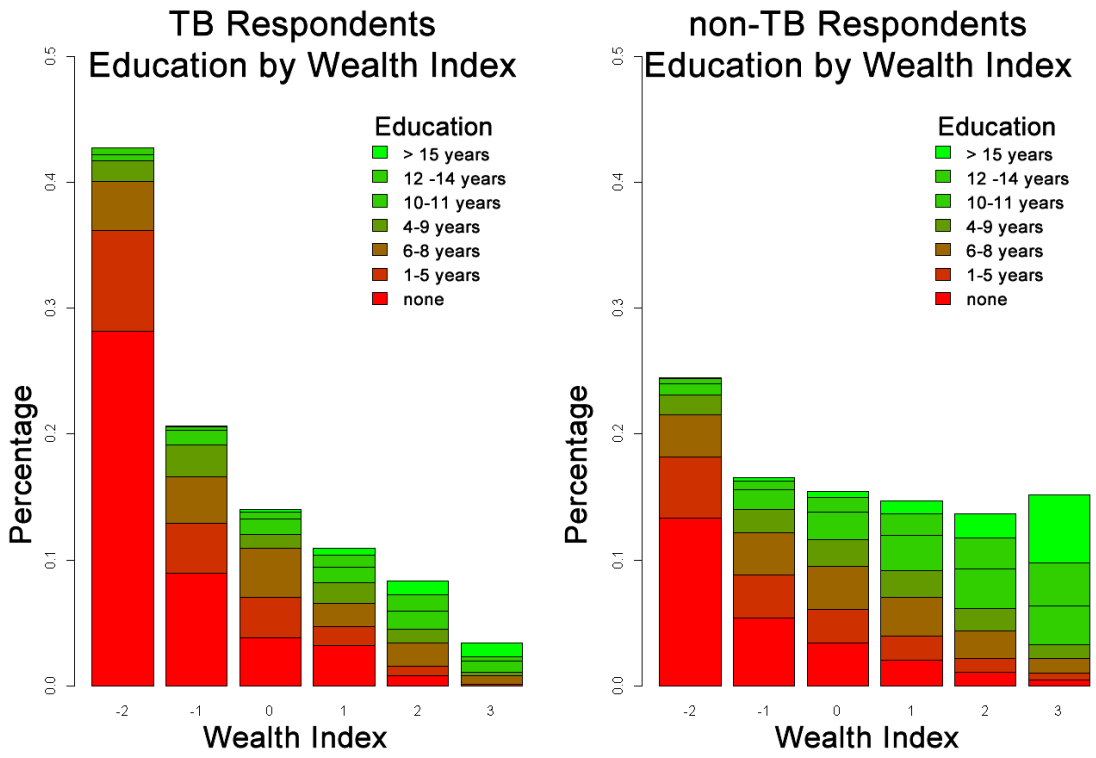

Figure 3.1: Respondents educational level by wealth bracket and TB status.

State	Number of persons surveyed who reported having TB	Total number of persons surveyed	Prevalence of TB (per 10,000) in surveyed population	Location
Jammu and Kashmir	3	4148	7	North
Himachal Pradesh	9	4134	22	North
Punjab	7	4862	14	North
Uttaranchal	13	3808	34	North
Haryana	14	3682	38	North
Delhi	18	4680	38	North
Rajasthan	25	5062	49	North
Bihar	40	4570	88	East
Sikkim	16	2865	56	East
Arunachal Pradesh	27	2236	121	East
Nagaland	57	7707	74	East
Manipur	53	8307	64	East
Mizoram	15	2437	62	East
Tripura	10	2509	40	East
Meghalaya	21	2812	75	East
Assam	34	5093	67	East
West Bengal	60	9082	66	East
Jharkhand	23	3808	60	East
Uttar Pradesh	146	22463	65	Central
Orissa	23	5833	39	Central
Chhattisgarh	18	4956	36	Central
Madhya Pradesh	35	8689	40	Central
Gujarat	34	4957	69	Central
Maharashtra	85	17517	49	South
Andhra Pradesh	40	13871	29	South
Karnataka	13	11116	12	South
Goa	7	4434	16	South
Kerala	12	4426	27	South
Tamil Nadu	57	11400	50	South
Total	915	191,464	48	

Table 3.10: Number and percentage of NFHS survey respondents with TB, by state.

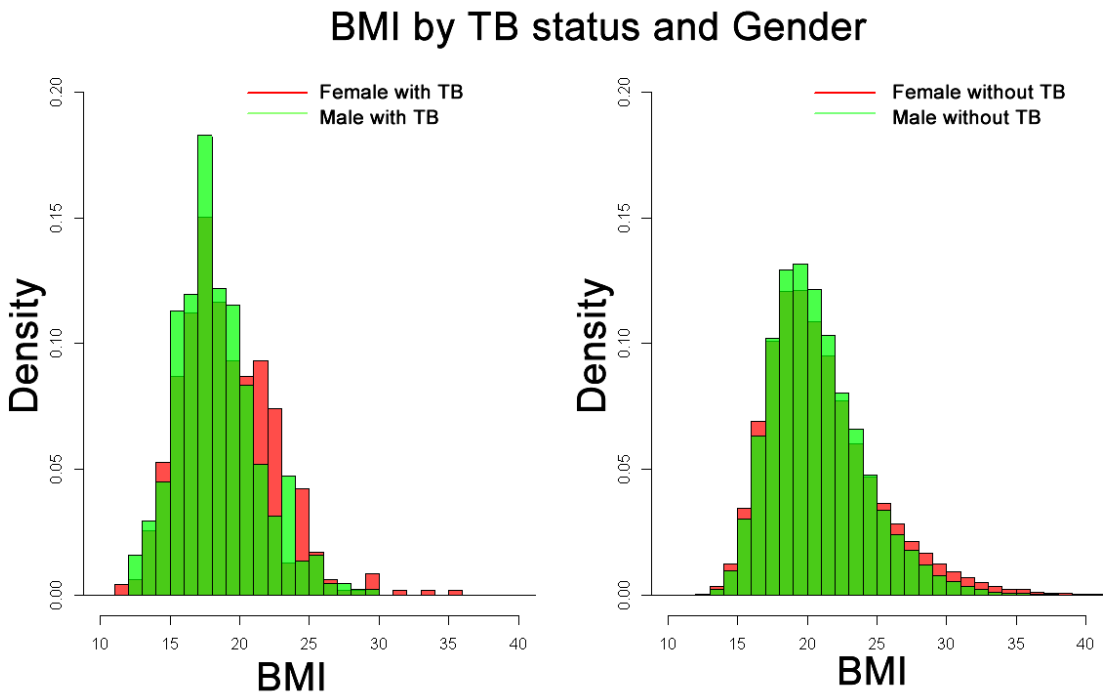

Figure 3.2: Respondents BMI split by gender and TB status.

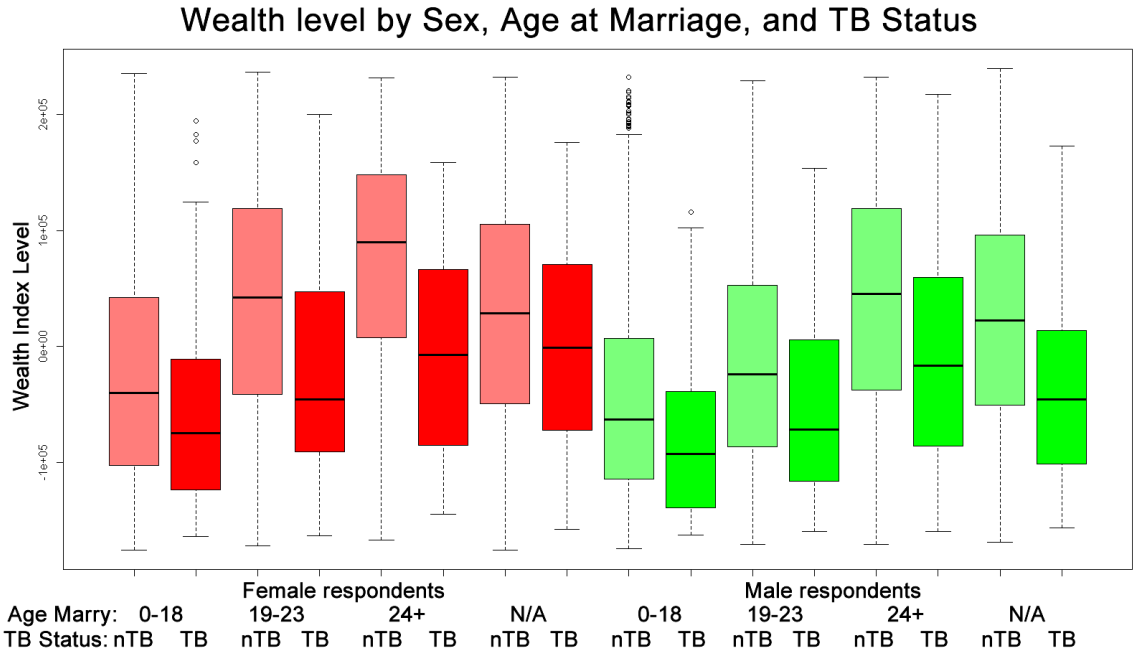

Figure 3.3: Respondents wealth index levels split by gender, age at marriage, and TB status.

3.2 TB and non-TB distribution functions

The TB and non-TB distribution functions were used to obtain an overview of the data as discussed in Section 2.2. These tests are described in detail in Sections 2.2.1, 2.2.2, and 2.2.3. The results from the KS test, Pearsons chi-squared test, and the permuted KS test were similar. This was expected and gives confidence to our test results. A summary of the distribution function testing is shown in Table 3.11. For a more detailed summary please see Tables 5.3 and 5.4 in the Appendix.

A selection of the variables have been shown with their ECDF plots (as described in 2.2.4) and histograms in Figures 3.4 to 3.8. The ECDF plots and histograms have been used as a visual aid to illustrate the differences between the distributions. When the KS permutation test was not significant at the 1% level the confidence bands on the ECDF for the distributions overlapped and the histograms were approximately equal. When the KS permuation test was significant at the 1% level the confidence bands on the ECDF did not overlap at all points. Differences between the TB and non-TB data were clearly shown for some variables in both the female and male datasets. The size of the TB population was much less than the non-TB population; both the female and male datasets had less than 500 respondents who reported as having TB. In contrast the female dataset had 100,000 respondents who reported not having TB and the male dataset had 73,000. Due to the differences in the TB and non-TB population sizes, the confidence band for the non-TB population was smaller than that for the TB population.

Wealth, education and age had significantly different distributions at the 1% level of the KS permutation test. This is shown in the CDF plots with the confidence bands for the TB and non-TB populations not overlapping. The histograms and density function also show two different distributions. The TB population had a significantly higher proportion of its respondents in the lower wealth brackets, with low or no education, and slightly older, than the non-TB population. As expected the female population was shown to have less education than the male population. As wealth was a continuous variable and education and age categorical, the CDF for wealth was smooth but not for education and age. The female and male CDF were not the same with a gender-specific effect being shown here. For respondents age the CDF for the female and male non-TB respondents showed a similar distribution with a convex positive unipolar line. However the respondents CDF for female TB population was close to linear while the CDF for the male population showed a concave positive unipolar line. The histogram for respondents age also shows a slight peaking every five years. This is most likely to due the respondents or surveyors rounding the ages to the nearest five. The age range for the female respondents was 15 to 49 years, while for the male respondents this was 15 to 54 years. Figures 3.4, 3.5, and 3.6 present the CDF and histograms/density function for these variables.

Haemoglobin level was significant in the KS permutation test at the 1% level for the female dataset, however it was not significant in all of the other tests. The TB and non-TB distributions were more different in the male dataset than in the female dataset. This was shown by the increased difference between the TB and non-TB distributions for the male dataset. The TB population was shown to have slightly lower haemoglobin levels than the non-TB population. The female population was also shown to have lower haemoglobin levels than the male population.

Plots where the TB and non-TB distributions overlapped were potentially from the same distribution. The CDF of the number of household members shows the intervals for TB and non-TB overlap in both the female and male datasets. In the histogram both the TB and non-TB distributions are in the same location and have the same shape. A difference between the TB and non-TB population was not seen in Figure 3.8.

The main results from this section are that the variables with the most different distribution functions between the TB and non-TB populations were:

- If anyone in the respondent's household suffered from TB
- The respondent's BMI
- The respondent's education
- The respondent's wealth index factor score
- The number of children born to the respondent.

Variables ranked by average t_{ks} for female and male data	Female data t_{ks}	TB mean	non-TB mean	Male data t_{ks}	TB mean	non-TB mean
Respondent suffers from TB	1	1.00	0.00	1	1.00	0.00
Any household resident has TB	0.983	1.00	0.02	0.984	1.00	0.02
Body mass index	0.232	18.99	21.06	0.313	18.44	20.78
Education in single years	0.245	3.38	6.08	0.28	4.93	7.98
Wealth index factor score	0.216	-396.45	78.39	0.255	-462.32	96.17
Number of children born	0.175	3.01	2.10	0.253	3.06	1.71
House type	0.181	2.16	2.03	0.23	2.14	2.47
Freq. watching tv	0.177	1.43	1.92	0.234	1.59	2.15
Freq reading newspaper/magazine	0.183	0.49	0.93	0.223	1.07	1.71
Number of living children	0.159	2.53	1.89	0.237	2.38	1.47
Current age of respondent	0.124	31.50	29.36	0.249	36.62	30.97
Number of sons at home	0.089	1.11	0.88	0.176	1.08	0.70
Haemoglobin level	0.076	113.82	116.45	0.182	131.07	139.25
Number of daughters at home	0.112	1.02	0.74	0.145	0.90	0.60
State *	0.099	*	*	0.13	*	*
Age at first marriage	0.109	17.04	17.98	0.113	21.57	22.74
City/Town/Countryside	0.067	2.16	2.03	0.14	2.21	1.86
Year Anganwadi began **	0.091	1993.60	1991.97	0.101	1992.93	1991.59
Age of household head	0.091	45.17	47.34	0.098	44.77	45.80
HIV weight	0.079	1,063	989	0.104	827	998
Cooking done under a chimney **	0.067	0.05	0.11	N/A	N/A	N/A
Number sexual partners	0.026	0.89	0.77	0.068	1.61	1.02
Believes TB can be cured ***	0.036	0.92	0.91	0.041	0.97	0.94
Number of household members	0.039	6.02	5.93	0.038	5.57	5.78
Food cooked on stove, open fire ****	0.014	****	****	N/A	N/A	N/A

Notes:
* Proportion of TB and non-TB population in each region
Of the TB population 44% of the females and 34% of the males came from the Eastern states.
Only 10% of the female and male TB population came from the Northern states.
** Excluding all PSU with no Anganwadi/ICDS center
*** Excluding all unknown responses
**** Proportion of TB and non-TB respondents cooking on Stove/ Chullah/ Open Fire.
All missing values were excluded (38% of TB cases and 21% of non-TB cases)
The percentages for the TB and non-TB population were nearly identical
with 6% cooking on a stove, 81% on a chullah and 13% on an open fire.

Table 3.11: Differences between the TB and non-TB respondents CDF for the female and male datasets. The maximum distance between the TB and non-TB CDF is shown, along with the mean of the TB and non-TB data for the variable. The highlighted cells show variables which were not significant at the 1% level.

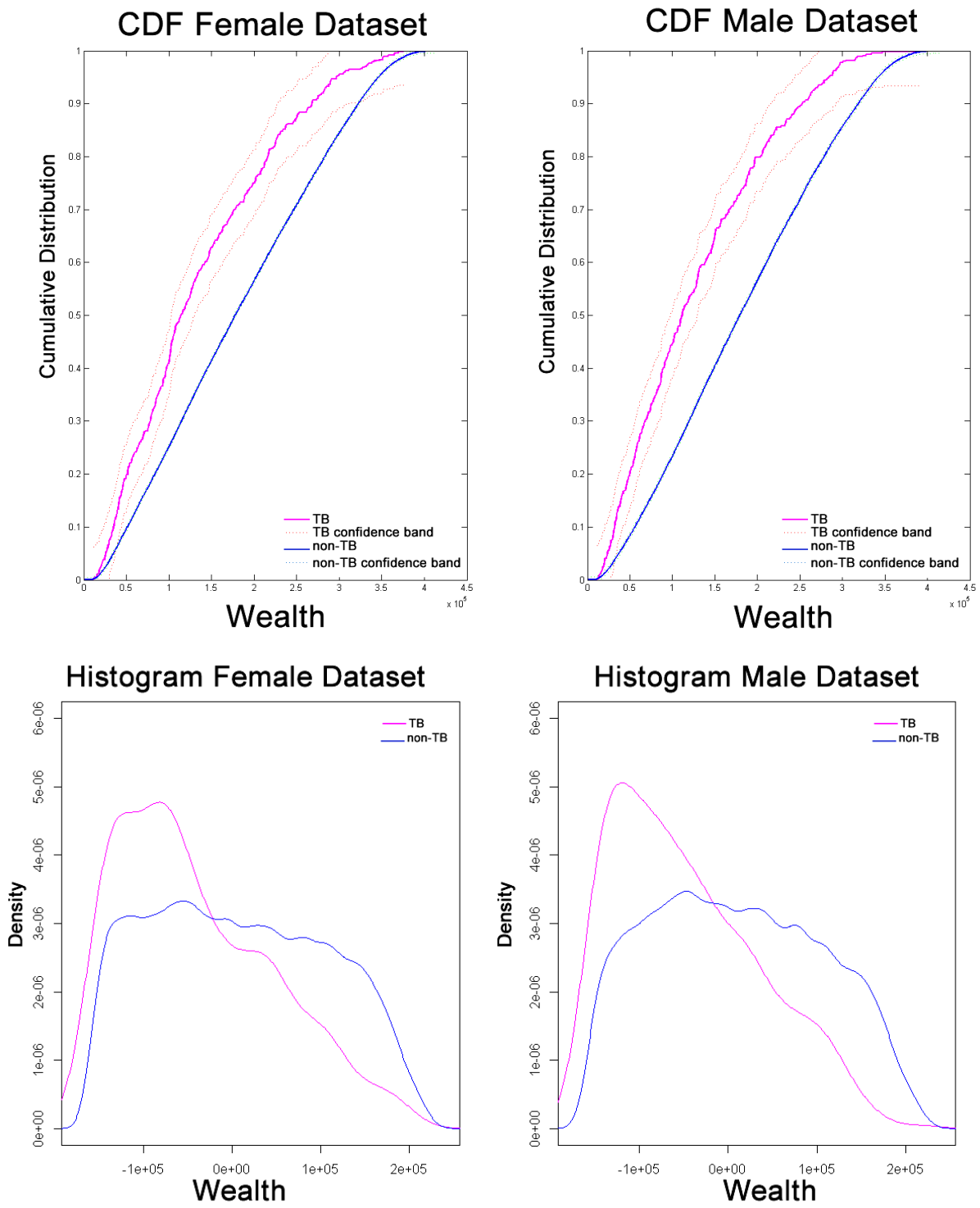

Figure 3.4: The CDF and density function for the wealth index factor score variable.

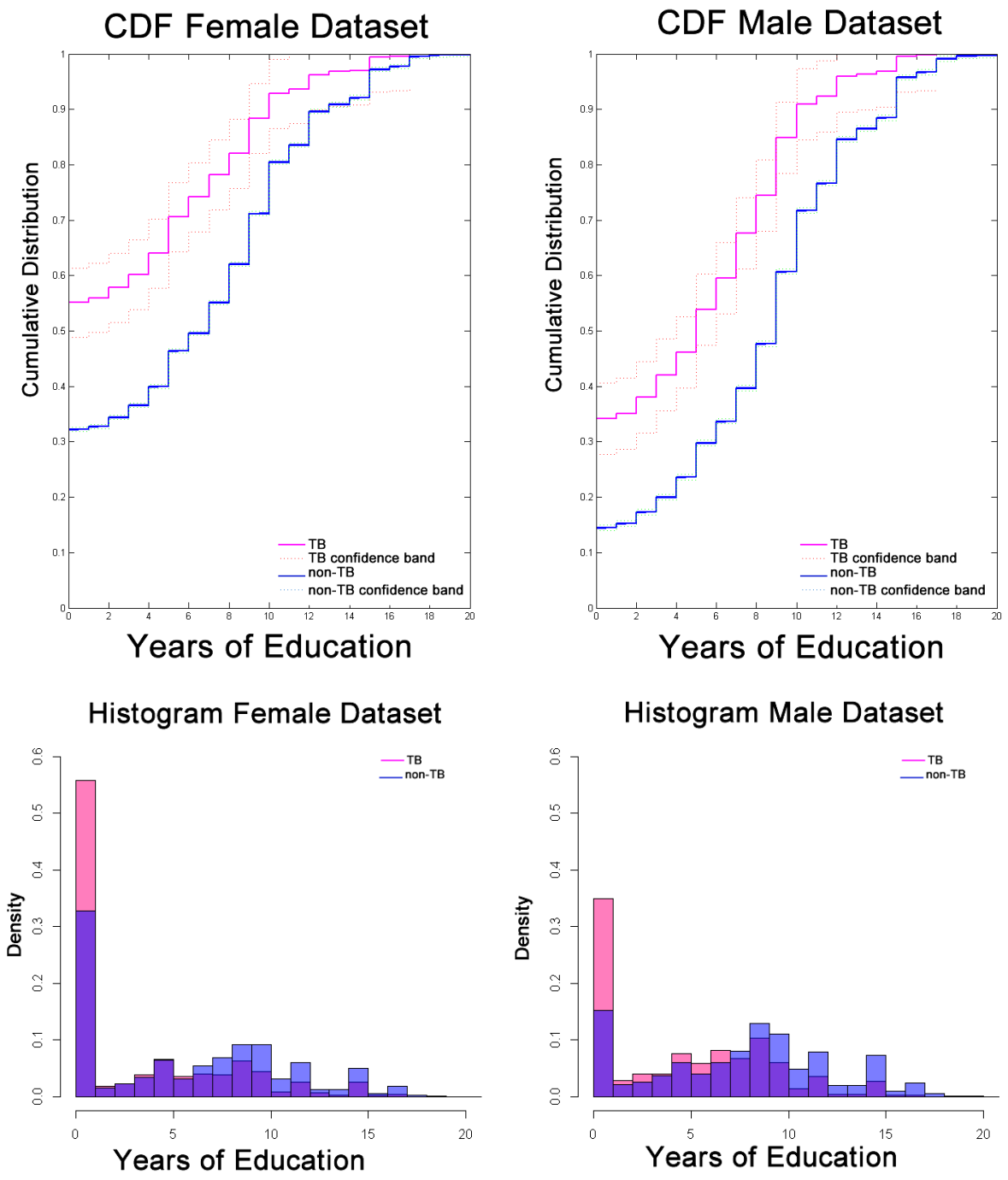

Figure 3.5: The CDF and density function for education.

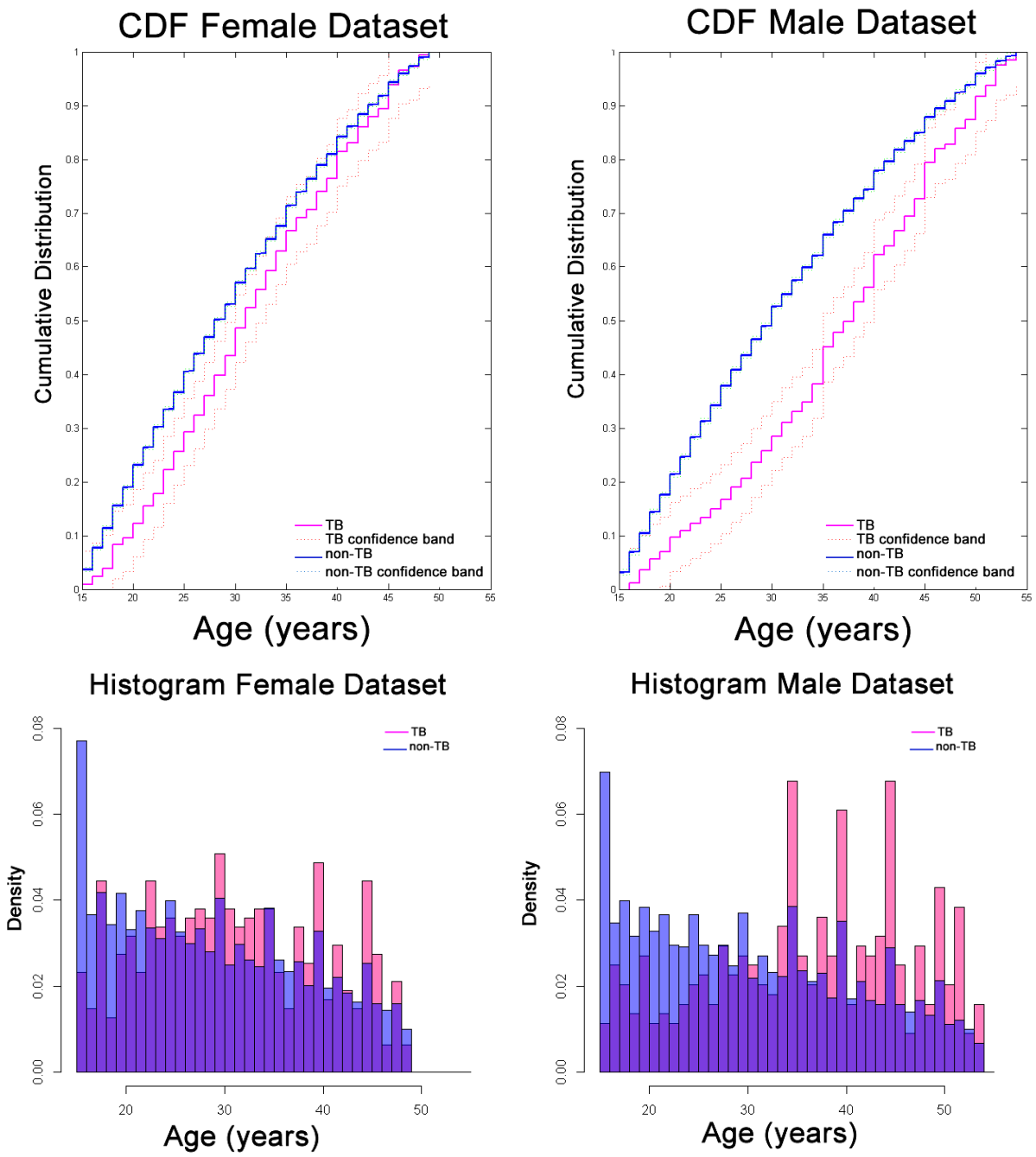

Figure 3.6: The CDF and density function for respondents age.

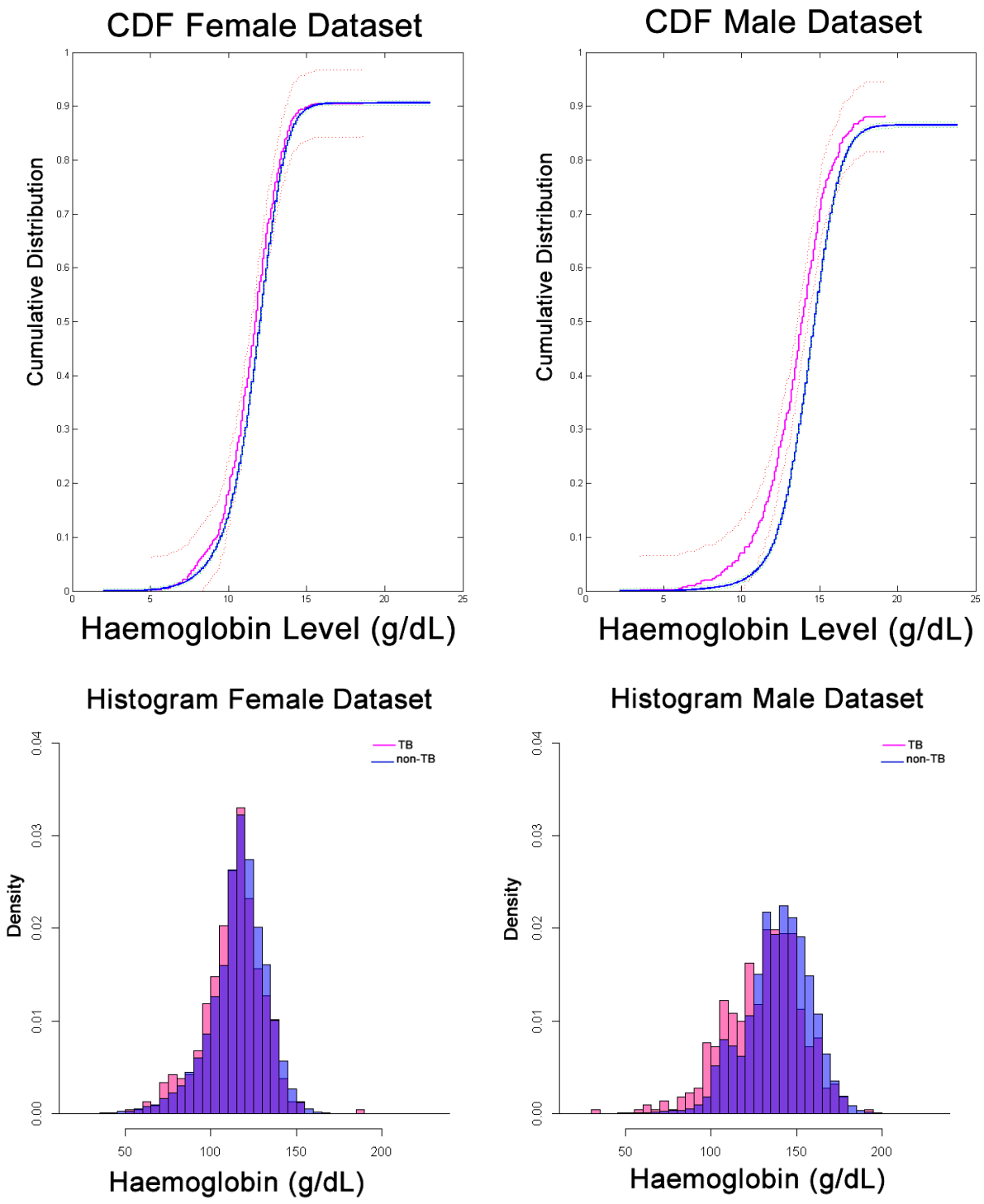

Figure 3.7: The CDF and density function for respondents haemoglobin levels.

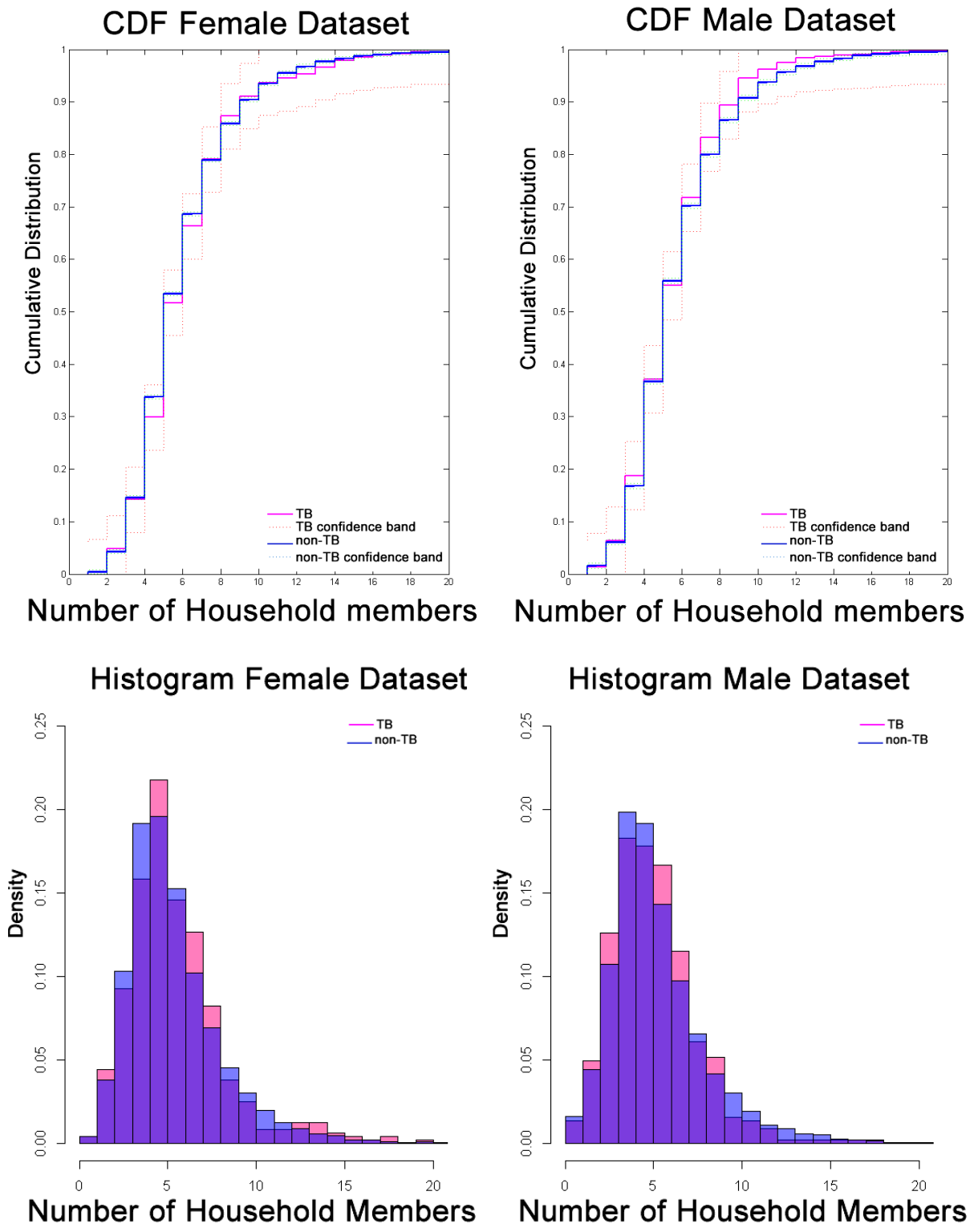

Figure 3.8: The CDF and density function for number of household members.

3.3 Nearest neighbour

Nearest neighbour classifies points based on the surrounding points as described in Section 2.3. Nearest neighbour analysis was used in this analysis to classify if a respondent was likely to have TB or not. Due to the differences between the populations not being extreme the nearest neighbour classifier did not perform well. Figure 3.9 shows the distribution of the TB and non-TB populations split by two of the best performing variables, wealth and BMI. This figure shows that the TB population density is centred at a lower BMI and at a lower wealth level than the non-TB population. The figure also shows that the TB and non-TB distributions significantly overlap each other, resulting in the expected performance to be poor. Re-substitution and cross-validation error, also described in Section 2.3, were performed on the data. A summary of the re-substitution loss error rate for 1, 3, 5, and 10 nearest neighbours is shown in Section 5.8 of the Appendix.

The female and male datasets had similar results. The best predictor variables from the nearest neighbour technique were:

- Wealth index factor score
- Body mass index
- Haemoglobin level
- Age
- State

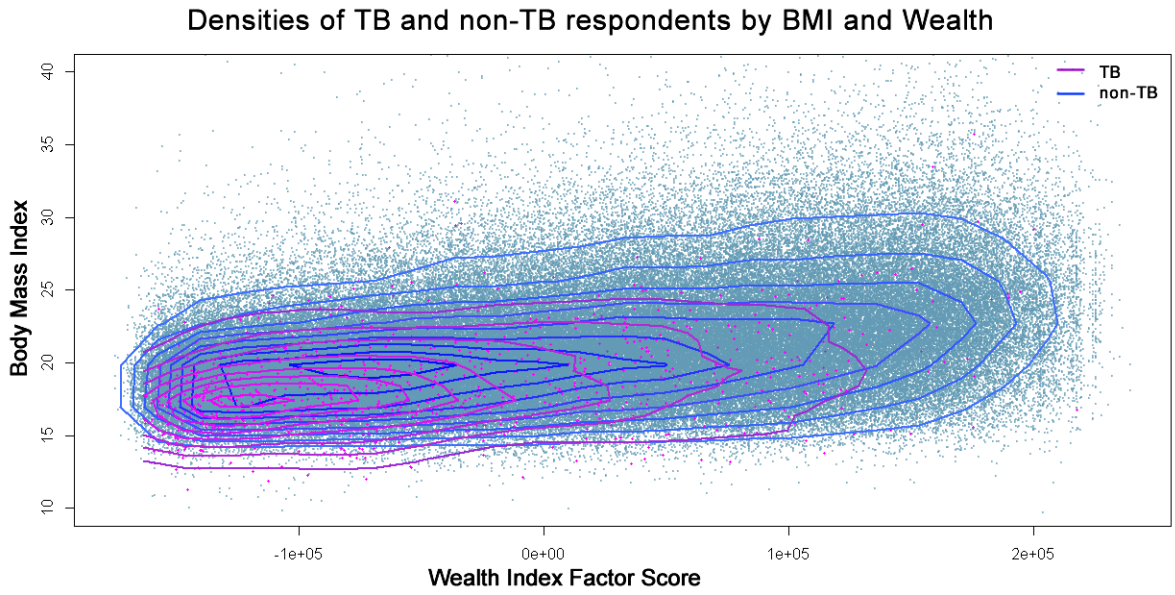

Figure 3.9: All respondents with the TB and non-TB densities split by wealth and BMI.

3.4 Classification trees

Classification trees, as described in Section 2.4 were used to identify which groups of respondents had TB. Cross-validation was carried out to determine the optimal depth of the classification tree. The results of cross-validation and re-substitution are shown in Figure 3.10 and the optimal classification tree is shown in Figure 3.11. The cross-validated results show that the best classifier was when there were 12 nodes in the classification tree. This separated six groups which were classified as having TB. These groups characteristics are described below:

- Respondents with a BMI less than 15.3, who were less than 25.5 years old, and had less than 7.5 years of education
- Respondents with a BMI less than 15.6, and who were older than 25.5 years
- Respondents with a BMI between 15.6 and 18.2, who were male, and who were older than 25.5 years
- Respondents with a BMI between 15.6 and 18.2, who were female, who were older than 25.5 years, who were from the Central or Eastern states, and who had given birth to five or more children
- Respondents with a BMI between 15.6 and 16.3, who were female, who were between 25.5 and 44.5 years old, who were from the Central or Eastern states, and who had given birth to less than 5 children
- Respondents with a BMI between 17.8 and 18.2, who were female, who were between 25.5 and 44.5 years old, who were from the Central or Eastern states, and who had given birth to less than 5 children.

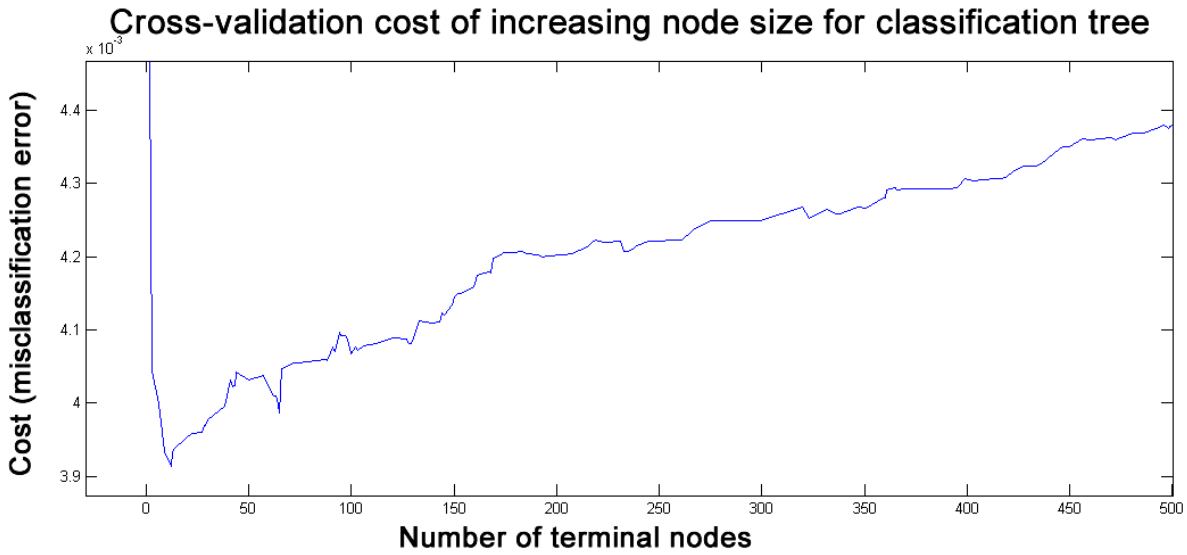

Figure 3.10: Regression Tree cross-validation.

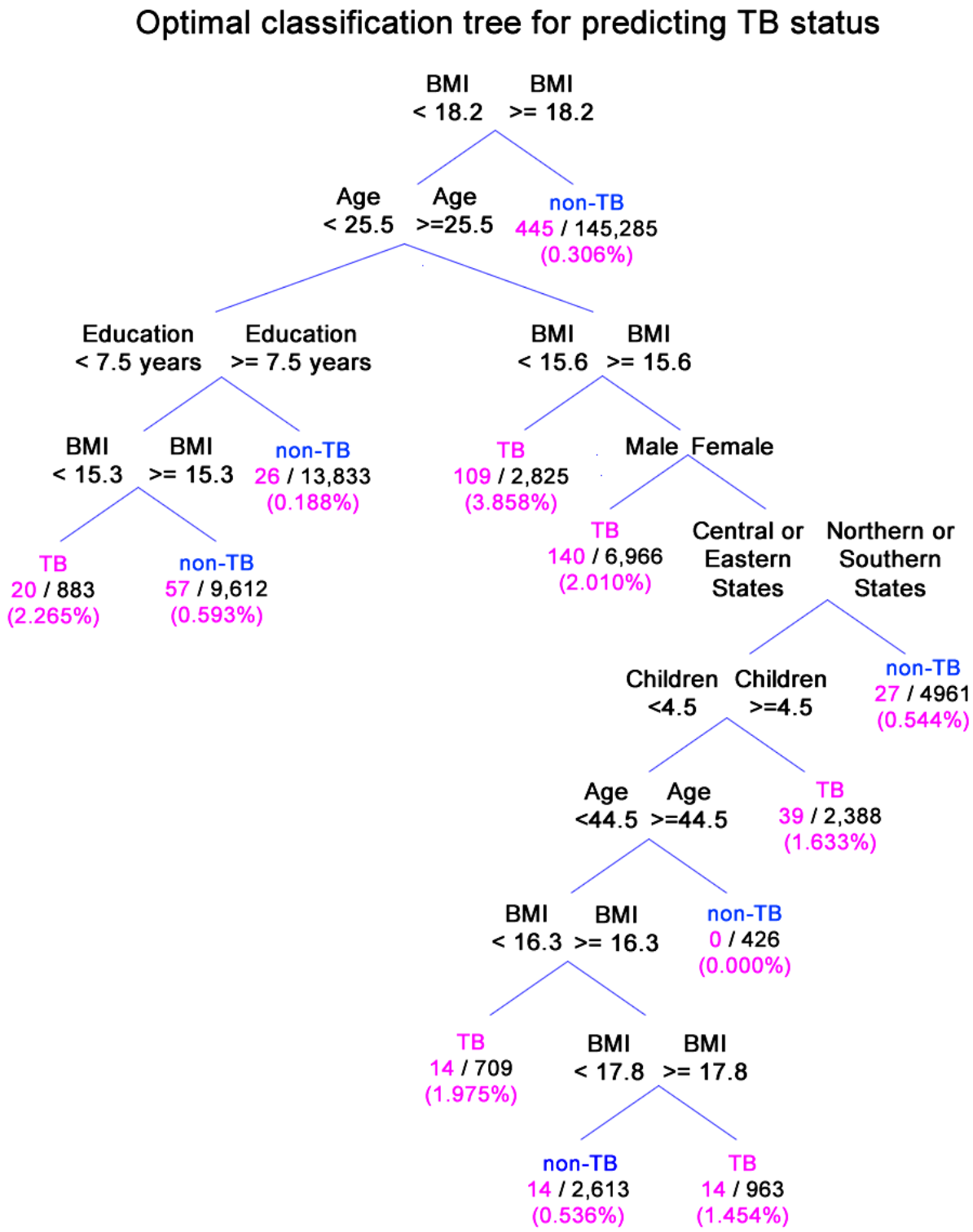

Figure 3.11: Optimal classification tree to determine TB status.
The TB status is shown, along with the number of TB cases in the node, and the percentage of TB cases in the node

3.5 Independence testing

Independence testing, as described in Section 2.2.2, was used to determine how independent the variables were. This was used to help determine which variables to include in the generalised linear models and to subset the DAGs. When all of the data was included in the test every pair of variables was deemed highly non-independent. For both the female and male datasets, 10 samples of size 1000 was taken. The independence tests were run on the 10 samples and the average of the P-value used. A summary of the results is in the Appendix in Section 5.9 The most non-independent variables were:

- Age at first intercourse and age at first marriage
- Age and total children born
- Cooking done under a chimney and what food cooked on (applicable to the female dataset only)
- Total number of sexual partners and age at first marriage
- Wealth index factor score and standard of living
- Number of children and number of sons at home.

3.6 Logistic regression

Logistic regression as described in Section 2.5 were used to predict which respondents had TB. They were used in both the initial and secondary analysis:

- Initial analysis – To help determine which variables to include in the secondary analysis
- Secondary analysis – To create a model predicting which respondents were at the greatest risk of having TB.

The results of the logistic regression from the initial data analysis are in the Appendix in Section 5.6. The variables used in the models needed to be easily obtainable if new respondents were to have their TB risk analysed from the same model. The model also needed to be simple, easy to understand, to predict the majority of TB cases, and to also predict the majority of non-TB cases.

For each model cross validation, as described in Section 2.3, was used to test how well the model performed on previously unseen data. The significance of each variable from the chosen model is shown in Table 3.12. The 10-fold cross validation results are in Section 5.5 of the Appendix. The optimal model as described in Table 3.12 correctly predicted 76% of the TB respondents as having TB, and 64% of the non-TB respondents as not having TB. This model is shown in Table 3.12.

Variable		Odds ratio	95% CI		Pr(≥|z|)		Variable Sig	
Intercept						*		
Wealth	very low †	1					0.005	**
Wealth	low	1.21	(0.92,	1.59)	0.181			
Wealth	average	1.06	(0.79,	1.42)	0.717			
Wealth	high	1.00	(0.73,	1.38)	0.976			
Wealth	higher	1.06	(0.75,	1.49)	0.758			
Wealth	very high	0.49	(0.30,	0.81)	0.005	**		
Years of Education	none †	1					0	***
Years of Education	1-5 years	0.72	(0.60,	0.88)	0.001	***		
Years of Education	6-8 years	0.71	(0.58,	0.87)	0.001	**		
Years of Education	9 years	0.62	(0.48,	0.81)	0	***		
Years of Education	10 - 11 years	0.46	(0.34,	0.62)	0	***		
Years of Education	12-14 years	0.45	(0.31,	0.65)	0	***		
Years of Education	15 + years	0.42	(0.27,	0.65)	0	***		
Age	15 - 24 †	1					0	***
Age	25 - 34	2.35	(1.63,	3.39)	0	***		
Age	35 - 44	3.23	(2.25,	4.64)	0	***		
Age	45 - 54	4.54	(3.12,	6.61)	0	***		
Location	Northern States †	1					0	***
Location	Central States	1.24	(0.85,	1.79)	0.265			
Location	Eastern States	2.85	(2.05,	3.96)	0	***		
Location	Southern States	1.36	(0.97,	1.91)	0.075	.		
Country/City	Large City †	1					0	***
Country/City	Small City	0.61	(0.43,	0.86)	0.005	**		
Country/City	Town	0.65	(0.51,	0.83)	0	***		
Country/City	Countryside	0.58	(0.47,	0.71)	0	***		
Gender	Male †	1					0.3	
Gender	Female	1.20	(0.85,	1.69)	0.304	***		
Haemoglobin		0.99	(0.99,	1.00)	0.009	**	0.009	**
BMI		0.66	(0.59,	0.74)	0	***	0	
BMI:male †		1					0	***
BMI: female		1.13	(1.08,	1.19)	0	***		
BMI: Northern States †		1					0	***
BMI: Central States		1.02	(0.92,	1.12)	0.73			
BMI: Eastern States		1.16	(1.06,	1.26)	0.001	**		
BMI: Southern States		1.18	(1.08,	1.29)	0	***		
BMI:Age 15 - 24 †		1					0.002	**
BMI:Age 25 - 34		0.96	(0.90,	1.03)	0.305			
BMI:Age 35 - 44		0.88	(0.82,	0.95)	0.001	***		
BMI:Age 45 - 54		0.95	(0.88,	1.02)	0.181			
Female: age 15 - 24 †		1					0	***
Female: age 25 - 34		0.68	(0.46,	1.01)	0.058	.		
Female: age 35 - 44		0.39	(0.27,	0.58)	0	***		
Female: age 45 - 54		0.39	(0.25,	0.62)	0	***		
BMI : wealth very low †		1					0.001	***
BMI : wealth low		1.08	(1.01,	1.17)	0.035	*		
BMI : wealth average		1.10	(1.02,	1.19)	0.014	*		
BMI : wealth high		1.09	(1.01,	1.19)	0.029	*		
BMI : wealth higher		1.09	(1.00,	1.18)	0.048	*		
BMI : wealth very high		1.26	(1.14,	1.38)	0	***		

notes:
†indicates the reference category
., *, **, ***, indicate increasing order of significance
Haemoglobin level was centered by decreasing the values by 130
BMI was centered by decreasing the values by 21

Table 3.12: Odds ratio, 95% CI, and variable significance from the optimal logistic regression model.

3.7 Directed acyclic graphs

The most statistically likely directed acyclic graphs as described in Section 2.6 were used to obtain a visual summary of how the variables were associated with each other. The most likely combinations of variables were determined by independence testing and using BIC. AIC was also looked at and generally agreed with the BIC. Examples of the calculations from the female dataset to determine the DAG with the lowest BIC are shown in Tables 3.13 and 3.14. This shows the different DAGs investigated for a group of nodes and the associated maximum likelihood (ML), degrees of freedom (DF), penalty term for the BIC calculation, bayes information criteria (BIC), akaike information criteria (AIC), and the euclidean distance (Euclidean). A further summary of the calculation is in the Appendix in Section 5.10. Moral graphs were derived from the directed acyclic graphs as discussed in Section 2.6.1.

The female and male datasets had similar, but not identical results. The female DAG has two extra variables than the male DAG: where the food is cooked and if the cooking is done under a chimney. As these variables were not available for the male dataset they were not analysed. For the male dataset a directed edge existed between the frequency of reading the paper and if they believed TB could be cured and if they would keep it a secret if a household member contracted TB; these edges were not in the female dataset. The female DAG also had an edge between housetype and years of education; this was not seen in the male DAG. These differences were also seen in the moral graphs. The directed acyclic graph including all the variables analysed are shown in Figure 3.12 and 3.13. The moral graphs for both the female and male dataset are shown in Figures 3.14 and 3.15.

# of living children, sons at home, daughters at home	ML	DF	Penalty	BIC	AIC	Euclidean
Full Model	(144,648)	95	241	144,889	144,743	
A - All independent	(206,923)	11	28	206,951	206,934	0.0896
B - Living children to sons and daughters	(161,375)	41	104	161,479	161,416	0.0116
C - Living children to sons to daughters	(177,963)	35	89	178,052	177,998	0.0274
D - sons and daughters to living children	(148,517)	86	218	**148,735**	**148,603**	0.0140
E - Living children to daughters to sons	(182,599)	35	89	182,688	182,634	0.0392

Table 3.13: Example directed acyclic graph calculation from the female dataset. The highlighted cells show the directed acyclic graphs with the lowest BIC and AIC values

Years of education (133), Keep it secret if a family member gets TB (476), Believes TB can be cured (475), Frequency of reading newspaper (157)	ML	DF	Penalty	BIC	AIC	Euclidean
Full Model	(203,135)	251	637	203,772	203,386	
A - 475 476 independent. 133 to 157	(217,036)	31	79	217,115	217,067	0.0089
B - 133 to 476, 157, 475	(212,275)	55	140	212,414	212,330	0.0087
C - 133 to 157 to 476 and 475	(213,125)	34	86	213,211	213,159	0.0083
D - 133 to 476, 157, 475. 157 to 476, 475	(211,889)	139	353	212,242	212,028	0.0088
E - 133 to 476, 475, 157. 475 to 476	(203,433)	83	211	**203,643**	**203,516**	0.0001
F - 133 to 157. 157 to 475, 576. 475 to 476	(204,012)	59	150	204,162	204,071	0.0007
G - 133 to 476, 157, 475. 157 to 476, 475. 475 to 476	(203,135)	242	614	203,749	203,377	0.0000
H - 133 to 476, 157, 475. 476 to 465	(203,433)	83	211	**203,643**	**203,516**	0.0001
I - 133 to 157. 157 to 476, 475. 476 to 475	(204,012)	59	150	204,162	204,071	0.0007
J - 133 to 476, 157, 475. 157 to 476, 475. 476 to 475	(203,135)	242	614	203,749	203,377	0.0000

Table 3.14: Example directed acyclic graph calculation from the female dataset. The highlighted cells show the directed acyclic graphs with the lowest BIC and AIC values

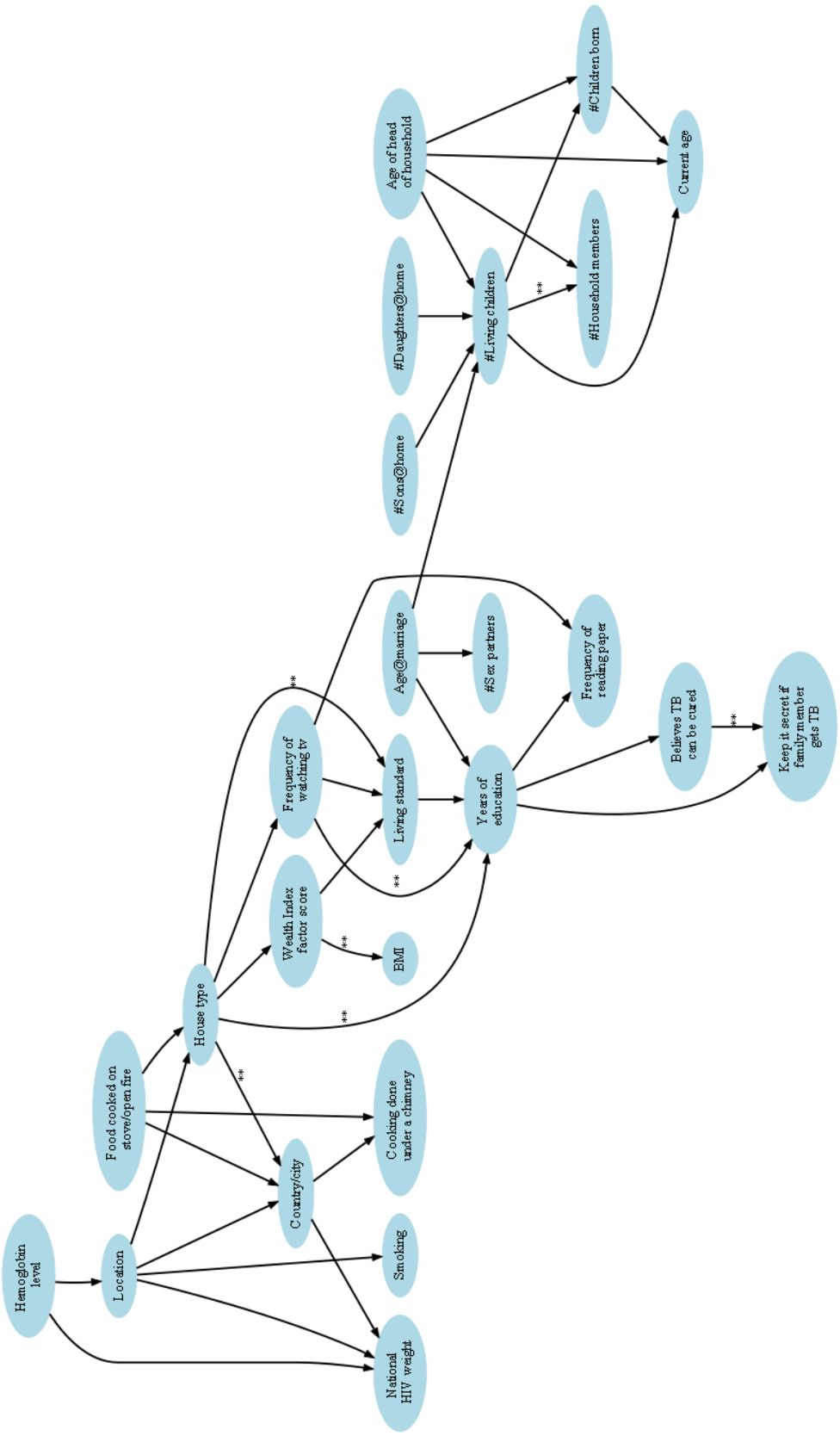

Figure 3.12: Most statistically likely directed acyclic graph for the female dataset.

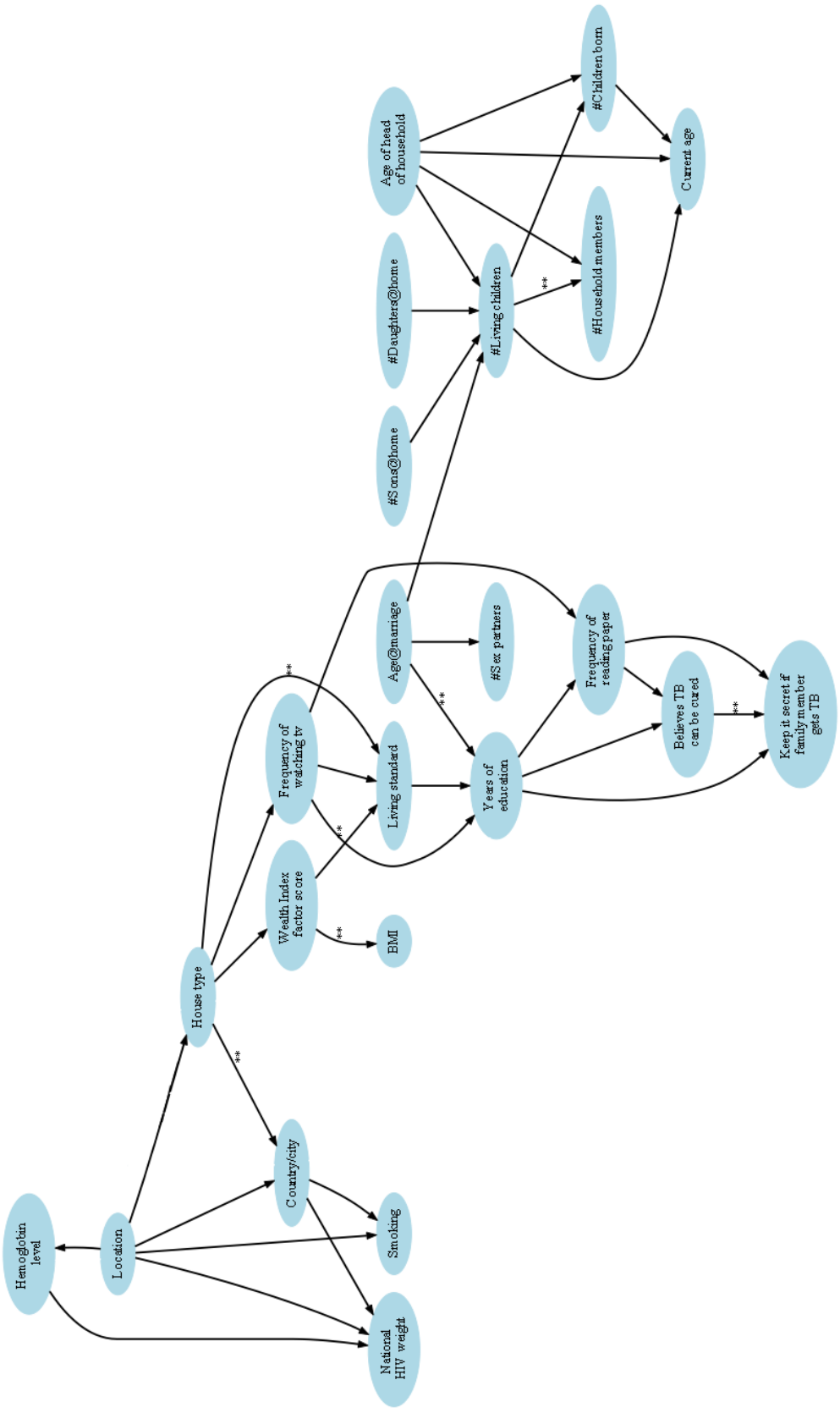

Figure 3.13: Most statistically likely directed acyclic graph for the male dataset.

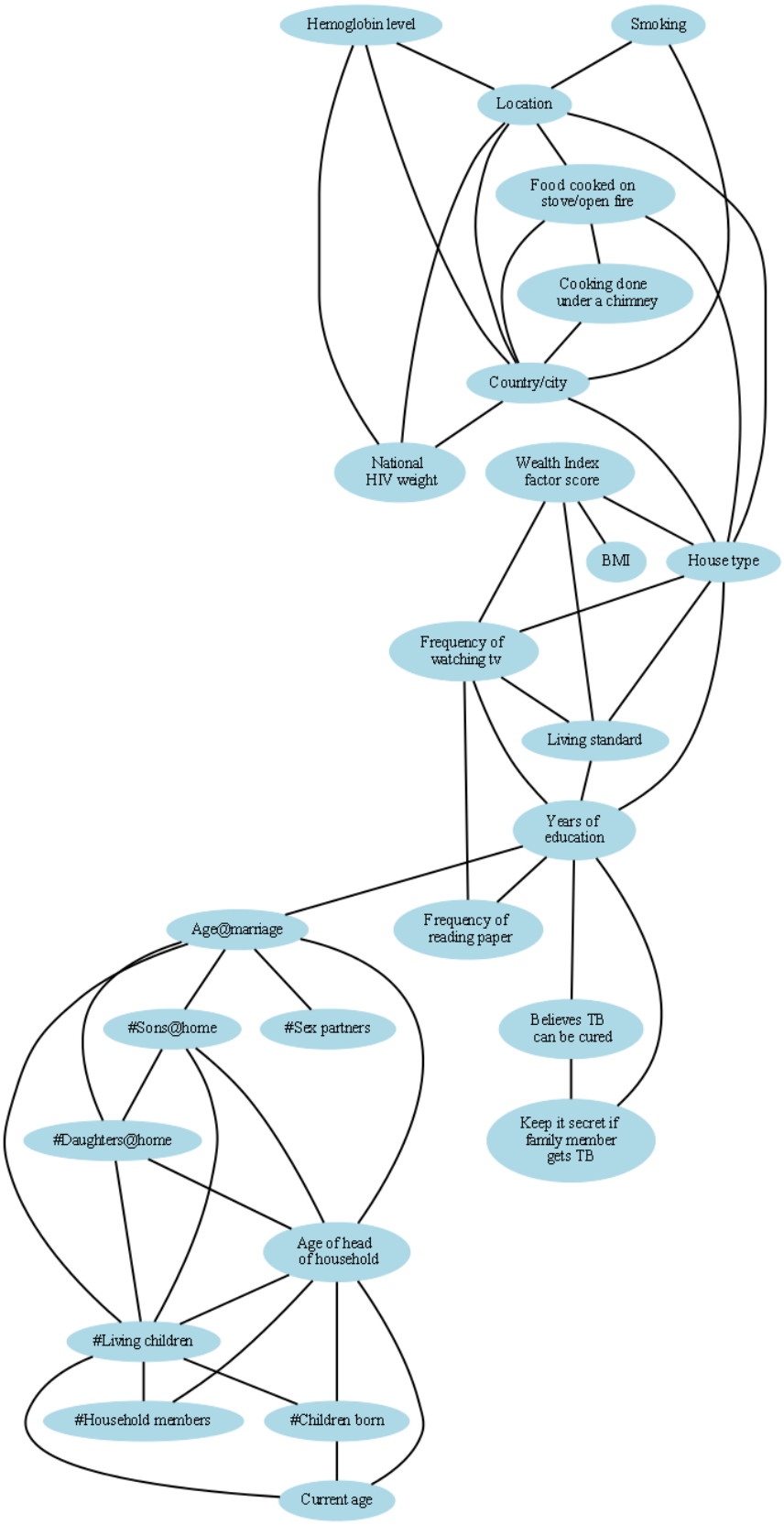

Figure 3.14: Moral graph calculated from the female most statistically likely directed acyclic graph.

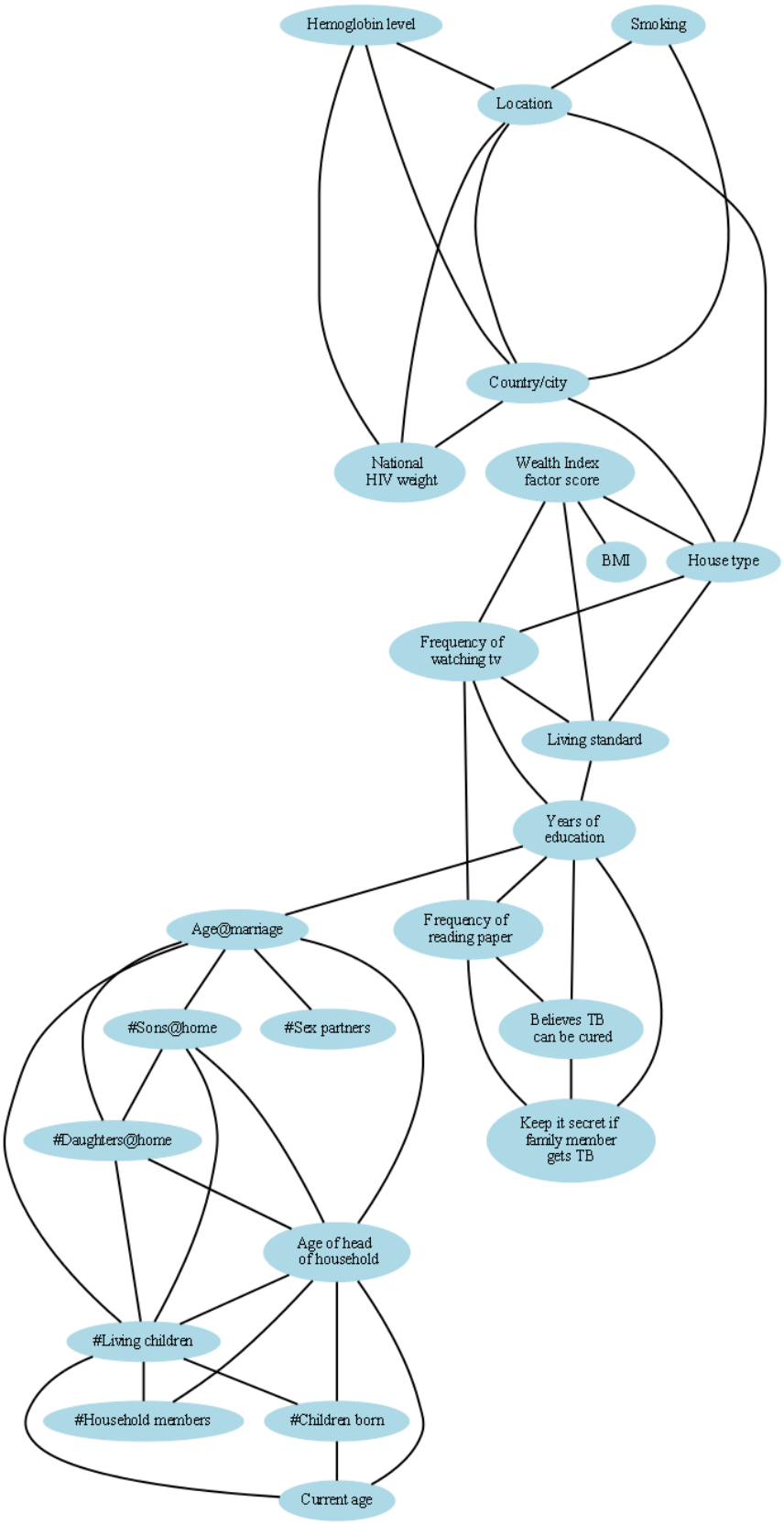

Figure 3.15: Moral graph calculated from the male most statistically likely directed acyclic graph.

[This page intentionally left blank]

Chapter 4

Discussion

This thesis used datasets available from the NFHS-3 to analysis the presence of tuberculosis in India. The differences in the TB and non-TB distributions were explained in Section 2.2 and presented in Section 3.2. Nearest neighbour classification, classification trees, and logistic regression were trialled as ways to predict TB cases as explained in Sections 2.3, 2.4, and 2.5 and presented in Sections 3.3, 3.4, and 3.6. The most statistically likely directed acyclic graphs were calculated for both the female and male datasets and used to visualise the most likely combination of variables as explained in Section 2.6 and presented in Section 3.7.

Many of the findings were expected and are presented in Section 4.2. Findings which were unexpected and not consistent with the literature review are explained in Section 4.3 and are further discussed in Section 4.5. The strengths, weaknesses, and potential follow up work for this thesis are discussed in Sections 4.7, 4.9, and 4.10.

4.1 Overview of key findings

After a thorough review of the literature, a selection of variables from the NFHS-3 dataset were selected. We found the variables with the largest difference in distribution between the TB and non-TB populations, as well as the most relevant variables to include in a classification tree and logistic regression model for predicting TB. Using a nearest neighbour approach to predict respondent's TB status was also trialled, however this was not successful. How the risk variables related to each other were investigated by finding the most statistically likely directed acyclic graphs. While the classification tree and logistic regression combined the female and male data, the difference in distribution functions and investigations with directed acyclic graphs were gender-specific.

The risk factors with the largest difference in distribution between the TB and non-TB populations were:

- BMI (See Figure 3.2)
- Wealth (See Figure 3.4)
- Education (See Figure 3.5).

These results were the same for both the female and male datasets.

Using crossvalidation we found that the optimal logistic regression model correctly classified 76% of respondents with TB; another 36% of non-TB respondents were also incorrectly classified as having TB. Similarly the classification tree correctly classified 37% of respondents with TB; another 0.5% of non-TB respondents were also incorrectly classified as having TB. The variables used for logistic regression were:

- Wealth
- Years of Education *
- Age *
- Location *
- Country/City
- Gender *
- Haemoglobin Level
- Body Mass Index *

where the variables marked with * were also used for the optimal classification tree. The optimal classification tree also included the variable 'number of children given birth to' which was not significant and dropped from the logistic regression analysis.

The directed acyclic graphs presented the most likely associations between the variables. While the literature review showed distinct differences between female and male lifestyles, the directed acyclic graph results were nearly identical. There were several groups in both the female and male graphs. The main groups and a summary of the variables in the groups were:

- General household information – number of household members, age at marriage, current age
- Education levels – years of education, belief that TB can be cured
- Wealth – wealth index factor score, living standard, house type
- Health – body mass index, haemoglobin levels, HIV weight, frequency of smoking.

4.2 Findings which were consistent with literature

From the existing literature we expected to identify high rates of TB in India due to its high population density [34, 47] and low health conditions [34, 35, 49, 50, 52]. These factors allow the relatively easy transmission of an infection which spreads via the airways and increase the risk of any TB infection progressing to active (and therefore contagious) TB. Respondents with low wealth were expected to be at an increased risk [35, 47, 53] as well as those with low/no formal education [52, 54]. There were some expected differences between the female and male population, such as lower education levels for females [77] which were observed. Although the level of contact respondents had with people infected with TB, their health, wealth, and education are associated with each other, we discuss these variables separately.

4.2.1 Distribution of variables

The distribution of the variables were similar for the female and male datasets apart from a few expected differences which related to education and age at marriage. These differences highlighted the difference in education levels and marriage expectations for females and males in India.

As discussed in Section 1.11 the male population is expected to obtain an education and to become literate. This was shown in the exploratory data analysis where 32% of females had no education compared to 15% of males, 17% of females read newspapers/magazines daily compared to 50% of males. The female population is also more likely to marry young compared to the male population. This was clearly shown with 47% of females married before the age of 18 compared to 13% of males.

Respondents who marry young were expected to be more likely to have TB due to the expectation that they had low wealth, low education, and from central/eastern India where TB rates are highest [75]. The relation between age at marriage, wealth, and gender are shown in Figure 3.3 for both the TB and non-TB populations. It is clear that TB respondents have lower wealth than non-TB respondents, and that the younger their age of marriage the lower their expected wealth index.

4.2.2 Contact with TB cases

Few variables were available to determine how much contact respondents had with people who had active TB. As TB is more dominant in areas with low wealth, variables relating to low wealth could also be indicators of contact with active TB cases. Wealth index factor score, housetype, and standard of living had significantly different distributions for the TB and non-TB populations. This could be explained by people with low wealth, in lower quality homes and with lower standards of living being in more contact

with people with active TB. These groups are also more likely to have weak immune systems which allow active TB to develop.

4.2.3 Health

From the analysis we found that respondents who had high BMI levels were the least likely to be infected with TB. This was expected as poor health and low immune systems are significant factors for TB. Malnutrition is also associated with weak immune systems and anaemia [80]. It was expected that the higher a respondent's haemoglobin level, the stronger their immune system and the greater the nourishment they received. This seems to be the case with the logistic regression where the risk of TB decreases with higher haemoglobin levels. The distribution functions also show this, as in Figure 3.7.

BMI is related to malnourishment and respondents with TB had lower BMI than respondents who did not have TB. As TB is associated with weight loss all weights of TB respondents will be a combination of their original weight, and any weight they have subsequently lost due to TB. BMI is related to TB for two reasons:

- People with low weight (and associated low immune systems) are more likely to contract TB
- Once someone has developed active TB they are expected to lose weight.

4.2.4 Wealth

As discussed in Section 1.7, low wealth is associated with low quality housing [40, 41], malnourishment [42, 43], the use of biomass fuels [44, 45], limited access to health care [40, 41], being of poor health [40, 41], having extremely limited leisure time [42], and having little or no education [42]. Low wealth is also associated with delays in travelling to a TB clinic [22], not keeping medical appointments [41], and defaulting on TB treatment [14, 25, 41].

Figure 3.4 clearly shows the difference in wealth distribution between the TB and non-TB populations. The wealth distribution was significantly denser at low levels in the TB population compared to the non-TB population. Of the respondents surveyed, those with TB were less likely to be wealthy compared to respondents who did not have TB. Half of the TB population had wealth levels less than 1.1 compared to half of the non-TB population having wealth levels greater than 1.8.

It was no surprise that low wealth was a significant risk factor for TB as the previous literature had identified a strong relation between low wealth and developing TB. While low wealth is associated to an increased likelihood of developing active TB, low wealth is also associated with TB treatment never commencing or being discontinued before completion (defaulting on treatment). Therefore, if a wealthy

and a poor respondent were both identified as having active TB one year prior to the survey, the wealthy respondent is now more likely to have been treated and answer 'no' to the survey question of 'do you have TB?'. Moreover, the poor person with still untreated TB is likely to be very sick and to have had to take days off work; further exacerbating their low wealth and health [46].

4.2.5 Education

People with low or no education were expected to be at increased risk of TB [52, 54]. As discussed in Sections 1.9 and 1.11, wealth and education were expected be strongly associated with each other due to the opportunities education provides [71]. Low wealth and low education were also expected to be associated with child marriage, especially for respondents in rural areas of Central and Eastern India [75]. As respondents who were married young were thought to be in a demographic group which was at higher risk of TB [75] we briefly investigated this association.

We found from our investigation that the more educated a respondent was the more likely she or he was to know that TB could be cured. This was consistent with previous findings that in addition to the cost of treatment, people defaulted on their TB medication due to their their low education and their limited knowledge of TB [14, 25, 27, 28, 29]. Figure 3.4 compared the female and male education levels for respondents with, and without TB. We found that respondents with TB were more likely to have no education than respondents without TB. We also found a dramatic difference between the education levels of females and males. Of the TB population 55% of females and 34% of males had no education. This dropped to 32% of females and 15% of males for the non-TB population.

Low education and low wealth were found to be strong indicators of how likely a respondent was to have TB. This was seen in the exploratory data analysis and in the directed acyclic graphs. The exploratory data analysis found that in the female dataset 64% of respondents in the lowest wealth bracket had no education and only 0.1% had over 15 years (for males this was 36% and 1.0%). For females in the highest wealth bracket 4% had no education and 33% had over 15 years of education (for males this was 1% and 39%). Figure 3.1 clearly shows the relation between low wealth and low education. The figure shows that 28% of TB respondents are in the lowest wealth bracket and have no education, compared to 13% of non-TB respondents.

The association between low education, low wealth, and a young age at marriage was observed in both the female and male datasets. Of female respondents with no education 75% were married before the age of 18 and 7% were still not married (for males this was 29% and 16%). For female respondents with over 15 years of education 6% were married before the age of 18 and 31% were still not married (for males this was 4% and 34%). Figure 3.3 shows the later the respondents marry the higher their expected wealth index.

Female respondents who were married before the age of 18 had an average wealth index level of -0.58 if they had TB; this increased to -0.25 for females without TB. For males who were married before the age of 18 the wealth index level was -0.80 for those with TB and -0.47 for those without TB. Female respondents who were married after the age of 24 and who had TB had an average wealth index level of -0.07, compared to 0.72 for the females who did not have TB. For males who married after the age of 24 those with TB had an average wealth index level of -0.13 and those without TB had an average wealth index level of 0.40.

4.3 Findings which were not consistent with the literature

While the majority of the findings were expected and consistent with the literature there were some unexpected findings. These related to our findings that respondents who smoked or cooked with biomass fuels were not at an increased risk of TB, that once all other factors were taken into account it was females who were at an increased risk of TB, and that the risk of TB was highest in the 45 - 54 year gap.

4.3.1 Pollution

It was reported that there was a large increase in the risk of TB for smokers compared to non smokers [34, 48]. Another article looked at the possibility of TB being associated with cooking smoke from biomass fuels [34]. We did not find any difference between the distribution of the TB and non-TB respondents by number of cigarettes smoked in the last 24 hours. It is also thought that people cooking with biomass fuels potentially increase their risk of TB [34]. We did not find any evidence of this association and we discuss it further in Section 4.5.

4.3.2 Gender and age

It was reported from studies in Spain, Sub-Saharan Africa, and India that males were at an increased risk of TB [49, 51, 54]. Another study from Africa found no difference in the risk of TB for females and males [52]. While we found a higher prevalence of TB in the male population the results from our logistic regression analysis (Section 3.6) found an increased risk of females for TB.

An article from Catalonia, Spain reported that people between the ages of 15 and 44 were most at risk of TB [49]. We did not observe this in our analysis but found that respondents aged between 45 and 54 were at the greatest risk of TB. Another article from South Africa stated that there was no relation between increasing age and the likelihood of contracting TB [52]. This was not consistent with our findings that increasing age leads to an increase in the risk of TB. As the NFHS-3 survey only included respondents between the ages of 15 and of 55 we could not compare with previous studies which stated that people

over the age of 55 were at an increased risk of TB [51, 54] or the study that people under the age of 15 were at a lesser risk of TB [49].

4.4 Findings which could not be verified

Not all the findings from the literature review could be investigated due to lack of data. For instance there was no information from the NFHS-3 dataset regarding if the respondent had used illicit drugs, been in prison, ate meat or fish daily, was deficient in vitamin D, E, C, or had a selenium deficiency. The relation between HIV/low immune systems and TB could not be verified due to the data not being at an individual level. This is discussed more in Section 4.4.1. While the number of household members, number of sons and daughters at home, number of living children, and number of children born were collected, the size of the household was not. This meant there was no information regarding the level of crowding within the household. The association between household crowding and TB was unable to be investigated. This is discussed more in Section 4.4.2. The amount of contact respondents had with people with active TB has been discussed briefly in Section 4.2.2 and is discussed more in Section 4.4.3.

4.4.1 HIV and immune levels

The HIV weight is a variable in the NFHS-3 dataset which estimates the HIV prevalence in the respondents region. A major problem with the HIV data was that due to confidentiality it was provided as a weight for each region instead of being specific for each respondent. Many of the HIV weights for regions were also missing, especially in the Northern (overall 75% missing) and Eastern (overall 68% missing) states. Due to the large amount of missing data and the HIV data at a regional instead of individual level, any HIV conclusions from this dataset are highly speculative. There was no variable available to indicate the level of the respondent's immune system. This is a difficult variable to measure. However, data on the number of doctors visits, days off work, days sick each year could help estimate the level of the respondents health. If a field was available which ranged from respondents who had strong immune systems to those with acquired immunodeficiency virus a more detailed analysis looking at HIV or immune levels could have been carried out.

The data provided by the NFHS-3 dataset for the HIV levels was not consistent with other existing literature. For instance from the dataset it appears that the Northern states have high HIV levels (53% of non-missing respondents in the highest HIV level) and the Southern states low HIV levels (only 2% of non-missing respondents in the highest HIV level). This is not consistent with other sources which have stated that the Eastern states of Nagaland and Manipur have the highest HIV/AIDS prevalence, followed

by the Southern states of Andhra Pradesh, Maharashtra, and Karnataka [89]. The results of the HIV testing were possibly due to the fact that HIV testing was voluntary. It may be that selection bias has occurred with respondents who knew that they had TB deciding not to be tested.

TB was expected be much more prevalent in the groups with high HIV weight [49, 97]. This is due to the strong correlation between weakened immune systems and active TB. The distribution function analysis did not present any difference between the TB and non-TB populations when looking at the HIV weights, possibly due to the amount of missing data. Due to the 'synergy from hell' [97] between people with HIV/AIDS and TB this was highly unexpected. Due to the highly unexpected results for HIV the logistic regression model was run with the HIV variable excluded. See Section 3.6 for the logistic regression results.

4.4.2 Household crowding

A study from Africa had found a significant increase in the rate of TB per household depending on the household size [51]. In the NFHS-3 data we did not find the number of household members to have significantly different distributions for the TB and non TB populations for either gender. Similarly the number of children born was expected to be significant due to people of lower wealth being associated with TB and multiple births, again this was not seen.

Household crowding has been cited as a significant variable associated with TB [e.g. 47, 51, 52]. Unfortunately there was no variable in the survey to indicate the level of household crowding and if the crowding was due to children or adults. While data on the number of household members was collected, information on the size of the respondent's dwelling, the number of rooms, and the number of people per square meter was not available. Potentially, households with more members also had more rooms in the household or larger dwellings. Therefore, we were unable to compare the levels of household crowding from our dataset.

4.4.3 Level of contact with TB

As discussed in Section 1.6, the health of household members are expected to be relatively similar. If one household member has a low immune system and is infected with TB, it would be expected that the other household members would also have low immune systems, putting them at a greater risk of also developing active TB. While each respondent was questioned as to whether anyone in their household was infected with TB the responses were such that it was not possible to distinguish between households where it was the respondent who had TB, and households where it was a family member who had TB. Due to this complexity every respondent who had TB also lived in a household which had TB. We mention in Section 4.2.2 that respondents in demographics known to have high TB rates were more likely to have an increased

contact with people with TB than respondents in demographics known to have low TB rates, however this is a biased measure.

4.5 New and surprising findings

While the majority of the findings were expected there were a few surprising results. In Section 4.5.1 we discuss that while the stereotypical expectations between females and males in India are different, the most statistically likely directed acyclic graphs for the female and male datasets had similar results. In Section 4.5.2 we discuss that the body mass index among females was slightly higher than males which was surprising given the gender-specific discrimination discussed in Sections 1.12. In Section 4.5.3 we discuss that the population with the lowest haemoglobin levels did not have the highest prevalence of TB, which was unexpected given the association between low haemoglobin levels, weakened immune systems, and TB. In Section 4.5.4 we discuss the unexpected results from the respondents age distribution, as well as how well the age of the household head predicted TB cases. In Section 4.5.5 we discuss biomass fuels and how cooking on a stove, chullah, or open fire did not affect the prevalence of TB.

4.5.1 Directed acyclic graphs

The differences between the most statistically likely directed acyclic graphs for the female and male populations were:

- In the female dataset haemoglobin was thought to affect the respondent's location by region, while for the male dataset this association was reversed
- In the female dataset smoking was only associated with their region, whereas for the male dataset smoking was associated with both the respondent's region and whether they lived in the country or the city
- In the female dataset the frequency of reading the newspaper or magazines was not directly related to their knowledge of TB and whether they would keep it a secret if one of their family members contracted TB. For males these variables were directly associated.

4.5.2 Body mass index

The distribution of respondent's height, weight, and body mass index (BMI) was investigated. Stunting occurs in malnourished respondents [80], and malnourishment is associated with TB. While the distribution of height was significant, it was far less significant than the distribution of weights. The average female

with TB was 159.28 cm, 50.56 kg, and had a BMI of 18.95. For females without TB these increased to 159.53 cm, 52.53 kg, and a BMI of 21.04. The average male with TB was 165.08cm, 52.51kg and had a BMI 18.45. For males without TB they were on average 165.65 cm, 56.11kg, and with a BMI of 20.83. Figure 3.2 shows the BMI histogram for females and males with and without TB.

It was also surprising that the average female BMI (20.32) was slightly higher than the average male BMI (20.24). We reported that Indian females were discriminated against [50], had less spent on their medical care [50], were sometimes fed after the males of the household [81], and fed less nutritious food [50] so a lower BMI for females was expected.

4.5.3 Haemoglobin levels

As discussed in Section 1.12, malnutrition and anaemia are strongly associated with people having low immune levels, which in turn are linked to the development of active TB. It was shown that 50% of females and 25% of males were anaemic so it was surprising that the levels of TB reported in the female population were lower than that reported in the male population (0.4% of the females had TB compared to 0.6% of the males). This could be investigated further. A few potential explanations for this finding are below:

- There may be a systematic measurement error in haemoglobin levels between females and males
- While females immune systems may be weaker they may also be less exposed to TB cases than the male population
- Fewer females than males may realize they need medical attention and so do not get tested, and subsequently do not report as having TB when they are actually infected
- Females may realise that they are sick but are choosing not to seek medical care and are unaware the cause of their sickness is TB
- Females may realise that they are sick but are prevented from seeking medical care so they are unaware they have TB
- Females may know they have TB but are more likely to incorrectly report that they are not infected.

4.5.4 Age

Age of household head was one of the best predictors in the female dataset in the nearest neighbour technique, however this was not the case for the male dataset. The difference in distributions between the TB and non-TB populations for the age of the household head was significant, but far less than many other variables.

The distribution for the respondent's age, shown in Figure 3.6, showed a gender-specific difference. The distribution for the female and male non-TB respondents age were roughly the same, however this was not the case for the TB respondents. The CDF of the female TB respondents age distribution was nearly linear with a slightly convex positive unipolar line. The CDF of the male TB respondents age distribution showed a concave positive unipolar line (See the CDF in Figure 3.6).

There was a spike in the frequency of respondents ages which ended in a 0 or 5 which was unexpected. This is possibly due to the respondent not knowing their precise age and rounding it to the closest 5. The male dataset had a much larger increase in frequency for every 5-th age group than the female dataset. This may be due to the questionnaire being designed with a focus on female respondents and the time being taken to get an accurate birth date for the female respondents, but not the male respondents.

4.5.5 Pollution

Biomass fuels and air pollution have been associated with people being of poor health [54], including TB [34, 54, 68]. Biomass fuels are thought to potentially increase the risk of TB [34]. There was no difference between the female TB and non-TB populations who cooked their food on a stove, chullah, or open fire. There was also no difference between the TB status of females who cooked their food under a chimney or not. As this variable was not available for the male population we have no information on it. This was surprising as air pollution and cooking smoke are associated with poor health and a link between biomass fuels and TB was expected to be found.

4.6 Inferences from this study

From this thesis it is clear that there is an association between poverty and TB. The reasons for this appear to be due to people in poverty being the most likely to have weak immune systems, to have low levels of education, and to have more contact with TB cases than their wealthier counterparts. These variables are all inter-related, however the separate roles they play is explained briefly below.

Respondents with weak immune systems are expected to have higher rates of active TB than those with strong immune systems [2]. This is due to the biology of TB, for active TB to occur in an infected person, their immune system must be weak enough that their body cannot resist the infection. A person with a strong immune system who has been infected with TB has latent TB (not contagious) and is unlikely to develop active TB (contagious). In comparison, a person with a weak immune system who is infected with TB is likely to develop active TB. In this thesis we observed that respondents with low BMI and haemoglobin values were more likely to develop TB.

Respondents with high levels of contact with people with active TB are at an increased risk of TB themselves. This is because TB is transmitted via the airways; for someone to become infected with TB they must have been in the same location as someone who had active TB [6]. We discussed in Section 1.6 that children in a household are expected to have similar levels of health [37]. If we continue this thought, we could also expect the respondent's social network to have similar immunity levels as the respondent. This would mean that people with weak immune systems mainly had contact with others who also had weak immune systems, and people with strong immune systems were mainly in contact with those who also had strong immune systems. If one member of a group with weak immunity developed active TB it would be expected that others would also develop active TB; the risk for the other members would increase over time. In comparison if one member of the group with strong immunity developed active TB it would not be expected that the others would also develop active TB; the risk for the other members would stay constant over time.

Respondents with low educational levels were at an increased risk of TB due to being less likely to believe TB could be cured, and therefore less likely to seek treatment. Respondents with low education levels were also more likely to be in the lowest wealth bracket, associated with poor health and nutrition. These associations make the costs of treatment difficult, and low immune levels increase the risk of TB. In comparison, highly educated respondents were more likely to know TB could be treated, and to be in the higher wealth brackets, associated with higher nutritional and health levels, and higher immune systems. These respondents are less likely to have contact with potential TB cases, to not develop active TB, and to seek treatment if they do. In this thesis we observed from the distribution functions, optimal classification tree, and optimal logistic regression model that respondents with low education were at a greater risk of TB than respondents with high levels of education.

From the optimal directed acyclic graphs calculated in this study the nature of interdependence between the variables associated with TB were found. We found that for both female and male respondents their educational level was associated with TB knowledge, frequency of reading the newspaper and watching tv, age at marriage, living standard, and house type. We found that wealth was associated with the respondents house type, body mass index (BMI), and living standard. The most statistically likely directed acyclic graphs showed that the respondents education was only indirectly associated with wealth. We found that the respondents location was associated with their haemoglobin level, house type, whether they lived in the country or the city, whether they smoked, and their HIV weight. The variables regarding the people in the household were nearly separated from the main graph, only connected by an association between age of marriage and number of living children. The number of living children was associated with all of the household variables.

4.7 Strengths of this study

By an initial thorough search of the existing literature this thesis has presented its findings in light of what is already known. By having an overview of the literature the main themes of contact with TB cases, health, wealth, and education were able to be established and focused on. The logistic regression and directed acyclic graph results are strengths of this study with the logistic regression model accurately predicting many of the TB respondent's TB status and the directed acyclic graphs showing the most likely conditional dependence structure among the variables.

There were a disproportionate number of non-TB cases (915 TB cases and 190,549 non-TB cases) and the TB and non-TB distributions significantly overlapped over the space of investigated variables however, the optimal logistic regression model accurately predicted over three quarters of the TB cases. As discussed in Section 4.8, while 36% of respondents were incorrectly predicted to have TB from this model, the error rate may in fact be a lower percentage. The logistic regression model was clear and simple to use, the variables did not use personal information and had clear values. Obtaining variable values for new respondents should not be difficult. Once the variables which were to be included in the logistic regression analysis were found every 2-way interaction was considered. This did not restrict our results to any smaller subset of interactions, all 2-way interactions were investigated and only discarded if they were insignificant from the ANOVA testing.

The final variables used for the logistic regression were the respondent's wealth, the number of years of their education, their age, which state they live in, if they live in the country or city, their gender, haemoglobin level, and BMI. Respondent's wealth and haemoglobin level are the variables which are potentially the hardest to obtain. Wealth was calculated by the NFHS from variables such as the respondent's water source, possessions, house type, and toilet facility. An estimate of the respondent's wealth should not be too hard to determine from basic questioning. Respondent's haemoglobin levels can be found instantly from a finger prick test.

Often directed acyclic graphs are used to infer conditional probability structures underpinning the variables without sound statistical justifications. We have shown that it is possible to find the most statistically likely directed acyclic graph for a given population using Bayesian Information Criteria (BIC). We were able to find differences between the associations among the variables for the female and male population. Using the same methodology this work could be continued to compare the directed acyclic graphs for other populations.

4.8 Reporting accuracy

The logistic regression model predicted many non-TB respondents as having TB. This may be partially due to some respondents who said they did not have TB, but were predicted to have TB, actually having TB. It seems plausible that of the respondents with active TB some did not report that they had TB. This could be due to the respondent being unaware that they were infected, expecting they were infected but not having any medical proof, or knowing that they were infected but choosing to state that they did not have TB. They could deny their TB status for multiple reasons, however the stigma associated with TB seems likely to be a main reason. See Section 1.8 for a brief discussion regarding the stigma surrounding TB.

Section 1.3 discussed how the female:male TB infection ratio changed between neighbouring Afghanistan, Pakistan, China, and Russia. We mentioned that while Afghanistan has more female than male TB cases reported, China and Russia have more male than female TB cases reported, and Pakistan has similar female and male TB rates [2]. Either the ratio of females and males infected with TB in each of these countries differs, or for some reason the reports from these countries are not reflecting the actual TB distribution. If the infection ratios are similar across neighbouring countries but the data available from these countries is not showing this then there is cause for further concern about who is being reported as having TB and who is not.

In Section 1.6 we discussed an article which suggested that the NFHS-3 showed some unbelievable results [37]. This article found that the percentage of children with similar malnutrition levels had decreased over the NFHS surveys, and that the time taken for each survey had also decreased. In the same section we also discussed how the wealth index may not identify the true wealth status of respondents and how it seems unusual that 80% of the slum population from the NFHS-3 survey were identified as being in the upper two quartiles of the wealth index [38].

4.9 Weaknesses of this study

As the data used is all from India the results from this study are specific to India. It is not appropriate to use the results found in this thesis for any country apart from India. While it is possible that very similar results would be found in other countries in the same region, such as Bangladesh, Nepal, Pakistan, and Nepal, caution must be exercised in such extrapolations.

For all the variables with missing values in the NFHS-3 dataset we used basic imputation methods and assumed that the missing values were missing completely at random. If the missing values were not missing

completely at random then the imputation method used was not necessarily appropriate. We have not tested how the results of this analysis change when other imputation methods are used.

The National HIV weight was not used for the logistic regression due to the large volume of missing data and due to the HIV weight being location, not respondent, specific. A more detailed investigation into the HIV weights could have been conducted with the possibility of determining why the results were unexpected, and how to include the HIV weights.

While the logistic regression model predicted over three quarters of the TB population to have TB a large number of non-TB respondents were also predicted to have TB. From the results of our initial logistic regression testing we did not find any 3-way interactions which were statistically significant. We limited our final logistic regression model to only include two-way interactions. Due to the large size of the dataset and the number of variables in the model, had we tested for three-way interactions the time to find the optimal model would have been large. If three-way interactions were investigated it is possible that a three-way interaction could have improved on our optimal logistic regression model. If this was the case more respondents could have had their TB status correctly classified.

The classification tree results misclassified a much smaller percentage of the non-TB population than the logistic regression model, however less TB respondents were also correctly classified. While different penalties for misclassifying TB and non-TB respondents were trialled a further investigation of the classification tree could have been carried out.

By using the same discretization levels for both the female and male dataset some of the levels had less values in them than ideal. For instance 68% of females were married before the age of 24 compared to 36% of males, and only 7% of females were married after the age of 24, compared to 26% of males. The difference in the percentage of respondents in each variables levels may have influenced the results of the most statistically likely directed acyclic graph.

The statistically likely directed acyclic graphs investigated the most likely associations between variables for all of India's female and male population samples. By only splitting the data by the respondent's gender there were around 100,000 respondents in each directed acyclic graph. If the data had been split into more specific sub-populations (such as wealth, education, state, or age), the associations between the variables for different demographic groups could have been investigated more specifically. This may have led to different statistically likely directed acyclic graphs depending on the sub-population.

4.10 Future directions

This thesis has investigated a summary of the variables thought to influence the prevalence of TB. The classification trees, logistic regression, and statistically likely directed acyclic graph work can easily be extended into more detailed, specific studies. This study used the respondents self-reported TB status, it could be extended using their medical history, TB symptoms, or by obtaining TB blood tests. Other variables can be trialled to be included, and the results can be compared across other countries.

The NFHS-3 survey was conducted for the purpose of obtaining an overview of India's health. It was not conducted for an in-depth analysis of the level of TB in India. As a results some variables (such as household crowding, vitamin D levels, level of immunization, weekly hours of free time, change in weight over the past year, HIV status, or savings level) were not included in the NFHS-3 survey. By including these variables in an analysis the accuracy of the logistic regression may improve. The classification tree may also perform better with these variables included. There may also be other variables which improve the models to create a more robust global TB indicator.

This analysis was specific to India, however it could be expanded to other countries. The NFHS-3 survey used in this analysis was co-ordinated with the Demographic and Health Surveys as described in Section 1.6. DHS have provided support for surveys in over 90 countries, some neighbouring India such as Pakistan, Nepal, Bangladesh, and Sri Lanka. The survey instruments used in these countries were similar to those used in India, however the TB status of the respondents was generally not available. The TB status of respondents is a vital field to know when testing how well a model can predict the TB status of respondents. A survey in Pakistan which includes TB questions is currently underway and may include questions as to whether the respondent has TB [98]. (The previous survey in Pakistan also included TB questions, however they asked respondents if they had heard of TB, how TB is spread, if TB can be cured, and did not ask if the respondent had TB.) If the TB status of the respondents was included, this survey would be an ideal country to analyse and to compare the results to India. There were some other countries whose TB status had been collected, however these surveys did not report the respondents height and weight.

Ideally, surveys which included people's TB status as well as the variables included in the logistic regression would be available from several countries. The optimal logistic regression model to predict respondents TB status could be applied to each country and the accuracy of the model tested using cross validation. It would be interesting to see for which countries the logistic regression is able to accurately predict the TB status of respondents and for which countries the logistic regression does not perform well. For countries where the logistic regression model found from the Indian dataset did not perform well, reasons as to why this is the case could be investigated.

The TB status provided was self reported and not from a medical test. There may be a large difference from what people said their TB status was, and their TB status from medical tests. If all the respondents had been tested for TB the number of respondents reported to have (or not have) TB may have changed enough to affect the classification tree and logistic regression model. If all respondents were asked if they had a persistent cough which had lasted for at least 4 weeks sometime in the past this could indicate respondents who had TB but did not realise it. The NFHS-3 survey also did not question if the respondent had had TB in the past and was now cured.

Understanding who is likely to be infected with latent TB, who is likely to develop active TB, and who is likely to be fully treated for TB helps to understand the pathways of TB in a population. Future studies into TB could use a longitudinal study to create predictions as to when respondents were likely to develop active TB. The optimal logistic regression model misclassified some TB and non-TB respondents. Determining which groups within the population were most likely to be misclassified could help to understand where the model had difficulties.

The statistically likely directed acyclic graphs were run for all of India in a gender-specific manner. The associations between variables may not be the same for sub-populations from specific states, educational brackets, age brackets, or TB status. By re-running the statistically likely directed acyclic graphs on a more detailed view of the data the associations for specific groups can be analysed, compared, and contrasted. In a similar way the statistically likely directed acyclic graphs can be obtained from data for different countries to determine if the associations across countries are similar. Altering the discretization method would affect the number of respondents in each category which potentially could affect which directed acyclic graph was the most likely.

[This page intentionally left blank]

Chapter 5

Appendix

5.1 Variable summary

The summary of the variables used in the secondary analysis of this thesis are provided below. For each variable the number and percentage of female and male respondents in each category are shown. For example table 1 shows that there were 118,385 female respondents and 72,164 male respondents in the NFHS-3 survey who did not have TB. This was 99.6% of the female population and 99.4% of the male population. Table 1 also shows that there were 472 female respondents and 443 male respondents from the NFHS-3 survey who did have TB. This was 0.4% of the female respondents and 0.6% of the male respondents.

For continuous variables the minimum, 1st quartile, median, mean, 3rd quartile and maximum value are shown, along with the number and percentage or respondents in each category. The discretization method for continuous variables involved having approximately the same number of respondents in each category, sensible limits for each category, and a sensible number of categories. For example table 23 shows the age of the household head. It was a continuous variable with a minimum recorded value of 14 and maximum recorded value of 95 (don't know and missing values were recorded as 98 and 99 and imputed). Age of household head was split into categories 15–34, 35–44, 45–54, 55–64, and $\geq$ 64. Using these splits each category had between 10% and 30

1. **Respondent suffers from TB**

	0 - no	1 - yes
Female	118385 (99.6%)	472 (0.4%)
Male	72164(99.4%)	443 (0.6%)

2. **Keep it a secret if a family member gets TB**

	0 - no	1 - yes	3 - unsure
Female	83880 (71%)	16011 (13%)	18966 (16%)
Male	54397 (75%)	10498 (14%)	7712 (11%)

3. **Respondent believes TB can be cured**

	0 - no	1 - yes	3 - unsure
Female	7980 (7%)	84809 (71%)	14820 (12%)
Male	3850 (5%)	57682 (79%)	11075 (15%)

4. **State**

	1 - northern	2 - central	3 - eastern	4 - southern
Female	22125 (19%)	28916 (24%)	34089 (29%)	33727 (28%)
Male	8251 (11%)	17982 (25%)	17337 (24%)	29037 (40%)

5. **Lives in city/town/country**

	0 - capital, lage city	1-small city	2-town	3-countryside
Female	25499 (21%)	8878 (7%)	20431 (17%)	64049 (54%)
Male	19997 (28%)	4905 (7%)	12565 (17%)	35140 (48%)

6. **If PSU covered, year Anganwadi/ICDS began operation**

	0 - not covered	1 - 1956-1985	2 - 1986-1995	3 - 1996-2006
Female	34031 (29%)	21010 (18%)	32442 (27%)	31374 (26%)
Male	20844 (29%)	14024 (19%)	19683 (27%)	18056 (25%)

7. **Education in single years**

	1 -none	2 - 1-5 years	3 - 6-8 yrs	4 - 9 yrs	5 - 10-11 yrs
Female	**38442 (32%)**	16863 (14%)	18548 (16%)	10899 (9%)	14614 (12%)
Male	**10655 (15%)**	11088 (15%)	13063 (18%)	9431 (13%)	11549 (16%)

	6 - 12-14 yrs	7 - $\geq$ 15 yrs
Female	10189 (9%)	9302 (8 %)
Male	8533 (12%)	8288 (11%)

8. **Partners highest year of education** – (data for female respondents only)

	1 - not applicable	2 - none	3 -1-5	4 - $\geq$ 6
Female	29010 (24%)	19828 (17%)	59646 (50%)	10373 (9%)

9. **National HIV weight**

	0 - lowest	1 - middle	2 - highest	3 - no information
Female	17359 (15%)	26405 (22%)	6912 (6%)	68181 (57%)
Male	22593 (31%)	24762 (34%)	6909 (10%)	18343 (25%)

10. **Total number of children ever born**

	0 - 0	1 -1	2 -2	3 - $\geq$ 3
Female	37668 (32%)	13804 (12%)	23414 (20%)	43971 (37%)
Male	32119 (44%)	6883 (9%)	11841 (16%)	21764 (30%)

11. **Number of living children**

	0 - 0	1 - 1	2 - 2	3 - 3	4 - 4	5 -$\geq$ 5
Female	38284 (32%)	15187 (13%)	25556 (22%)	19153 (16%)	10797 (9%)	9880 (8%)
Male	32440 (45%)	7573 (10%)	12987 (18%)	9463 (13%)	5166 (7%)	4978 (7%)

12. **Number of sons at home**

	0 - 0	1 -1	2 -2	3 - 3+
Female	54880 (45%)	24437 (29%)	21329 (18%)	8211 (7%)
Male	41578 (57%)	16574 (23%)	10354 (14%)	4101 (6%)

13. **Number of daughters at home**

	0 - 0	1 -1	2 -2	3 - $\geq$ 3
Female	63867 (54%)	32507 (27%)	15120 (13%)	7363 (6%)
Male	45542 (63%)	15785 (22%)	7524 (10%)	3756 (5%)

14. **Number of household members**

	1 - 1-3	2 - 4	3 - 5	4 - 6
Female	17349 (15%)	22799 (19%)	23339 (20%)	18153 (15%)
Male	12227 (17%)	14430 (20%)	13912 (19%)	10436 (14%)

	5 - 7-8	6 - ≥ 9
Female	20458(17%)	16759 (14%)
Male	11852 (16%)	9750 (13%)

15. **Age at first marriage**

	0 - 0-18	1 - 19-23	2 - 24+	3 - not married
Female	**55790 (47%)**	25524 (21%)	7875 (7%)	**29668 (25%)**
Male	**9636 (13%)**	16921 (23%)	18574 (26%)	**27476 (38%)**

16. **Age at first intercourse**

	0 - 0-18	1 - 19-23	2 - 24+	3 - virgin
Female	**55788 (47%)**	25785 (22%)	**7872 (7%)**	29412 (25%)
Male	**12276 (17%)**	18383 (25%)	**17973 (25%)**	23975 (33%)

17. **Total number of sexual partners**

	0 - 0	1-1	2-≥ 2
Female	29690 (25%)	**87330 (73%)**	1837 (2%)
Male	23975 (33%)	**39182 (54%)**	9450 (13%)

18. **Cooking done under a chimney** – (data for female respondents only)

	0 - No	1 - Yes	9 - Missing
Female	65439 (55%)	8324 (7%)	45094 (38%)

19. **Frequency of waching television**

	0 - not at all	1 -less than weekly	2 - weekly	3 - daily
Female	29859 (25%)	12175 (10%)	14360 (12%)	62463 (53%)
Male	9031 (12%)	11460 (16%)	11660 (16%)	40456 (56%)

20. **Frequency of reading newspaper or magazine**

	0 - not at all	1 -less than weekly	2 - weekly	3 - daily
Female	**64953 (55%)**	18082 (15%)	15640 (13%)	**20182 (17%)**
Male	**19632 (27%)**	10740 (15%)	13403 (18%)	**28832 (40%)**

21. **Food cooked on stove chullah open fire** - (data for female respondents only)

	1 - Stove	2 - Chullah	3 - Open fire	9 -Other/Missing
Female	4719 (4%)	60957 (51%)	8557 (7%)	44624 (38%)

22. **House type**

	1 - kaccha	2 - semi-pucca	3 -pucca
Female	11127 (9%)	44093 (37%)	63637 (54%)
Male	6,246 (9%)	25,937 (36%)	40,424 (56%)

23. **Age of household head**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	14	36	45	46.17	55	95
Male	15	36	45	45.8	54	95

	15 –34	35 – 44	45 – 54	55 – 64	≥65
Female	20525 (17%)	33720 (28%)	33860 (28%)	18417(15%)	12335 (10%)
Male	13674 (19%)	19309 (27%)	21883 (30%)	10745 (15%)	6996 (10%)

24. **Current age of respondent**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	15	21	28	29.37	37	49
Male	15	22	30	34.01	40	54

	15-24	25-34	35 - 44	44-54
Female	43532 (37%)	36823 (31%)	28816 (24%)	9686 (8%)
Male	24767 (34%)	20227 (28%)	16665 (23%)	10948 (15%)

25. **Wealth index factor score**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	-1.75	-0.76	0.02	0.08	0.89	2.37
Male	-1.74	-0.70	0.04	0.09	0.85	2.40

	-1.75 – -0.75	-0.75 – -0.25	-0.25 – 0.25	0.25 – 0.75
Female	30195 (25%)	19360 (16%)	17867 (15%)	16984 (14%)
Male	16899 (23%)	12332 (17%)	11748 (16%)	11115 (15%)

	0.75 – 1.25	1.25 – 2.5
Female	16114 (14%)	18337 (15%)
Male	9960 (14%)	10553 (15%)

26. **Standard of living index**

	1 - low	2 - medium	3 - high	9 - missing
Female	20302 (17%)	35538 (30%)	60271 (51%)	2746 (2%)
Male	11125 (15%)	22443 (31%)	37241 (51%)	1798 (2%)

27. **BMI**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	5.40	18.24	20.32	21.06	23.07	68.03
Male	6.06	18.32	20.24	20.77	22.67	74.77

	<18.5	18.5-20	20-22	22-25	>25
Female	33,545 (28%)	21,580 (18%)	24,245 (20%)	21,869 (18%)	17,618 (15%)
Male	19,704 (27%)	14,410 (20%)	16,329 (22%)	14,089 (19%)	8,075 (11%)

28. **Weight in kg**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	15.1	41.7	47.1	48.84	54.1	160.9
Male	16.3	48.9	54.7	56.31	62.1	173.0

29. **Height in cm**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	100.3	148.4	152.1	152.2	155.8	199.3
Male	80.0	160.3	164.5	164.5	168.6	199.1

30. **Hemoglobin test result adjusted for altitude**

	Min	1st Qu	Median	Mean	3rd Qu	Max
Female	20	108	118	116.4	127	229
Male	22	128	141	139.2	152	238

	≤120	120-130	130-140	140-150	≥150
Female	**66227 (56%)**	28734 (24%)	16760 (14%)	5699 (5%)	**1437 (1%)**
Male	**11511 (16%)**	9303 (13%)	14939 (21%)	15797 (22%)	**21057 (29%)**

5.2 Example of survey instrument used

NATIONAL FAMILY HEALTH SURVEY, INDIA 2005-2006 (NFHS- 3)
MAN'S QUESTIONNAIRE [STATE NAME]
IDENTIFICATION
STATE
DISTRICT
TEHSIL/TALUK
CITY/TOWN/VILLAGE
MEGA CITY/LARGE CITY/SMALL CITY/LARGE TOWN/SMALL TOWN/RURAL
(MEGA CITY=1, LARGE CITY=2, SMALL CITY=3, LARGE TOWN =4, SMALL TOWN=5, RURAL=6)
PSU NUMBER
HOUSEHOLD NUMBER
NAME AND LINE NUMBER OF MAN
ADDRESS OF HOUSEHOLD

SECTION 1. RESPONDENT'S BACKGROUND

INTRODUCTION AND INFORMED CONSENT

Namaste. My name is (INSERT NAME) and I am working with (NAME OF ORGANIZATION). We are conducting a national survey about the health of men, women and children. We would very much appreciate your participation in this survey. Several different health-related topics will be discussed including use of health services, the quality of health care, marital and sexual relationships, and infectious diseases.This information will help the government to assess health and information needs and to better plan health services. The survey usually takes about 30 minutes to complete. Whatever information you provide will be kept strictly confidential and will not be shown to other persons.

Participation in this survey is voluntary and if you choose to participate, you may withdraw at any time. However, we hope that you will take part in this survey since your participation is important.

At this time, do you want to ask me anything about the survey?

ANSWER ANY QUESTIONS AND ADDRESS RESPONDENT'S CONCERNS.

In case you need more information about the survey, you may contact the person listed on the card that has already been given to your household.

May I begin the interview now?

Signature of interviewer: Date:

RESPONDENT AGREES TO BE INTERVIEWED, 1
RESPONDENT DOES NOT AGREE TO BE INTERVIEWED, 2, END

NO.	QUESTIONS AND FILTERS		CODE	SKIP
101	RECORD THE TIME.	Hour		
		Minutes		
102	How long have you been living continuously in (NAME OF	Years		
	CURRENT PLACE OF RESIDENCE)?	Always	95	104
	(IF LESS THAN ONE YEAR, RECORD '00' YEARS)	Visitor	96	104
103	Just before you moved here, did you live in a city, in a town, or in	City	1	
	the countryside?	Town	2	
		Countryside	3	
104	In the last 12 months, on how many separate occasions have you	Number of trips away		
	traveled away from your home community for at least one night?	None	0	106
105	In the last 12 months, have you been away from your home.	Yes	1	
	community for more than 1 month at a time?	No	2	
106	In what month and year were you born?	Month		
		Don't know month	98	
		Year		
		Don't know year	9998	
114	Do you read a newspaper or magazine almost every day, at least	Almost every day	1	
	once a week, less than once a week or not at all?	At least once a week	2	
		Less than once a week	3	
		Not at all	4	
115	Do you listen to the radio almost every day, at least once a week,	Almost every day	1	
	less than once a week or not at all?	At least once a week	2	
		Less than once a week	3	
		Not at all	4	
116	Do you watch television almost every day, at least once a week,	Almost every day	1	
	less than once a week or not at all?	At least once a week	2	
		Less than once a week	3	
		Not at all	4	
201	Now I would like to ask about any children you have had during	Yes	1	
	your life. I am interested only in the children that are biologically	NO	2	
	yours. Have you ever fathered any children with any woman?	Don't know	8	206

202	Do you have any sons or daughters that you have fathered who are now living with you?	Yes No	1 2	 204
203	How many sons live with you? And how many daughters live with you? IF NONE, RECORD 00'.	Sons at home Daughters at home		
204	Do you have any sons or daughters you have fathered who are alive but do not live with you?	Yes No	1 2	 206
205	How many sons are alive but do not live with you? And how many daughters are alive but do not live with you? IF NONE, RECORD 00'.	Sons elsewhere Daughters elsewhere		
206	Have you ever fathered a son or a daughter who was born alive but later died? DON'T KNOW IF NO, PROBE: Any baby who cried or showed signs of life but did not survive?	Yes No Don't know	1 2 8	 208
207	How many boys have died? And how many girls have died? IF NONE, RECORD 00'. GIRLS DEAD	Boys dead Girls dead		
208	(In addition to the children that you have just told me about), do you have: a. Any other living sons or daughters who are biologically your children but who are not legally yours or do not have your last/family name? b. Any other sons or daughters who died who were biologically your children but who were not legally yours or did not have your last/family name? NO TO BOTH, CONTINUE OTHER, PROBE AND CORRECT 201 -207 AS NECESSARY			
209	SUM ANSWERS TO 203, 205, AND 207, AND ENTER TOTAL IF NONE, RECORD '00'	Total children		
407	Have you been married once or more than once?	Once More than once	1 2	409 409a
409	In what month and year did you get married?	Month Don't know month		
409A	Now I would like to ask about when you married your first wife. In what month and year was that?	Year Don't know year	411 9998	
410	How old were you when you (first) got married?	Age		
608	Do you currently smoke cigarettes or bidis?	Yes No	1 2	 610
609	In the last 24 hours, how many cigarettes or bidis did you smoke? IF NONE, RECORED '00'	Cigarettes/bidis		
610	Do you currently smoke or use tobacco in any other form?	Yes No	1 2	 612
611	In what other form do you currently smoke or use tobacco? Any other form? RECORD ALL MENTIONED.	Cigar/pipe Paan masala Ghutka Other chewing tobacco Snuff Other (specify)	A B C D E F	
612	Do you drink alcohol?	Yes No	1 2	 612
613	How often do you drink alchol: almost every day, about once a week, or less often?	Almost every day About once a week Less often	1 2 3	

614	Have you ever herd of an illness called tuberculosis or TB?	Yes	1	
		No	2	618
615	How does tuberculosis spread from one person to another?	Through the air when coughing of sneezing	A	
		Through sharing utensils	B	
		Through touching a person with TB	C	
		Through food	D	
		Through sexual contact	E	
		Through mosquito bites	F	
		Other (specify)	X	
		Don't know	Z	
616	Can tuberculosis be cured?	Yes	1	
		No	2	
		Don't know	8	
617	If a member of your family got tuberculosis, would you want it to remain a secret or Not?	Yes, remain a secret	1	
		No	2	
		Don't know/not sure/ /depends	8	

Notes:
CODE stands for the coding categories used.

5.3 Full list of variables used in the initial analysis

Female dataset:

- Wealth index factor score
- Any usual resident of the household suffers from TB
- Body Mass Index
- Rohrer's Index
- Has received medical treatment for TB
- Woman's weight in kilograms
- Woman's height in centimetres
- Education in single years
- Literacy
- Cooking done under a chimney
- Highest year of education
- Haemoglobin test result adjusted for altitude
- State
- Haemoglobin test result
- Years since first marriage
- Highest educational level
- Type of cooking fuel
- Wealth index
- Standard of Living Index
- Partner's education level

- Partner's educational attainment
- Current age - respondent
- Partner's highest year of education
- Age of household head
- Food cooked on stove, chullah, open fire
- Age at first marriage Marital duration (grouped)
- Frequency of reading newspaper or magazine
- Reduce chance of AIDS: have 1 sex partner
- If PSU covered, year Anganwadi/ICDS centre started
- Get AIDS from mosquito bites
- Age at first intercourse
- Acres of agricultural land
- Reduce chances of AIDS by always using a condom
- Get AIDS by sharing food with person
- House has windows with glass
- Can a healthy person have AIDS
- Daughters at home
- Can tuberculosis be cured
- Tuberculosis spread by: don't know
- Anaemia level (from V456) adjusted for altitude
- Number of household members
- Ever heard of AIDS
- Reduce risk of getting AIDS by not having unprotected sex
- Sons at home
- House has windows with curtains or shutters
- Smokes other chewing tobacco
- Current marital status
- Currently/formerly/never married
- Number of unions
- Keep secret when family member gets TB
- House has windows with screens
- Smokes nothing
- Tuberculosis spread through food
- Tuberculosis spread by: sharing clothes
- Tuberculosis spread by: smoking/bidis
- Tuberculosis spread by: air when coughing
- Tuberculosis spread by: mosquito bite
- Heard of Tuberculosis or TB
- Tuberculosis spread by: sharing utensils

- House has any windows
- Number of eligible women in household
- De facto place of residence
- Number of other wives
- Tuberculosis spread by: touching a person
- Tuberculosis spread by: sexual contact
- Tuberculosis spread by: blood/blood transfusion
- Tuberculosis spread by: spit/sputum/stepping on spit
- Tuberculosis spread by: other
- Type of place of residence
- Husband lives in house
- Household in PSU covered by Anganwadi/ICDS centre
- Number of cigarettes in last 24 hours
- Number of children 5 and under in household
- Sex of household head
- Smokes cigarettes/bidis
- Frequency of alcohol use
- Drinks alcohol
- Do you have: diabetes
- Smokes paan masala
- Smokes ghutka
- Smokes other
- Uses snuff
- Smokes pipe/cigar
- Wife rank number
- Chewing tobacco

Male dataset:

- Wealth Index factor score
- Primary sampling unit
- Any usual resident of the household suffers from TB
- Body Mass Index
- Rohrer's Index
- Date of birth (CMC)
- Man's weight in kilograms
- Man's height in centimetres
- Years since first marriage
- Man's age in years

- Current age - respondent
- Respondent's year of birth
- Haemoglobin test result
- Haemoglobin level adjusted for altitude
- Total children ever born
- Number of living children
- Age at first intercourse
- Education in single years
- Relationship to household head
- Age of household head
- Marital duration (grouped)
- Age in 5-year groups
- Highest year of education
- State
- Region
- Women fathered children with
- Educational attainment
- Length of interview in minutes
- House type (as defined in NFHS-2)
- Literacy
- Native language of respondent
- Sons at home
- If yes, year Anganwadi/ICDS centre began operation
- Anaemia level (from MV456)
- Wealth Index
- Standard of Living Index
- Highest educational level
- Ever participated in a literacy program outside
- Frequency of watching television
- Frequency of reading newspaper or magazine
- Number of wives, partners
- Daughters at home
- Number of unions
- City/Town/Countryside
- Recent sexual activity
- Intend to wait until married to have sex
- Tuberculosis spread through: food
- Smokes other chewing tobacco
- Times away from home in last 12 months

- Wife/partner lives with respondent
- Number of wives, partners
- Type of caste or tribe
- Type of place of residence
- De facto place of residence
- Type of place of residence
- De facto place of residence
- Tuberculosis can be cured
- Number of cigarettes in last 24 hours
- Smokes nothing
- Number of eligible men in household
- Number of household members
- Caste or tribe
- Tuberculosis spread by: sharing utensils
- Years lived in place of residence
- Tuberculosis spread by: air when coughing or sneezing
- Respondent's month of birth
- Keep secret when family member gets TB
- Tuberculosis spread by: touching a person with TB
- PSU covered by Anganwadi/ICDS centre
- Slum designation by supervisor
- Frequency of listening to radio
- Religion
- PSU altitude in meters
- Heard of Tuberculosis or TB
- Slum designation by census
- Smokes cigarettes
- Don't know how tuberculosis is spread
- Away for more than one month
- Type of place of previous residence
- Household structure
- Result of measurement - haemoglobin
- Result of measurement - HIV
- Smokes ghutka
- Tuberculosis spread by: sexual contact
- Tuberculosis spread by: mosquito bites
- Smokes other
- Smokes pipe
- Smokes paan masala

- Tuberculosis spread by: blood/blood transfusion
- Smokes snuff
- Tuberculosis spread by: smoking/bidis/cigarette
- Tuberculosis spread by: other
- Tuberculosis spread by: sharing clothes/bed/towel
- Tuberculosis spread by: spit/sputum/stepping on spit
- Sex of household head
- Primary sampling unit
- Childhood place of residence
- Ethnicity
- Usual resident or visitor
- Current pregnancy wanted
- Married to mother when first
- Chewing tobacco
- Have ever been married

5.4 Calculated fields

The wealth index was calculated by principal components analysis using the following variables:

- Drinking and non-drinking water source
- Household electrification and possessions
- Main floor, roof and wall material
- Type of windows and cooking fuel
- Number of members per sleeping room
- Household member having a bank or post office account
- Domestic servant in household
- Ownership of agricultural land or house
- Toilet facility

Housetype was defined as either kaccha, pucca or semi-pucca as in Table 5.2.

House Type	Main Material
Kaccha	
material on floor	Mud/clay/earth, sand, dung, raw wood planks, palm, bamboo
material on wall	No walls, cane/palm/trunks, mud, grass/reeds/thatch, bamboo with mud, stone with mud, plywood, cardboard, unburnt brick
material on roof	thatch/palm leaf, mud, sod/mud and grass mixture, plastic/polythene sheeting, rustic mat, palm/bamboo, raw wood planks/timber, unburnt bricks, loosely packed stone
Pucca	
material on floor	Brick, stone, parquet, polished wood, vinyl, asphalt strips, ceramic tiles, cement, carpet, polished stone/marble/granite
material on wall	Cement/concrete, stone with lime/cement, burnt brickes, cement blocks, wood planks/shingles, GI/metal/asbostos
material on roof	Metal/GI, wood, calamine/cement fiber, asbestos, cement/concrete, roofing shingles, tiles, slate, burnt brick
Semi-Pucca	Any combination of Kaccha and Pucca materials
Missing	If any of the floor, exterior wall, roof materials were missing

Table 5.2: House type calculation.

5.5 TB and non-TB distribution functions

The full summary of the variables included in the distribution function testing in Table 3.11 in Section 3.2 is shown below. The tests which did not show significantly different distributions for the TB and non-TB populations have been shaded. The KS test was not designed for categorical variables, these results have been marked with a *. The results of the permuted KS test are shown for both 10,000 and 100,000 permutations. The band width of the TB and non-TB data and the maximal distance between the TB and non-TB CDF for both the 5% and 1% level of testing are shown. If the maximal distance was negative, the null hypothesis that the TB and non-TB data came from the same distribution was failed to be rejected.

5.6 Modelling TB cases

Both the female and male datasets were approached in the same way. Using the variables identified from the literature review three types of models were created.

- The first model did not include any interaction terms or transformations of the variables.
- The second model did not include any interaction terms but allowed transformations of the variables.
- The third model had interaction terms and transformations.

As there were many variables available they were split into similar groups and each group was initially analysed separately. The groups were ‘basic information’, ‘health information’, ‘wealth information’.

variable name	maximum CDF dist apart	**Chi-Square Test**	**KS test** P value at 1%	**Permuted KS test** P value, perms = 10,000	100,000	**CDF confidence bands** band widths at 5% level TB	non-TB	Dist	band widths at 1% level TB	non-TB	Dist
Respondent suffers from TB	1.000	***	0.000*	0.000	0.000	0.063	0.004	0.934	0.075	0.005	0.920
Any household resident has TB	0.983	***	0.000*	0.000	0.000	0.063	0.004	0.917	0.075	0.005	0.904
Education in single years	0.245	***	0.000*	0.000	0.000	0.063	0.004	0.179	0.075	0.005	0.166
Body mass index	0.232	***	0.000	0.000	0.000	0.064	0.004	0.165	0.076	0.005	0.152
Wealth index factor score	0.216	***	0.000	0.000	0.000	0.063	0.004	0.150	0.075	0.005	0.137
Freq reading newspaper/magazing	0.183	**	0.000*	0.000	0.000	0.063	0.004	0.117	0.075	0.005	0.104
House type	0.181	***	0.000*	0.000	0.000	0.063	0.004	0.114	0.075	0.005	0.101
Freq. watching tv	0.177	**	0.000*	0.000	0.000	0.063	0.004	0.110	0.075	0.005	0.097
Number of children born	0.175	***	0.000*	0.000	0.000	0.063	0.004	0.109	0.075	0.005	0.096
Number of living children	0.159	**	0.000*	0.000	0.000	0.063	0.004	0.093	0.075	0.005	0.080
Partners highest year of education	0.154	.	0.000*	0.000	0.000	0.068	0.005	0.081	0.082	0.006	0.066
Current age of respondent	0.124	*	0.000	0.000	0.000	0.063	0.004	0.058	0.075	0.005	0.045
Number of daughters at home	0.112	*	0.000*	0.000	0.000	0.063	0.004	0.045	0.075	0.005	0.032
Age at first marriage	0.109	.	0.000 *	0.000	0.000	0.068	0.005	0.036	0.082	0.006	0.022
State	0.099	**	0.000 *	0.000	0.0001	0.063	0.004	0.032	0.075	0.005	0.019
Age of household head	0.091	.	0.001	0.000	0.0004	0.063	0.004	0.024	0.075	0.005	0.011
Number of sons at home	0.089	*	0.001 *	0.000	0.0002	0.063	0.004	0.022	0.075	0.005	0.009
Years Anganwadi centre began	0.091	*	0.005 *	0.005	0.004	0.072	0.005	0.014	0.086	0.006	-0.001
Hemoglobin level	0.076	.	0.008	0.005	0.005	0.066	0.004	0.006	0.079	0.005	-0.007
Place of residence	0.067	.	0.027 *	0.004	0.004	0.063	0.004	0.001	0.075	0.005	(0.012)
Cooking done under a chimney	0.067	***	0.071 *	0.001	0.000	0.070	0.005	-0.009	0.084	0.006	-0.024
Number sexual partners	0.026	*	0.942 *	0.002	0.002	0.068	0.005	-0.046	0.082	0.006	-0.061
HIV weight	0.079	.	0.145	0.082	0.082	0.094	0.006	-0.021	0.113	0.007	-0.041
Number of household members	0.039		0.458 *	0.189	0.192	0.063	0.004	-0.027	0.075	0.005	-0.040
Believes TB can be cured	0.036	*	0.618 *	0.027	0.028	0.064	0.004	-0.033	0.077	0.005	-0.047
Food cooked on stove, open fire	0.014	*	1.000 *	0.616	0.613	0.070	0.005	-0.061	0.084	0.006	-0.076

Table 5.3: Female TB and non-TB distribution function results for the KS test, permuted KS test, CDF confidence bands, and Chi-squared test. Variables have been ordered by level of significance. Variables which did not have significantly different distributions between the TB and non-TB distributions have been shaded. The chi-squared results are shown in terms of significance where *** is the highest level.

variable name	maximum CDF dist apart	**Chi-Square Test**	**KS test** P value at 1%	**Permuted KS test** P value, perms = 10,000	100,000	**CDF confidence bands** band widths at 5% level TB	non-TB	Dist	band widths at 1% level TB	non-TB	Dist
Respondent suffers from TB	1.000	***	0.000*	0.000	0.000	0.065	0.005	0.930	0.077	0.006	0.917
Any usual resident has TB	0.984	***	0.000*	0.000	0.000	0.065	0.005	0.915	0.077	0.006	0.901
Body Mass Index	0.313	***	0.000	0.000	0.000	0.066	0.005	0.242	0.079	0.006	0.228
Education in single years	0.280	***	0.000*	0.000	0.000	0.065	0.005	0.211	0.077	0.006	0.197
Wealth Index factor score	0.255	***	0.000	0.000	0.000	0.065	0.005	0.186	0.077	0.006	0.172
Total children ever born	0.253	***	0.000*	0.000	0.000	0.065	0.005	0.183	0.077	0.006	0.169
Current age - respondent	0.249	***	0.000	0.000	0.000	0.065	0.005	0.179	0.077	0.006	0.166
Number of living children	0.237	***	0.000*	0.000	0.000	0.065	0.005	0.167	0.077	0.006	0.153
Frequency of watching television	0.234	***	0.000*	0.000	0.000	0.065	0.005	0.164	0.077	0.006	0.151
House type	0.230	***	0.000*	0.000	0.000	0.065	0.005	0.160	0.078	0.006	0.147
Freq. reading newspaper/magazine	0.223	***	0.000*	0.000	0.000	0.065	0.005	0.153	0.077	0.006	0.139
Sons at home	0.176	**	0.000*	0.000	0.000	0.065	0.005	0.107	0.077	0.006	0.093
Hemoglobin level	0.182	***	0.000*	0.000	0.000	0.069	0.005	0.107	0.082	0.007	0.093
Daughters at home	0.145	*	0.000*	0.000	0.000	0.065	0.005	0.076	0.077	0.006	0.062
City/Town/Countryside	0.140	*	0.000*	0.000	0.000	0.065	0.005	0.070	0.077	0.006	0.057
State	0.130	*	0.000*	0.000	0.000	0.065	0.005	0.060	0.077	0.006	0.046
Age at first marriage	0.113	.	0.000	0.000	0.001	0.071	0.006	0.035	0.085	0.008	0.020
Age of household head	0.098	.	0.000	0.000	0.000	0.065	0.005	0.029	0.077	0.006	0.015
National mens HIV weight	0.104		0.002	0.000	0.001	0.075	0.006	0.023	0.090	0.007	0.007
Year Anganwadi center began	0.101	.	0.002	0.002	0.002	0.073	0.006	0.022	0.088	0.007	0.006
Total number of sexual partners	0.068	***	0.057*	0.004	0.003	0.070	0.006	-0.008	0.083	0.007	-0.023
Tuberculosis can be cured	0.041	*	0.474*	0.003	0.003	0.065	0.005	-0.030	0.078	0.006	-0.044
Number of household members	0.038	.	0.530*	0.240	0.235	0.065	0.005	-0.031	0.077	0.006	-0.045

Table 5.4: Male TB and non-TB distribution function results for the KS test, permuted KS test, and CDF confidence bands. Variables have been ordered by level of significance. Variables which did not have significantly different distributions between the TB and non-TB distributions have been shaded

Female Data set Information variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
V024	State	x	x	x
V025	Type of place of residence	x	x	x
V151	Sex of household head	x	x	x
V152	Age of household head	x	x	x
V136	Number of household members	x	x	
V137	Number of children 5 and under in household	x		x
V138	Number of eligible women in household	x	x	
V202	Sons at home	x		x
V203	Daughters at home	x	x	x
V501	Current marital status	x		
V503	Number of unions		x	
V504	Husband lives in house	x		
V511	Age at first marriage	x	x	
V512	Years since first marriage	x	x	x
V525	Age at first intercourse	x	x	x
V012	Current age - respondent	x	x	

Female Data set Health variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
V457	Anaemia level (from V456) adjusted for pregnancy	x	x	
V437	Woman's weight in kilograms	x		x
V438	Woman's height in centimetres		x	
V445	Body Mass Index	x	x	x
V446	Rohrer's Index			
SANGAYN	Household in PSU covered by Anganwadi/ICDS centre			x
SANGAYR	If PSU covered, year Anganwadi/ICDS centre began			x
V456	Haemoglobin test result adjusted for altitude			x
V754DP	Reduce chance of AIDS by only having 1 sex partner	x	x	
V754WP	Get AIDS by sharing food with person who has AIDS	x		
V463A	Smokes cigarettes/bidis	x		
V463D	Uses snuff		x	
V463E	Smokes paan masala		x	
V463F	Smokes ghutka		x	
V463G	Smokes other chewing tobacco	x		
V463X	Smokes other		x	
V463Z	Smokes nothing		x	
V474B	Tuberculosis spread by: sharing utensils	x		
V474D	Tuberculosis spread through food	x	x	
V475	Can tuberculosis be cured			x

Female Data set Wealth variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
SSLI	Standard of Living Index		x	
V190	Wealth index			
V191	Wealth index factor score	x	x	x
V702	Partner's highest year of education	x		
V133	Education in single years	x	x	x
S49	Food cooked on stove, chullah, open fire	x		
S50	Cooking done under a chimney		x	x

Male Data set Information variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
SMANGYR	If yes, year Anganwadi/ICDS centre began operation	x		
MV009	Respondent's month of birth			
MV010	Respondent's year of birth	x	x	x
MV012	Current age - respondent	x	x	x
MV024new	State grouped	x	x	
MV104	Years lived in place of residence	x		
MV105	Type of place of previous residence	x		
MV138	Number of eligible men in HH	x	x	x
MV151	Sex of household head		x	x
MV201	Total children ever born	x	x	x
MV202	Sons at home	x		
MV203	Daughters at home	x		
MV218	Number of living children	x	x	x
MV504	Wife/partner lives with respondent		x	
MV505	Number of wives, partners	x		
MV512	Years since first marriage	x	x	
MV525	Age at first intercourse	x	x	x
MV536	Recent sexual activity		x	x
SM025	City/Town/Countryside	x	x	x
MV040	PSU altitude in meters	x	x	
SMSTRUC	Household structure	x	x	
SM30	Any usual resident of the household suffers from TB	x		
MV476	Keep secret when family member gets TB	x	x	

Male Data set Health variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
MV463F	Smokes ghutka	x	x	
MV463G	Smokes other chewing tobacco	x	x	x
MV463Z	Smokes nothing	x	x	x
MV474A	Tuberculosis spread by: air when coughing or sneezing		x	x
MV474B	Tuberculosis spread by: sharing utensils	x	x	x
MV474C	Tuberculosis spread by: touching a person with TB	x	x	x
MV474D	Tuberculosis spread through: food	x	x	x
MV474H	Tuberculosis spread by: blood/blood transfusion	x		
MV475	Tuberculosis can be cured	x	x	x
MV438	Man's height in centimetres		x	x
MV445	Body Mass Index	x	x	x
MV456	Haemoglobin level adjusted for altitude	x	x	x
MV457	Anaemia level		x	x

Male Data set Wealth variables included in initial models				
Var Code	Variable Name	model 1	model 2	model 3
MV133	Education in single years	x	x	x
MV155	Literacy		x	x
MV159	Frequency of watching television	x	x	
MV191	Wealth Index factor score	x		
SMSLI	Standard of Living Index		x	x
SMNFHS2	House type		x	x
MV149	Educational attainment			

5.7 Optimal model cross validation results

The cross validated results from the optimal logistic regression are shown below. 10-fold cross validation was run and the results from each 10-th of the data are shown.

	Proportion of TB cases assigned as TB									
	1	2	3	4	5	6	7	8	9	10
0.0010	0.95	0.99	0.99	0.98	0.96	1.00	0.97	0.97	0.99	0.99
0.0015	0.91	0.97	0.96	0.93	0.91	0.97	0.89	0.91	0.92	0.96
0.0020	0.90	0.90	0.93	0.88	0.85	0.90	0.89	0.90	0.90	0.92
0.0025	0.84	0.87	0.86	0.80	0.81	0.86	0.86	0.89	0.88	0.87
0.0030	0.84	0.84	0.79	0.77	0.75	0.81	0.82	0.83	0.85	0.86
0.0035	0.82	0.82	0.75	0.73	0.70	0.80	0.82	0.79	0.84	0.84
0.0040	0.79	0.78	0.74	0.71	0.67	0.76	0.77	0.77	0.79	0.80
0.0045	0.75	0.74	0.70	0.68	0.64	0.69	0.73	0.70	0.77	0.77
0.0050	0.71	0.72	0.68	0.65	0.60	0.64	0.70	0.66	0.77	0.69
0.0055	0.70	0.67	0.63	0.58	0.59	0.57	0.64	0.62	0.74	0.63
0.0060	0.67	0.65	0.63	0.58	0.56	0.54	0.60	0.53	0.68	0.62
0.0065	0.63	0.64	0.61	0.57	0.53	0.49	0.57	0.53	0.64	0.56
0.0070	0.60	0.62	0.60	0.54	0.53	0.47	0.56	0.49	0.57	0.55
0.0075	0.59	0.60	0.57	0.52	0.48	0.47	0.53	0.46	0.53	0.54
0.0080	0.57	0.59	0.50	0.49	0.47	0.46	0.49	0.43	0.50	0.51
0.0085	0.54	0.55	0.47	0.48	0.47	0.46	0.46	0.42	0.50	0.51
0.0090	0.53	0.53	0.45	0.45	0.47	0.43	0.46	0.39	0.45	0.46
0.0095	0.52	0.52	0.42	0.42	0.47	0.41	0.43	0.39	0.39	0.46
0.0100	0.50	0.52	0.40	0.42	0.47	0.37	0.41	0.35	0.39	0.45

	Proportion of non-TB cases assigned as non-TB									
	1	2	3	4	5	6	7	8	9	10
0.0010	0.18	0.18	0.19	0.19	0.19	0.18	0.18	0.18	0.19	0.18
0.0015	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
0.0020	0.38	0.38	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.37
0.0025	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
0.0030	0.52	0.53	0.53	0.53	0.53	0.53	0.52	0.53	0.53	0.53
0.0035	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
0.0040	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
0.0045	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
0.0050	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
0.0055	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
0.0060	0.77	0.78	0.78	0.77	0.78	0.77	0.78	0.78	0.78	0.78
0.0065	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
0.0070	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
0.0075	0.83	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.84	0.84
0.0080	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
0.0085	0.86	0.86	0.86	0.86	0.87	0.86	0.87	0.86	0.87	0.86
0.0090	0.87	0.87	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
0.0095	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.88	0.89	0.89
0.0100	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.89	0.90	0.90

Table 5.5: Results of 10-fold cross validation from the optimal logistic regression model.

5.8 Nearest neighbour results

Variable	1 nearest neighbour	3 nearest neighbours	5 nearest neighbours	10 nearest neighbours
Wealth index factor score	12.22	12.22	12.22	13.50
Body Mass Index	12.71	12.71	12.71	13.43
Rohrer's Index	13.37	13.37	13.37	13.98
Woman's weight in kilograms	14.42	14.42	14.42	14.83
Woman's height in centimeters	16.94	16.94	16.94	16.95
Hemoglobin test result (adj. altitude)	25.40	25.40	25.40	22.84
Current age - respondent	29.44	29.44	29.44	26.10
Age of household head	27.75	27.75	27.75	25.63
State	28.40	28.40	28.40	25.63
Years since first marriage	32.58	32.58	32.58	30.67
Year Anganwadi/ICDS centre began	31.94	31.94	31.94	30.29
Education in single years	34.17	34.17	34.17	32.14
Acres of agricultural land	32.67	32.67	32.67	31.03
Age at first marriage	34.81	34.81	34.81	33.65
Number of household members	36.02	36.02	36.02	34.70
Wealth index	37.59	37.59	37.59	36.35
Partner's highest year of education	36.86	36.86	36.86	35.71
Marital duration	37.43	37.43	37.43	36.46
Type of cooking fuel	36.81	36.81	36.81	36.05
Cooking done under a chimney	37.76	37.76	37.76	37.06
Food cooked on stove, chullah, open fire	37.57	37.57	37.57	36.94
Can tuberculosis be cured	38.90	38.90	38.90	37.32
Frequency of reading newspaper or magazine	38.57	38.57	38.57	37.96
State	38.64	38.64	38.64	37.78
Number of eligible women in household	39.13	39.13	39.13	38.39
Sons at home	38.50	38.50	38.50	37.71
Keep secret when family member gets TB	38.87	38.87	38.87	37.83
De facto place of residence	39.18	39.18	39.18	38.28
Can a healthy person have AIDS	39.11	39.11	39.11	38.18
Standard of Living Index	39.04	39.04	39.04	38.02
Daughters at home	38.98	38.98	38.98	38.18
Husband lives in house	40.00	40.00	40.00	39.45
Heard of Tuberculosis or TB	40.04	40.04	40.04	39.14
Type of place of residence	40.13	40.13	40.13	39.53
Household covered by Anganwadi/ICDS centre	40.43	40.43	40.43	39.85
Smokes nothing	40.73	40.73	40.73	40.29
Sex of household head	40.71	40.71	40.71	40.21
Do you have: diabetes	40.78	40.78	40.78	40.18
Frequency of alcohol use	40.85	40.85	40.85	40.26

Table 5.6: Summary of the female dataset nearest neighbour re-substitution results.
Results have been ordered in descending order from the most reliable indicator of TB. The results from increasing the neighbourhood size between 1, 3, 5, and 10 nearest neighbours are shown. The values shown are the sum of the re-substitution errors for each variable when compared with all the the other variables.

Variable	1 nearest neighbour	3 nearest neighbours	5 nearest neighbours	10 nearest neighbours
Wealth index factor score	0.25	15.71	19.32	24.12
Body Mass Index	8.63	18.9	22.15	25.61
Rohrer's Index	9.94	18.17	22.21	26.51
Man's weight in kilograms	14.49	19.43	21.86	24.26
Man's height in centimeters	22.91	25.9	27.85	29.83
State	41.53	39.9	37.08	32.28
Hemoglobin test result (adj. altitude)	33.51	33.36	32.54	30.32
Current age - respondent	39.06	38.82	37.92	34.31
Year Anganwadi/ICDS centre began	41.94	40.87	39.7	35.06
Education in single years	39.94	37.75	36.89	35.03
Age of household head	38.49	38.02	37.37	35.12
Years since first marriage	41.02	41.27	40.6	38.48
Years lived in place of residence	41.67	41.22	40.78	38.84
Keep secret when family member gets TB	44.75	44.23	43.5	40.91
Tuberculosis can be cured	45.05	44.72	43.82	42.16
Number of household members	43.7	43.48	43.1	41.4
Wealth Index	45.21	43.92	43.63	42.5
City/Town/Countryside	44.89	44.31	43.72	42.83
Slum designation by supervisor	45.31	44.58	44.3	43.15
Total children ever born	44.76	44.7	44.61	42.91
State - grouped	45.91	45.45	44.59	43.11
Marital duration	44.96	44.79	43.55	42.28
Number of eligible men in HH	45.19	45.07	44.69	42.98
Standard of Living Index	45.63	45.51	44.78	42.86
Number of living children	45.08	44.96	44.8	43.28
Frequency of reading newspaper or magazine	45.34	45.04	44.26	42.77
House type	46.11	46.11	45.67	44.31
Relationship to household head	44.82	44.9	44.82	43.3
Frequency of watching television	46.27	46.2	45.67	44.05
Sons at home	46.04	46.01	46.03	45.39
Frequency of listening to radio	46.07	45.95	45.3	44.66
Number of cigarettes in last 24 hours	43.94	44.31	44.5	43.55
Literacy	46.4	46.48	46.41	45.69
Smokes nothing	46.44	46.48	46.25	45.57
Smokes cigarettes	46.67	46.79	46.84	46.21
Household covered by Anganwadi/ICDS centre	46.45	45.75	45.56	44.65
PSU altitude in meters	44.52	44.62	44.7	45.1
Daughters at home	46.03	46.31	46.26	45.12
Type of place of residence	46.64	46.69	46.61	45.85

Table 5.7: Summary of the male dataset nearest neighbour re-substitution results.
Results have been ordered in descending order from the most reliable indicator of TB. The results from increasing the neighbourhood size between 1, 3, 5, and 10 nearest neighbours are shown. The values shown are the sum of the re-substitution errors for each variable when compared with all the the other variables.

5.9 Independence testing

Variable 1	Variable 2	Indep.
Age	Total children ever born	****
Standard of Living Index	Wealth Index factor score	****
Body Mass Index	Man's weight in kilograms	****
Total number sexual partners	Age	****
Total children ever born	Sons at home	****
Frequency of reading newspaper or magazine	Education in single years	***
Total children ever born	Daughters at home	***
Total children ever born	Age	***
House type	Wealth Index factor score	***
Total number sexual partners	Total children ever born	***1
Age	Age of household head	***
Age	Sons at home	***
Total number sexual partners	Age	***
Keep secret when family member gets TB	Tuberculosis can be cured	**
Sons at home	Age	**
Total number sexual partners	Sons at home	**
House type	Standard of Living Index	**
Age	Daughters at home	**
National mens HIV weight	State	**
Education in single years	Wealth Index factor score	**
Total number sexual partners	Daughters at home	**
Frequency of watching television	Wealth Index factor score	**
National mens HIV weight	De facto place of residence	**
Daughters at home	Age	**
De facto place of residence	Wealth Index factor score	**
Frequency of reading newspaper or magazine	Wealth Index factor score	**
Daughters at home	Sons at home	**
Frequency of watching television	Standard of Living Index	*
House type	De facto place of residence	*
Wealth Index factor score	Man's weight in kilograms	*
Education in single years	Standard of Living Index	*
Man's height in centimeters	Man's weight in kilograms	*
Frequency of reading newspaper or magazine	Frequency of watching television	*
Frequency of reading newspaper or magazine	House type	*
Frequency of watching television	House type	*
Frequency of reading newspaper or magazine	Standard of Living Index	*
Frequency of watching television	Education in single years	*
Education in single years	House type	*
Number of household members	Total children ever born	*

Table 5.8: The most nonindependent variables for male dataset.
The independence level has been calculated from the P-value methods as described in Section 2.2.2. The results have been ordered in terms of non-independence with the smallest values being the least independent.

Variable 1	Variable 2	Indep.
Age at first intercourse	Age at first marriage	****
Food cooked on stove, chullah, open fire	Cooking done under a chimney	****
Total number of sexual partners	Age at first marriage	****
Wealth index factor score	Standard of Living Index	****
Age at first marriage	Total children ever born	****
Total children ever born	Sons at home	****
Total number of sexual partners	Total children ever born	****
Frequency of reading newspaper or magazine	Education in single years	***
Wealth index factor score	Cooking done under a chimney	***
Food cooked on stove, chullah, open fire	Wealth index factor score	***
Wealth index factor score	House type	***
Total children ever born	Daughters at home	***
Keep secret when family member gets TB	Can tuberculosis be cured	***
Age	Total children ever born	***
Partner's highest year of education	Age at first marriage	***
Partner's highest year of education	Total number of sexual partners	**
Wealth index factor score	Education in single years	**
Age at first marriage	Sons at home	**
House type	Standard of Living Index	**
Total number of sexual partners	Sons at home	**
Age	Age at first marriage	**
Age	Total number of sexual partners	**
Frequency of reading newspaper or magazine	Wealth index factor score	**
Food cooked on stove, chullah, open fire	De facto place of residence	**
Wealth index factor score	Frequency of watching television	**
Partner's highest year of education	Total children ever born	**
Education in single years	Age at first marriage	**
National womens HIV weight	State	**
Food cooked on stove, chullah, open fire	Education in single years	**
Cooking done under a chimney	Standard of Living Index	**
Food cooked on stove, chullah, open fire	Standard of Living Index	**
Age at first marriage	Daughters at home	**
Total number of sexual partners	Daughters at home	**
Cooking done under a chimney	Education in single years	**
Age	Sons at home	**
Food cooked on stove, chullah, open fire	House type	**
Cooking done under a chimney	De facto place of residence	**
Frequency of reading newspaper or magazine	Food cooked on stove, chullah, open fire	**
Cooking done under a chimney	House type	**
Food cooked on stove, chullah, open fire	Frequency of watching television	**
Frequency of reading newspaper or magazine	Cooking done under a chimney	**
House type	De facto place of residence	**
Wealth index factor score	De facto place of residence	**
Frequency of watching television	Standard of Living Index	**
Education in single years	Standard of Living Index	*

Table 5.9: The most non-independent variables for female dataset.
The independence level has been calculated from the methods as described in Section 2.2.2. The more * shown the stronger the association between the variables. The results have been ordered in terms of non-independence with the smallest values being the least independent.

5.10 Determining the most statistically likely directed acyclic graph

The calculations used to obtain the most statistically likely graph are shown below. The maximum likelylihood estimator (MLE), degrees of freedom (DF), penalty term (Pen.), bayes information criteria (BIC), akaike information criteria (AIC), and the euclidean distance (Eucl.) are shown as well.

Results from Female dataset

A 218 (living children), 202 (sons at home), 203 (daughters at home)	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(144,648)	95	241	144,889	144,743	
Opt A - All independent	(206,923)	11	28	206,951	206,934	0.089
Opt B - 218 to 202 and 203.	(161,375)	41	104	161,479	161,416	0.011
Opt C - 218 to 202 to 203.	(177,963)	35	89	178,052	177,998	0.027
Opt D - 202 and 203 to 218.	(148,517)	86	218	148,735	148,603	0.014
opt E - 218 to 203 to 202.	(182,599)	35	89	182,688	182,634	0.039

B 201 (# children born) with A	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(153,569)	383	972	154,541	153,952	
Opt A - 201 independent. 202 and 203 to 218	(215,731)	89	226	215,957	215,820	0.091
Opt B - 201 to 202 and203. 202 and 203 to 218	(174,571)	107	272	174,843	174,678	0.011
Opt C - 202 and 203 to 218 to 201	(157,528)	104	264	157,791	157,632	0.012
Opt D - 201 and 202 and 203 to 218	(204,182)	329	835	205,017	204,511	0.073

C 511 (age marriage), 836 (# sex partners) into B	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(141,318)	287	728	142,047	141,605	
Opt A - 511 to 836 to 201. 218 to 201.	(163,379)	70	178	163,557	163,449	0.051
Opt B - 511 to 836 and 201. 218 to 201.	(163,099)	88	223	163,323	163,187	0.051
Opt C - 836 to 511 to 201. 218 to 201.	(163,099)	88	223	163,323	163,187	0.051
Opt D - 836 to 511 and 201. 218 to 201.	(163,379)	70	178	163,557	163,449	0.051
Opt E - 511 to 836 to 218. 218 to 201.	(144,451)	38	96	144,547	144,489	0.002
Opt F - 511 to 836 and 218. 218 to 201	(142,233)	42	107	142,339	142,275	0.000
Opt G - 836 to 511 to 218. 218 to 201.	(142,233)	49	124	142,357	142,282	0.000
Opt H - 836 to 511 and 218. 218 to 201.	(144,451)	44	112	144,562	144,495	0.002

D 136 (# household members) into C	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(179,948)	143	363	180,311	180,091	
Opt A - 136 independent	(187,515)	28	71	187,586	187,543	0.007
Opt B - 136 and 218 to 201	(242,065)	118	299	242,364	242,183	0.02
Opt C* - 136 to 218 to 201	(180,130)	53	134	180,265	180,183	0.000
Opt D - 218 to 201 to 136	(183,818)	43	109	183,709	183,928	0.002
Opt E ** - 218 to 201 and 136	(180,130)	53	134	180,265	180,183	0.000

E 133 (education level) into D	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(219,349)	671	1,703	221,052	220,020	
Opt A - 133 independent of 511 to 218 to 201	(230,447)	65	165	230,612	230,512	0.007
Opt B - 133 to 511 to 218 to 201	(222,513)	65	165	222,678	222,578	0.001
Opt C - 511 and 133 to 218 to 201	(228,383)	167	424	228,807	228,550	0.006
Opt D - 511 to 218 to 201. 133 also to 201	(229,728)	155	393	230,121	229,883	0.007
Opt E - 511 to 218 to 201. 511 also to 133	(222,513)	65	165	222,678	222,578	0.001
Opt F - 511 to 218 to 201 and 133	(224,501)	77	195	224,696	224,578	0.001
Opt G - 511 to 218 to 201 to 133	(224,575)	65	165	224,740	224,640	0.001

F 026 (country/city), smnfhs2 (housetype), smsli (liv. Standard)	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(145,671)	47	119	145,790	145,718	
Opt A - Assuming all independent	(164,055)	8	20	164,075	164,063	0.029
Opt B - 026 to smnfhs2 and smsli	(152,774)	23	58	152,832	152,797	0.009
Opt C - smsli to smnfhs2 and 026	(149,274)	23	58	149,333	149,297	0.004
Opt D- smsli to 026 to smnfhs2	(152,774)	23	58	152,832	152,797	0.009
Opt E- 026 to smsli to smnfhs2	(149,274)	23	58	149,333	149,297	0.004
Opt F- smnfhs2 to smsli and 026	(146,220)	20	51	146,271	146,240	0.000
Opt G- 026 to smnfhs2 to smsli	(146,220)	20	51	146,271	146,240	0.000
Opt H- smnfhs2 and 026 to smsli	(152,839)	41	104	152,943	152,880	0.016
Opt I - smsli and 026 to smnfhs2	(149,784)	38	96	149,881	149,822	0.009
Opt J - smsli to smnfhs2 to 026	(146,220)	20	51	146,271	146,240	0.000
Opt K- smnfhs2 to smsli to 026	(149,274)	23	58	149,333	149,297	0.004
Opt L - smnfhs2 to 026 to smsli	(152774)	23	58	152,832	152,797	0.009
opt M - smnfhs2 and smsli to 026	)149274)	41	104	149,378	149,315	0.004

G 024 (state) into F	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(211,283)	191	485	211,768	211,474	
Opt 0 - state independent, smnfhs2 to 026 and ssli	(217,066)	23	58	217,124	217,089	0.003
Opt A -state to smnfhs2. smnfhs2 to 026 and ssli	(214,601)	29	74	214,675	214,630	0.001
Opt B - state to 026. smnfhs2 to 026 and ssli	(215,061)	50	127	215,188	215,111	0.002
Opt C - state to smsli. smnfhs2 to 026 and ssli	(216,558)	50	127	216,685	216,608	0.003
Opt D - state from smnfhs2. smnfhs2 to 026 and ssli	(214,601)	29	74	214,675	214,630	0.001
Opt E - state from 026. smnfhs2 to 026 and ssli	(214,836)	32	81	214,917	214,868	0.002
Opt F - state from smsli. smnfhs2 to 026 and ssli	(216,205)	32	81	216,287	216,237	0.002
Opt G - state to smnfhs2 and 026. smnfhs2 to smsli and 026	(212,597)	54	137	212,734	212,651	0.000
Opt H - smnfhs2, 026 to state. Smnfhs2 to smsli and 026.	(212,597)	56	142	212,739	212,653	0.000
Opt I - smnsfhs2 and smsli to state. Smnfhs2 to smsli	(214,094)	56	142	214,236	214,150	0.001
Opt J - state to smnfhs2 and 026. smnfhs2 to smsli	(219,539)	32	81	219,620	219,571	0.005
Opt K - smnsfhs2 and smsli to state excl smnfhs2 to smsli	(224,408)	32	81	224,489	224,440	0.006

H 159 (tv) into G	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(259,650)	767	1,946	261,596	260,417	
Opt 0 - tv independent.	(273,191)	59	150	273,341	273,250	0.004
Opt A - tv from state	(271,735)	68	173	271,908	271,803	0.003
Opt B - tv from 026	(268,417)	68	173	268,590	268,485	0.002
Opt C - tv from housetype	(267,255)	65	165	267,420	267,320	0.001
Opt D tv from ssli	(264,504)	68	173	264,677	264,572	0.001
Opt E tv to ssli	(269,442)	86	218	269,660	269,528	0.002
Opt F - tv to smnfhs2	(267,905)	83	211	268,115	267,988	0.001
Opt G - tv to 024	(271,735)	68	173	271,908	271,803	0.003
Opt H - tv to 026	(277,590)	167	424	278,014	277,757	0.005
Opt I smnfhs2 to 159 to smsli	(263,506)	92	233	263,740	263,598	0.000
Opt J smnfhs2 to 159 to smsli (excl smnfhs2 to smsli)	(269,235)	92	233	269,469	269,327	0.001

I133 (edu) into E	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(277,310)	1343	3,408	280,718	278,653	
Opt A - assume edu independent	(291,013)	62	157	291,171	291,075	0.004
Opt B - edu to country	(289,964)	116	294	290,258	290,080	0.004
Opt C - country to edu	(287,765)	80	203	287,968	287,845	0.003
Opt D - edu to housetype	(286,232)	81	206	286,437	286,313	0.002
Opt E - housetype to edu	(286,232)	74	188	286,420	286,306	0.002
Opt F - edu to ssli	(288,689)	278	705	289,395	288,967	0.003
Opt G - ssli to edu	(282,716)	74	188	282,904	282,790	0.001
Opt H - edu to tv	(286,891)	116	294	287,185	287,007	0.002
Opt I - tv to edu	(283,780)	80	203	283,983	283,860	0.001
Opt J - smn and 159 to edu	(282,109)	128	325	282,434	282,237	0.000
Opt K - smn and ssli to edu	(288,400)	128	325	288,724	288,528	0.002
Opt L - ssli and 159 to edu	(285,509)	152	386	285,895	285,661	0.001
Opt M - smn and 159 from 133	(282,109)	128	325	282,434	282,237	0.000
Opt N - smn and ssli from 133	(283,907)	290	736	284,643	284,197	0.001
Opt O - ssli and 159 from 133	(284,567)	332	842	285,409	284,899	0.002

J s49 (fire) into I	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(179,215)	191	485	179,700	179,406	
Opt A - assume 49 independent	(199,273)	23	58	199,332	199,296	0.022
Opt B - assume 49 from 026	(187,390)	32	81	187,472	187,422	0.005
Opt C - assume 49 from snfhs2	(190,717)	29	74	190,791	190,746	0.009
Opt D - assume 49 from ssli	(188,665)	32	81	188,746	188,697	0.009
Opt E - assume 49 from 026 and smnfhs2	(184,250)	56	142	184,392	184,306	0.002
Opt F - assume 49 from smnfhs2 and smsli	(185,832)	56	142	185,974	185,888	0.006
Opt G - assume 49 to 026	(192,805)	50	127	192,932	192,855	0.017
Opt H- assume 49 to smnfhs2	(190,717)	29	74	190,791	190,746	0.009
Opt I- assume 49 to smsli	(199,332)	50	127	199,459	199,382	0.024
Opt J- assume 49 to 026 and smnfhs2	(184,250)	56	142	184,392	184,306	0.002
Opt K- assume 49 to smnfhs2 and smsli	(190,776)	56	142	190,918	190,832	0.010

K s50 (cooking under chimney) into J	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(150,713)	143	363	151,076	150,856	
Opt A - assume s50 independent	(183,986)	49	124	184,111	184,035	0.042
Opt B - assume s50 from 026	(175,478)	53	134	175,613	175,531	0.013
Opt C - assume 50 from snfhs2	(176,531)	53	134	176,666	176,584	0.020
Opt D - assume 50 from s49	(151,349)	55	140	151,488	151,404	0.000
Opt E - assume50 to s49	(151,349)	55	140	151,488	151,404	0.000
Opt E - assume50 to 026	(183,826)	121	307	184,133	183,947	0.048
Opt F - assume50 to smnfhs2	(183,512)	65	165	183,677	183,577	0.039
Opt G - excl s49 to 026. s49 to s50 to 026	(153,668)	46	117	153,785	153,714	0.001
Opt H - excl s49 to 026. s50 to 026 and s49	(153,668)	46	117	153,785	153,714	0.001
Opt I - s49 and026 to s50	(151,285)	79	200	151,485	151,364	0.000
Opt J - s49 to s50 to 026	(151,188)	127	322	151,510	151,315	0.000
opt K - excl smnfhs2 to 026. smnfhs2 to s50 to 026	(178,220)	53	134	178,355	178,273	0.025
Opt L - excl smnfhs2 to 026. s50 to 026 and smnfhs2	(185,201)	65	165	185,366	185,266	0.044
Opt M - excl smnfhs2 to 026. smnfhs2 and 026 to s50	(174,136)	47	119	174,255	174,183	0.011
Opt N - s50 to s49 and 026	(151,188)	131	332	151,520	151,319	0.000

L v157 (freq. reading mag/paper) into K	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(188,746)	111	282	189,028	188,857	
Opt A - 157 independent. 159 to 133.	(208,395)	30	76	208,471	208,425	0.017
Opt B - 159 to 133 and 157	(202,102)	39	99	202,201	202,141	0.007
Opt C - 159 to 133 to 157	(189,676)	48	122	189,798	189,724	0.000
Opt D - 159 to 133 and 157. 133 to 157	(188,746)	111	282	189,028	188,857	0.000
Opt E - 157 and 159 to 133	(195,039)	102	259	195,298	195,141	0.008
Opt F - 157 to 159 to 133	(202,656)	39	99	202,755	202,695	0.067
Opt G - 157 to 159 and 133. 159 to 133	(189,301)	111	282	189,583	189,412	0.104

M sb69 (HIV weight) into L	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(359,565)	16127	40922	400,488	375,692	
Opt A - HIV independent	(418,045)	91	231	418,276	418,136	0.001
Opt B - HIV from state	(396,314)	100	254	396,567	396,414	0.001
Opt C - HIV from country/city	(401,105)	100	254	401,359	401,205	0.001
Opt D - HIV from housetype	(405,508)	97	246	405,754	405,605	0.001
Opt E - HIV from living standard	(405,036)	100	254	405,289	405,136	0.001
Opt F - HIV from sex partners	(405,739)	97	246	405,985	405,836	0.001
Opt G - HIV from edu level	(405,140)	109	277	405,416	405,249	0.001
Opt H - HIV from state and country/city	(389,077)	136	345	389,422	389,213	0.001
Opt I - HIV from country/city and housteype	(400,632)	124	315	400,947	400,756	0.001
Opt J - HIV from state and housteype	(400,631)	136	345	400,976	400,767	0.001
Opt K - HIV to country/city	(418,052)	127	322	418,374	418,179	0.001
Opt I - HIV to state	(418,042)	100	254	418,295	418,142	0.001

152 (Hhhead age), into 218, 201, 136	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(251,793)	719	1,748	253,540	252,512	
Opt A - age independent	(270,469)	42	102	270,571	270,511	0.006
Opt B - age to 218 to 201 and 136	(268,337)	57	139	268,476	268,394	0.005
Opt C - age and 218 to 201. 218 also to 136	(326,684)	129	314	326,998	326,813	0.009
Opt D - age and 218 to 136. 218 also to 201	(264,103)	177	430	264,534	264,280	0.003
Opt E - 218 to 152 and 201 and 136	(268,337)	62	151	268,488	268,399	0.005
Opt F -218 to 201 and 136. 201 to 152	(268,481)	60	146	268,627	268,541	0.006
Opt G - 218 to 136 and 201. 136 to 152	(267,317)	77	187	267,504	267,394	0.005
Opt H - 152 to 218 and 201. 218 to 136	(268,233)	77	187	268,421	268,310	0.005
Opt I - 152 to 218,201, 136. 218 to 201, 136	(251,972)	269	654	252,626	252,241	0.000
Opt J - 218 and 152 to 201 and 136	(254,104)	249	605	254,709	254,353	0.000
Opt K - 218 to 201 and 136. 218 and 201 to 152	(268,233)	134	326	268,559	268,367	0.005
Opt L - 218 to 201 and 136. 218 and 136 to 152	(262,667)	197	479	263,146	262,864	0.004
Opt m - 218 to 201 and 136. 201 and 136 to 152	(263,561)	134	326	263,887	263,695	0.004
Opt n - 152 to 218,201, 136	(318,186)	69	168	318,354	318,255	0.008

012 (age) into 152, 218, 201	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(216,136)	479	1,215	217,352	216,615	
Opt A - age independent	(232,024)	122	310	232,334	232,146	0.0230
Opt B - 012 to 152 to 218 and 201. 218 to 201	(225,333)	134	340	225,673	225,467	0.020
Opt C - 012 to 218 to 201. 152 to 218 and 201	(223,161)	197	500	223,661	223,358	0.003
Opt D - 012 to 201. 152 to 218 and 201. 218 to 201.	(231,691)	392	995	232,686	232,083	0.022
Opt E - 152 to 012 and 218 and 201. 218 to 201	(225,333)	134	340	225,673	225,467	0.020
Opt F - 152 to 218 and 201. 218 to 201 and 012	(216,913)	137	348	217,260	217,050	0.012
Opt G - 152 to 218 and 201. 218 to 201. 201 to 012	(217,185)	131	332	217,518	217,316	0.012
Opt H - 152 to 218 to 012 to 201. 152 to 201. excl 218 to 201	(257,528)	107	272	257,799	257,635	0.012
Opt I - 152 to 218 to 012 to 201. 152 to 201. 218 to 201.	(216,579)	407	1,033	217,612	216,986	0.012
Opt j - 152 to 218 and 201. 012 to 218 and 201	(233,090)	167	424	233,514	233,257	0.037
Opt K - 152 to 218 to 201. 152 to 201. 012 to 218 and 201	(222,828)	467	1,185	224,013	223,295	0.003
Opt L - 152 to 218 and 201. 218 and 201 to 012	(291,619)	116	294	291,914	291,735	0.018
Opt M - 152 to 218 and 201. 218 to 201 nd 012. 201 to 012.	(216,609)	191	485	217,093	216,800	0.012
Opt N - 218 to 012. 012 to 201 and 152	(253,689)	51	129	253,818	253,740	0.017
Opt O - 152 and 218 to 012. 012 to 201	(252,987)	93	236	253,223	253,080	0.014
Opt P - 152 to 218 and 201 and 012. 218 to 012 and 201	(250,134)	179	454	250,588	250,313	0.013
Opt Q - 152 to 218 and 012 and 201	(300,343)	59	150	300,493	300,402	0.028
Opt R - 152 to 218 and 201 and 012. 218 to 201	(225,333)	134	340	225,673	225,467	0.020
Opt S - 012 to 152 to 218. 152 to 201	(300,343)	59	150	300,493	300,402	0.028
Opt T - 152 to 218, 012, 201. 218 to 201 and 012. 201 to 012	(209,519)	479	1,215	210,734	209,998	0.013
Opt U - 152 to 218, 012, 201. 218 to 201. 012 to 201 and 218.	(216,469)	179	454	216,924	216,648	0.000
Opt V - 152 to 218 and 201. 012 to 218 and 152 and 201	(257,084)	179	454	257,538	257,263	0.006

191 (wealth) into smnfhs2, smsli, 133	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(231,770)	503	1,276	233,046	232,273	
Opt A - assume 191 independent	(269,566)	40	102	269,668	269,606	0.012
Opt B - 191 to smnfhs2 to smsli to 133	(252,500)	50	127	252,627	252,550	0.005
Opt C - smnfhs2 and 191 to smsli to 133	(254,343)	85	216	254,559	254,428	0.009
Opt D - smnfhs2 to smsli to 133. 191 to 133.	(265,678)	160	406	266,084	265,838	0.011
Opt E - smnfhs2 to 191. smnfhs2 to smsli to 133	(252,500)	50	127	252,627	252,550	0.005
Opt F - smnfhs2 and 191 to smsli to 133	(244,202)	55	140	244,342	244,257	0.003
Opt G - smnfhs2 to smsli to 133 to 191	(257,600)	55	140	257,739	257,655	0.006
Opt H - smnfhs2 and smsli to 191. smsli to 133	(247,257)	89	226	247,483	247,346	0.007
Opt I - smnfhs2 and smsli to 191. smnfhs2 to smsli to 133	(236,590)	95	241	236,831	236,685	0.002
Opt J 191 to smnfhs2 and smsli. Smsli to 133	(237,116)	59	150	237,266	237,175	0.002
Opt K - 191 to smsnfhs2 and smsli. Smnfhs2 to smsli to 133	(236,590)	119	302	236,892	236,709	0.002
Opt L - smnfhs2 to 191 and smsli. 191 to smsli to 133	(236,590)	95	241	236,831	236,685	0.002

438 (height), 437(weight), and 191	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(158,464)	95	241	158,705	158,559	
Opt A - all independent	(164,436)	11	28	164,464	164,447	0.010
Opt B - 191 to 437 and 438	(159,262)	41	104	159,366	159,303	0.000
Opt C - 437 to 191 and 438	(158,590)	35	89	158,678	158,625	0.000
Opt D - 438 to 191 and 437	(163,059)	35	89	163,147	163,094	0.007
Opt E - 191 to 437 to 438	(158,700)	35	89	158,788	158,735	0.000
Opt F - 191 to 438 to 437	(163,169)	35	89	163,257	163,204	0.008
Opt G - 438 to 437 to 191	(158,590)	35	89	158,678	158,625	0.000
Opt H - 438 to 191 to 437	(159,152)	41	104	159,256	159,193	0.000
Opt I - 437 to 438 to 191	(163,059)	35	89	163,147	163,094	0.007
Opt J - 437 to 191 to 438	(159,152)	41	104	159,256	159,193	0.000
Opt K - 191 to 437 and 438. 437 to 438	(158,574)	95	241	158,815	158,669	0.000
Opt L - 191 to 437 and 438. 438 to 437	(158,574)	95	241	158,815	158,669	0.000
Opt M - 437 to 191 and 438. 191 to 438	(158,464)	95	241	158,705	158,559	0.000
Opt N - 437 to 191 and 438. 438 to 191	(158,464)	95	241	158,705	158,559	0.000
Opt O - 438 to 191 and 437. 437 to 191	(158,464)	95	241	158,705	158,559	0.000
Opt P - 438 to 191 and 437. 191 to 437	(158,464)	95	241	158,705	158,559	0.000

456 (Haemoglobin) into 024, sb69	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(176,310)	79	(192)	176,118	176,389	
Opt A - 456 indep of 024 to sb69	(177,128)	19	(46)	177,081	177,147	0.001
Opt B - 024 to 456 and sb69	(176,775)	31	(75)	176,700	176,806	0.000
Opt C - 456 to 024 to sb69	(176,775)	31	(75)	176,700	176,806	0.000
Opt D - 024 and 456 to sb69	(176,663)	67	(163)	176,500	176,730	0.000
Opt E 024 to sb69 to 456	(176,898)	31	(75)	176,823	176,929	0.000
Opt F - 456 to 024 and sb69	(186,086)	34	(83)	186,003	186,120	0.010
Opt G - 024 to 456 and sb69. 456 to sb69	(176,898)	79	(192)	176,706	176,977	0.000
Opt H - 024 to 456 and sb69. sb69 to 456	(176,898)	79	(192)	176,706	176,977	0.000
Opt I - 456 to 024 and sb69. 024 to sb69	(176,310)	79	(192)	176,118	176,389	0.000

133, 476, 475, 157	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(203,135)	251	637	203,772	203,386	
Opt A - 475 476 independent. 133 to 157	(217,036)	31	79	217,115	217,067	0.008
Opt B - 133 to 476, 157, 475	(212,275)	55	140	212,414	212,330	0.008
Opt C - 133 to 157 to 476 and 475	(213,125)	34	86	213,211	213,159	0.008
Opt D - 133 to 476, 157, 475. 157 to 476, 475	(211,889)	139	353	212,242	212,028	0.008
Opt E - 133 to 476, 475, 157. 475 to 476	(203,433)	83	211	203,643	203,516	0.000
Opt F - 133 to 157. 157 to 475, 576. 475 to 476	(204,012)	59	150	204,162	204,071	0.000
Opt G - 133 to 476, 157, 475. 157 to 476, 475. 475 to 476	(203,135)	242	614	203,749	203,377	0.000
Opt H - 133 to 476, 157, 475. 476 to 465	(203,433)	83	211	203,643	203,516	0.000
Opt I - 133 to 157. 157 to 476, 475. 476 to 475	(204,012)	59	150	204,162	204,071	0.000
Opt J - 133 to 476, 157, 475. 157 to 476, 475. 476 to 475	(203,135)	242	614	203,749	203,377	0.000

Wealth, TV, living standard, and education	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(279,076)	2015	5,113	284,190	281,091	
Opt A - snfhs2 to ssli, 191, 133. 191 to ssli. 159 to ssli, 133.	(290,385)	314	797	291,182	290,699	0.002
Opt B - sfh2, 159, sli to 133. 159, sfs2, 191 to sli. sfs2 to 191	(288,061)	530	1,345	289,406	288,591	0.002
Opt C - snfhs2 to ssli, 191, 133. 191 to ssli. 159 to ssli to 133	(290,305)	296	751	291,056	290,601	0.002
Opt D - snfhs2 to ssli, 191, 133. 191 to ssli. 159, ssli to 133	(288,318)	368	934	289,252	288,686	0.002

Results from Male dataset

A 218 (living children), 202 (sons at home), 203 (daughters at home)	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(77,766)	95	241	78,007	77,861	
Opt A - all independent	(115,092)	11	28	115,120	115,103	0.113
Opt B - 218 parent of 202 and 203.	(86,251)	41	104	86,356	86,292	0.007
Opt C - 218 parent of 202. Assume 202 parent of 203.	(95,450)	35	89	95,539	95,485	0.020
Opt D - 202. and 203 parent of 218.	(81,727)	86	218	81,945	81,813	0.021

B 201 (# children born) with **A**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(82,431)	383	972	83,403	82,814	
Opt A - 201 independent. 202 and 203 to 218	(120,861)	89	226	121,086	120,950	0.119
Opt B - 201 to 202 and203. 202 and 203 to 218	(94,177)	107	272	94,448	94,284	0.008
Opt C - 202 and 203 to 218 to 201	(86,420)	104	264	86,684	86,524	0.019
Opt D - 201 and 202 and 203 to 218	(128,061)	329	835	128,896	128,390	0.109

C 511 (age marriage), 836 (# sex partners) into **B**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(93,980)	287	728	94,709	94,267	
Opt A - 511 to 836 to 201. 218 to 201.	(133,534)	70	178	133,712	133,604	0.076
Opt B - 511 to 836 and 201. 218 to 201.	(129,905)	88	223	130,128	129,993	0.075
Opt C - 836 to 511 to 201. 218 to 201.	(129,905)	88	223	130,128	129,993	0.075
Opt D - 836 to 511 and 201. 218 to 201.	(133,534)	70	178	133,712	133,604	0.076
Opt E - 511 to 836 to 218. 218 to 201.	(98,214)	38	96	98,310	98,252	0.002
Opt F - 511 to 836 and 218. 218 to 201.	(94,656)	42	107	94,763	94,698	0.000
Opt G - 836 to 511 to 218. 218 to 201.	(94,656)	49	124	94,781	94,705	0.000
Opt H - 836 to 511 and 218. 218 to 201.	(98,214)	44	112	98,325	98,258	0.002

D 136 (# household members) into **C**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(105,164)	143	363	105,526	105,307	
Opt A - 136 independent. 218 to 201	(109,496)	28	71	109,567	109,524	0.006
Opt B - 136 and 218 to 201	(141,720)	118	299	142,020	141,838	0.029
Opt C - 136 to 218 to 201	(105,252)	53	134	105,386	105,305	0.000
Opt D - 218 to 201 to 136	(107,301)	43	109	107,410	107,344	0.002
Opt E- 218 to 201 and 136	(105,252)	53	134	105,386	105,305	0.000

E 133 (education level) into **D**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(136,979)	671	1,703	138,682	137,650	
Opt A - 133 independent of 511 to 218 to 201	(140,397)	47	119	140,517	140,444	0.003
Opt B - 133 to 511 to 218 to 201	(138,023)	65	165	138,188	138,088	0.000
Opt C - 511 and 133 to 218 to 201	(139,933)	167	424	140,357	140,100	0.002
Opt D - 511 to 218 to 201. 133 also to 201	(140,136)	155	393	140,529	140,291	0.002
Opt E - 511 to 218 to 201. 511 also to 133	(138,023)	65	165	138,188	138,088	0.000
Opt F - 511 to 218 to 201 and 133	(138,734)	77	195	138,930	138,811	0.000
Opt G - 511 to 218 to 201 to 133	(138,693)	65	165	138,858	138,758	0.000

F 026 (country/city), smnfhs2 (housetype), smsli (liv. Standard)	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(89,501)	47	119	89,621	89,548	
Opt A - Assuming all independent	(100,311)	8	20	100,331	100,319	0.030
Opt B - 026 to smnfhs2 and smsli	(93,154)	23	58	93,212	93,177	0.006
Opt C - smsli to smnfhs2 and 026	(92,218)	23	58	92,277	92,241	0.006
Opt D- smsli to 026 to smnfhs2	(93,154)	23	58	93,212	93,177	0.006
Opt E- 026 to smsli to smnfhs2	(92,218)	23	58	92,277	92,241	0.006
Opt F- smnfhs2 to smsli and 026	(89,867)	20	51	89,918	89,887	0.000
Opt G- 026 to smnfhs2 to smsli	(89,867)	20	51	89,918	89,887	0.000
Opt H- smnfhs2 and 026 to smsli	(94,256)	41	104	94,360	94,297	0.017
Opt I - smsli and 026 to smnfhs2	(91,904)	38	96	92,001	91,942	0.009
Opt J - smsli to smnfhs2 to 026	(89,867)	20	51	89,918	89,887	0.000
Opt K- smnfhs2 to smsli to 026	(92,692)	23	58	92,751	92,715	0.010
Opt L - smnfhs2 to 026 to smsli	(93,154)	23	58	93,212	93,177	0.006
Opt M - smnfhs2 and smsli to 026	(92,692)	41	104	92,796	92,733	0.010

G 024 (state) into **F**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(126,675)	191	485	127,160	126,866	
Opt 0 - state independent, smnfhs2 to 026 and ssli	(130,901)	23	58	130,959	130,924	0.005
Opt A -state to smnfhs2. smnfhs2 to 026 and ssli	(129,252)	29	74	129,326	129,281	0.002
Opt B - state to 026. smnfhs2 to 026 and ssli	(129,165)	50	127	129,292	129,215	0.002
Opt C - state to smsli. smnfhs2 to 026 and ssli	(130,568)	50	127	130,694	130,618	0.005
Opt D - state from smnfhs2. smnfhs2 to 026 and ssli	(129,252)	29	74	129,326	129,281	0.002
Opt E - state from 026. smnfhs2 to 026 and ssli	(128,807)	32	81	128,888	128,839	0.001
Opt F - state from smsli. smnfhs2 to 026 and ssli	(130,632)	32	81	130,713	130,664	0.004
Opt G - state to smnfhs2 and 026. smnfhs2 to smsli and 026	(127,517)	54	137	127,654	127,571	0.000
Opt H - smnfhs2 and 026 to state. Smnfhs2 to smsli and 026.	(127,517)	56	142	127,659	127,573	0.000
Opt J - smnsfhs2 and smsli to state. Smnfhs2 to smsli	(128,919)	56	142	129,061	128,975	0.002
Opt K - state to smnfhs2 and 026. smnfhs2 to smsli	(131,913)	32	81	131,994	131,945	0.006
Opt L - smnsfhs2 and smsli to state excl smnfhs2 to smsli	(134,672)	32	81	134,754	134,704	0.006

H 159 (tv) into **G**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(157,345)	767	1,946	159,291	158,112	
Opt 0 - tv independent.	(164,417)	59	150	164,567	164,476	0.004
Opt A - tv from state	(163,366)	68	173	163,539	163,434	0.002
Opt B - tv from 026	(162,051)	68	173	162,224	162,119	0.001
Opt C - tv from housetype	(161,490)	65	165	161,655	161,555	0.001
Opt D tv from ssli	(160,375)	68	173	160,547	160,443	0.001
Opt E tv to ssli	(162,681)	86	218	162,899	162,767	0.002
Opt F - tv to smnfhs2	(161,912)	83	211	162,122	161,995	0.007
Opt G - tv to 024	(163,366)	68	173	163,539	163,434	0.002
Opt H - tv to 026	(168,541)	167	424	168,965	168,708	0.006
Opt I smnfhs2 to 159 to smsli	(159,754)	92	233	159,987	159,846	0.000
Opt J smnfhs2 to 159 to smsli (excl smnfhs2 to smsli)	(163,137)	92	233	163,370	163,229	0.001

I133 (edu) into **E**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(176,305)	1343	3,408	179,713	177,648	
Opt A - assume edu independent	(183,092)	62	157	183,249	183,154	0.004
Opt B - edu to country	(182,756)	116	294	183,051	182,872	0.004
Opt C - country to edu	(181,918)	80	203	182,121	181,998	0.004
Opt D - edu to housetype	(180,974)	81	206	181,180	181,055	0.003
Opt E - housetype to edu	(180,974)	74	188	181,162	181,048	0.003
Opt F - edu to ssli	(181,286)	278	705	181,991	181,564	0.004
Opt G - ssli to edu	(178,645)	74	188	178,832	178,719	0.003
Opt H - edu to tv	(181,552)	116	294	181,847	181,668	0.004
Opt I - tv to edu	(180,291)	80	203	180,494	180,371	0.003
Opt J - smn and 159 to edu	(179,435)	128	325	179,759	179,563	0.003
Opt K - smn and ssli to edu	(178,405)	128	325	178,730	178,533	0.003
Opt L - ssli and 159 to edu	(177,826)	152	386	178,212	177,978	0.003
Opt M - smn and 159 from 133	(179,435)	128	325	179,759	179,563	0.003
Opt N - smn and ssli from 133	(179,168)	290	736	179,904	179,458	0.003
Opt O - ssli and 159 from 133	(179,747)	332	842	180,589	180,079	0.003

L v157 (freq. reading mag/paper) into **I**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(124,325)	111	282	124,606	124,436	
157 independent. 159 to 133.	(136,552)	30	76	136,628	136,582	0.011
159 to 133 and 157	(133,147)	39	99	133,246	133,186	0.006
159 to 133 to 157	(125,399)	48	122	125,521	125,447	0.000
159 to 133 and 157. 133 to 157	(124,325)	111	282	124,606	124,436	0.000
157 and 159 to 133	(127,730)	102	259	127,989	127,832	0.004
157 to 159 to 133	(137,711)	39	99	137,810	137,750	0.012
157 to 159 and 133. 159 to 133	(128,889)	111	282	129,170	129,000	0.009

M sb69 (HIV weight) into **L**	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(235,434)	16127	40922	276,357	251,561	
Opt A - HIV independent	(256,778)	91	231	257,009	256,869	0.000
Opt B - HIV from state	(241,289)	100	254	241,542	241,389	0.001
Opt C - HIV from country/city	(306,650)	100	254	306,903	306,750	0.001
Opt D - HIV from housetype	(310,701)	97	246	310,948	310,798	0.001
Opt E - HIV from living standard	(310,584)	100	254	310,837	310,684	0.001
Opt F - HIV from sex partners	(318,173)	97	246	318,419	318,270	0.001
Opt G - HIV from edu level	(313,367)	109	277	313,643	313,476	0.001
Opt H - HIV from state and country/city	(293,584)	136	345	293,929	293,720	0.001
Opt I - HIV from country/city and housteype	(306,453)	124	315	306,767	306,577	0.001
Opt J - HIV from state and housteype	(306,674)	136	345	307,019	306,810	0.001
Opt K - HIV to country/city	(245,733)	127	322	246,055	245,860	0.001
Opt I - HIV to state	(241,281)	100	254	241,535	241,381	0.001

152 (Hhhead age), into 218, 201, 136	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(359,565)	16127	39197	398,762	375,692	
Opt A - HIV independent	(418,043)	91	221	418,264	418,134	0.001
Opt B - HIV from state	(396,315)	100	243	396,558	396,415	0.001
Opt C - HIV from country/city	(401,104)	100	243	401,347	401,204	0.001
Opt D - HIV from housetype	(405,507)	97	236	405,743	405,604	0.001
Opt E - HIV from living standard	(405,035)	100	243	405,278	405,135	0.001
Opt F - HIV from sex partners	(405,737)	97	236	405,973	405,834	0.001
Opt G - HIV from edu level	(405,135)	109	265	405,400	405,244	0.001
Opt H - HIV from state and country/city	(389,078)	136	331	389,408	389,214	0.001
Opt I - HIV from country/city and housteype	(400,631)	124	301	400,933	400,755	0.001
Opt J - HIV from state and housteype	(400,630)	136	331	400,960	400,766	0.001
Opt K - HIV to country/city	(418,052)	127	309	418,360	418,179	0.001
Opt I - HIV to state	(418,042)	100	243	418,285	418,142	0.001

012 (age) into 152, 218, 201	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(120,632)	479	1,164	121,797	121,111	
Opt A - age independent	(141,841)	122	297	142,138	141,963	0.038
Opt B - 012 to 152 to 218 and 201. 218 to 201.	(131,608)	134	326	131,934	131,742	0.033
Opt C - 012 to 218 to 201. 152 to 218 and 201	(131,039)	197	479	131,518	131,236	0.008
Opt D - 012 to 201. 152 to 218 and 201. 218 to 201.	(141,668)	392	953	142,621	142,060	0.034
Opt E - 152 to 012 and 218 and 201. 218 to 201	(131,608)	134	326	131,934	131,742	0.033
Opt F - 152 to 218 and 201. 218 to 201 and 012	(129,898)	137	333	130,231	130,035	0.027
Opt G - 152 to 218 and 201. 218 to 201. 201 to 012	(129,951)	131	318	130,269	130,082	0.026
Opt H - 152 to 218 to 012 to 201. 152 to 201. excl 218 to 201	(151,439)	107	260	151,699	151,546	0.023
Opt I - 152 to 218 to 012 to 201. 152 to 201. 218 to 201.	(129,724)	407	989	130,714	130,131	0.027
Opt j - 152 to 218 and 201. 012 to 218 and 201	(147,619)	167	406	148,025	147,786	0.013
Opt K - 152 to 218 to 201. 152 to 201. 012 to 218 and 201	(130,866)	467	1,135	132,001	131,333	0.008
Opt L - 152 to 218 and 201. 218 and 201 to 012	(162,271)	116	282	162,553	162,387	0.024
Opt M - 152 to 218 and 201. 218 to 201 and 012. 201 to 012.	(129,715)	191	464	130,180	129,906	0.027
Opt N - 218 to 012. 012 to 201 and 152	(144,544)	51	124	144,668	144,595	0.007
Opt O - 152 and 218 to 012. 012 to 201	(143,288)	93	226	143,514	143,381	0.026
Opt P - 152 to 218 and 201 and 012. 218 to 012 and 201	(142,201)	179	435	142,636	142,380	0.026
Opt Q - 152 to 218 and 012 and 201	(164,164)	59	143	164,307	164,223	0.035
Opt R - 152 to 218 and 201 and 012. 218 to 201	(131,608)	134	326	131,934	131,742	0.033
Opt S - 012 to 152 to 218. 152 to 201	(164,164)	59	143	164,307	164,223	0.035
Opt T - 152 to 218, 012, 201. 218 to 201 and 012. 201 to 012.	(120,660)	479	1,164	121,824	121,139	0.036
Opt U - 152 to 218, 012, 201. 218 to 201. 012 to 201, 218.	(120,806)	179	435	121,241	120,985	0.000
Opt V - 152 to 218 and 201. 012 to 218 and 152 and 201	(142,347)	179	435	142,782	142,526	0.006

191 (wealth) into smnfhs2, smsli, 133	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(147,508)	503	1,223	148,731	148,011	
Opt A - assume 191 independent	(169,314)	40	97	169,412	169,354	0.009
Opt B - 191 to smnfhs2 to smsli to 133	(159,311)	50	122	159,432	159,361	0.004
Opt C - smnfhs2 and 191 to smsli to 133	(159,869)	85	207	160,076	159,954	0.006
Opt D - smnfhs2 to smsli to 133. 191 to 133.	(167,605)	160	389	167,994	167,765	0.008
Opt E - smnfhs2 to 191. smnfhs2 to smsli to 133	(159,311)	50	122	159,432	159,361	0.004
Opt F - smnfhs2 and 191 to smsli to 133	(154,520)	55	134	154,654	154,575	0.002
Opt G - smnfhs2 to smsli to 133 to 191	(163,487)	55	134	163,621	163,542	0.004
Opt H - smnfhs2 and smsli to 191. smsli to 133	(155,316)	89	216	155,532	155,405	0.004
Opt I - smnfhs2 and smsli to 191. smnfhs2 to smsli to 133	(149,626)	95	231	149,857	149,721	0.001
Opt J 191 to smnfhs2 and smsli. Smsli to 133	(149,967)	59	143	150,110	150,026	0.001
Opt K - 191 to smsnfhs2 and smsli. Smnfhs2 to smsli to 133	(149,626)	119	289	149,915	149,745	0.001
Opt L - smnfhs2 to 191 and smsli. 191 to smsli to 133.	(149,626)	95	231	149,857	149,721	0.001

438 (height), 437(weight and 191	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(135,735)	95	231	135,966	135,830	
Opt A - all independent	(142,581)	11	27	142,608	142,592	0.005
Opt B - 191 to 437 and 438	(138,256)	41	100	138,356	138,297	0.001
Opt C - 437 to 191 and 438	(135,991)	35	85	136,076	136,026	0.000
Opt D - 438 to 191 and 437	(138,181)	35	85	138,266	138,216	0.001
Opt E - 191 to 437 to 438	(135,991)	35	85	136,076	136,026	0.000
Opt F - 191 to 438 to 437	(138,181)	35	85	138,266	138,216	0.001
Opt G - 438 to 437 to 191	(135,991)	35	85	136,076	136,026	0.000
Opt H - 438 to 191 to 437	(138,256)	41	100	138,356	138,297	0.001
Opt I - 437 to 438 to 191	(138,181)	35	85	138,266	138,216	0.001
Opt J - 437 to 191 to 438	(138,256)	41	100	138,356	138,297	0.001
Opt K - 191 to 437 and 438. 437 to 438	(135,735)	95	231	135,966	135,830	0.000
Opt L - 191 to 437 and 438. 438 to 437	(135,735)	95	231	135,966	135,830	0.000
Opt M - 437 to 191 and 438. 191 to 438	(135,735)	95	231	135,966	135,830	0.000
Opt N - 437 to 191 and 438. 438 to 191	(135,735)	95	231	135,966	135,830	0.000
Opt O - 438 to 191 and 437. 437 to 191	(135,735)	95	231	135,966	135,830	0.000
Opt P - 438 to 191 and 437. 191 to 437	(135,735)	95	231	135,966	135,830	0.000

456 (Haemoglobin) itno 024, sb69	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(121,471)	79	192	121,663	121,550	
Opt A - 456 indep of 024 to sb69	(124,226)	19	46	124,272	124,245	0.001
Opt B - 024 to 456 and sb69	(122,583)	31	75	122,658	122,614	0.000
Opt C - 456 to 024 to sb69	(122,583)	31	75	122,658	122,614	0.000
Opt D - 024 and 456 to sb69	(123,114)	67	163	123,276	123,181	0.000
Opt E - 024 to sb69 to 456	(123,202)	31	75	123,278	123,233	0.000
Opt F - 456 to 024 and sb69	(128,960)	34	83	129,042	128,994	0.010
Opt G - 024 to 456 and sb69. 456 to sb69	(123,202)	79	192	123,394	123,281	0.000
Opt H - 024 to 456 and sb69. sb69 to 456	(123,202)	79	192	123,394	123,281	0.000
Opt I - 456 to 024 and sb69. 024 to sb69	(121,471)	79	192	121,663	121,550	0.000

133, 476, 475, 157	MLE	DF	Pen.	BIC	AIC	Eucl.
Reality	(127,336)	251	610	127,946	127,58	
Opt A - 475 476 independent. 133 to 157	(134,168)	31	75	134,243	134,199	0.003
Opt B - 133 to 476, 157, 475	(132,051)	55	134	132,184	132,106	0.002
Opt C - 133 to 157 to 476 and 475	(132,252)	34	83	132,335	132,286	0.002
Opt D - 133 to 476, 157, 475. 157 to 476, 475	(172,526)	139	338	172,864	172,665	0.042
Opt E - 133 to 476, 475, 157. 475 to 476	(127,564)	83	202	127,766	127,647	0.000
Opt F - 133 to 157. 157 to 475, 576. 475 to 476	(127,736)	59	143	127,879	127,795	0.000
Opt G - 133 to 476, 157, 475. 157 to 476, 475. 475 to 476	(127,336)	242	588	127,924	127,578	0.000
Opt H - 133 to 476, 157, 475. 476 to 465	(127,564)	83	202	127,766	127,647	0.000
Opt I - 133 to 157. 157 to 476, 475. 476 to 475	(127,736)	59	143	127,879	127,795	0.000
Opt J - 133 to 476, 157, 475. 157 to 476, 475. 476 to 475	(127,336)	242	588	127,924	127,578	0.000
Opt K - 475 476 133 157independent.	(145,320)	13	32	145,352	145,333	0.013
Opt L - 133 to 476, 475	(143,203)	31	75	143,279	143,234	0.012
Opt M - 133 independent. 157 to 476 and 475	(143,405)	25	61	143,465	143,430	0.013
Opt N - 133 to 476, 475. 157 to 476, 475	(183,679)	121	294	183,973	183,800	0.039
Opt O - 133 to 476, 475. 475 to 476	(138,717)	65	158	138,875	138,782	0.011
Opt P - 133 independent. 157 to 475, 576. 475 to 476	(138,889)	41	100	138,988	138,930	0.012
Opt Q - 133 to 476, 475. 157 to 476, 475. 475 to 476	(138,489)	233	566	139,055	138,722	0.012
Opt R - 133 to 476, 475. 476 to 465	(138,717)	65	158	138,875	138,782	0.011
Opt S - 133 independent. 157 to 476, 475. 476 to 475	(138,889)	41	100	138,988	138,930	0.012
Opt T - 133 to 476, 475. 157 to 476, 475. 476 to 475	(138,489)	233	566	139,055	138,722	0.012

5.11 Sample of coding used

Data was taken off the SAS datasets using the following SAS code:

```
#female dataset code
PROC SQL;
   CREATE TABLE SASUSER.QUERY_FOR_IAIR52FL_SD2_0002 AS
   SELECT t1.CASEID,
          t1.S31A,
          t1.S30,
          t1.V475,
          t1.V024,
          t1.V026,
          t1.V152,
          t1.SANGAYR,
          t1.V012,
          t1.V201,
          t1.V202,
          t1.V203,
          t1.V218,
          t1.V136,
          t1.V511,
          t1.V836,
          t1.V437,
          t1.V438,
          t1.V445,
          t1.V456,
          t1.SA69,
          t1.V191,
          t1.SSLI,
          t1.SNFHS2,
          t1.V133,
          t1.V159,
          t1.V157,
          t1.S49,
          t1.S50,
          t1.V702
      FROM EC100014.IAIR52FL AS t1
      WHERE t1.S31A NOT = 7
      ORDER BY t1.S31A DESCENDING;
QUIT;

#male dataset code
PROC SQL;
   CREATE TABLE SASUSER.QUERY_FOR_IAMR52FL_SD2_0001 AS
   SELECT t1.MCASEID,
          t1.SM31A,
          t1.SM30,
          t1.MV476,
          t1.MV024,
          t1.SM025,
          t1.MV152,
          t1.SMANGAYR,
          t1.MV012,
          t1.MV202,
          t1.MV203,
          t1.MV201,
          t1.MV218,
          t1.MV136,
          t1.SM410C,
          t1.MV836,
          t1.MV437,
          t1.MV438,
          t1.MV445,
          t1.MV456,
          t1.SB69,
          t1.MV191,
          t1.SMSLI,
          t1.SMNFHS2,
          t1.MV133,
          t1.MV159,
```

```
          t1.MV157
      FROM EC100009.IAMR52FL AS t1
      WHERE t1.SM31A NOT = 7
      ORDER BY t1.SM31A DESCENDING;
QUIT;
```

Kolmogorov-Smirnov test coding:

```
%running Kirminov Smirnoff test
%runs thru all variables
%selects TB and non TB data
%copies if H0, H1 chosen, maximal distance, p-value to matrix
%prints matrix


matIRks=[1,1,1,1]
for i=1:1:29
TBi=sort(data(1:472,i));
nTBi=sort(data(473:118857,i));
[h,p,ks2stat] = kstest2(TBi,nTBi,0.05,'unequal');
matIRks=[matIRks; i,h,p,ks2stat];
end
matIRks


matIRks=[1,1,1,1]
for i=1:1:29
TBi=sort(data(1:472,i));
nTBi=sort(data(473:118857,i));
[h,p,ks2stat] = kstest2(TBi,nTBi,0.01,'unequal');
matIRks=[matIRks; i,h,p,ks2stat];
end
matIRks
```

Kolmogorov-Smirnov permuted test coding:

```



matIR_perm_ks=[1,1,1,1,1]
ApproxPVal=[99,99]
for i=1:1:29
TBi=sort(data(1:472,i))';
nTBi=sort(data(473:118857,i))';
[h, p,dist]= kstest2(TBi,nTBi,0.05,'unequal');% observed test statistic
ObsKSvalue=dist;

n_TBi=length(TBi); % sample size of the TB data
n_nTBi=length(nTBi); % sample size of the non TB data
n_tot=n_TBi+n_nTBi; % sample size of the pooled data
tot=[TBi nTBi]; % observed data -- ordered: TB data followed by non TB data

B=100000; % number of bootstrap replicates
BResult=zeros(1,B); % vector of zeros for the bootstrapped test statistics
ApproxPValue=0; % initialise an accumulator for approximate p-value

for b=1:B
% use MATLAB's randperm function to get a random permutation of indices
PermutedIndices=randperm(n_tot);
% use the first n_TBi of PermutedIndices for the bootstrapped TB data
B_TB=tot(PermutedIndices(1:n_TBi));
% use the last n_nTBi of the PermutedIndices for bootstrapped non TB data
B_nTB=tot(PermutedIndices(n_TBi+1:n_tot));

% compute the test statistic for the bootstrapped data
[h, p,dist]= kstest2(B_TB,B_nTB,0.05,'unequal');

% increment the ApproxPValue by 1/B if bootstrapped dist > ObsKSvalue
if(dist>ObsKSvalue)
ApproxPValue=ApproxPValue+(1/B);
```

```
end
end
ApproxPVal=[ApproxPVal; i, ApproxPValue];
end
ApproxPVal % report the Approximate p—value
```

CDF confidence bands coding:

```
% CDF confidence band widths for each variable.
% Excluding NaN values from length of variables which alters CI

n_var=[99 99]
n_NANvar=[99 99]
Epsn_var=[99 99 99]
Alpha=0.05              %Level of test at 5%

for i=1:1:29
    TBi=sort(data(1:472,i));            %TB values
    nTBi=sort(data(473:118857,i));      %non TB values

    NANplaceTB = isnan(data(1:472,i));  %Finding TB NaN values
    NANcountTB = sum(sum(NANplaceTB))   %counting TB NaNs
    n_TBi=length(TBi) — NANcountTB;     %# of TB values (excl NaNs)

    NANplacenTB = isnan(data(473:118857,i));%Finding nTB NaN values
    NANcountnTB = sum(sum(NANplacenTB))     %counting nTB NaNs
    n_nTBi=length(nTBi) — NANcountnTB;      %# of nTB values (excl NaNs)

    n_NANvar=[n_NANvar; NANcountTB NANcountnTB]  %showing # NaNs
    n_var=[n_var; n_TBi n_nTBi];                 %showing # data values

    Epsn_TB = sqrt((1/(2*n_TBi))*log(2/Alpha));  %calculating TB CI width
    Epsn_nTB = sqrt((1/(2*n_nTBi))*log(2/Alpha));%calculating nTB CI width
    Epsn_var=[Epsn_var; Epsn_TB Epsn_nTB Epsn_TB+Epsn_nTB]; %showing CI
end

n_NANvar
n_var
Epsn_var
```

Nearest neighbour coding

```
% Coding for nearest neighbour

[m,n] =size(data)
matsumresub_x=zeros(n,1);
matminresub_x=zeros(n,1);
matsumcval_x=zeros(n,1);
matmincval_x=zeros(n,1);

%Number of nearest neighbours looked at
x=5

for i=1:n
    for j=1:n;
mrnnl = ClassificationKNN.fit([data(:,i) data(:,j)], TBind(:,1),...
    'NumNeighbors',x);
rloss = resubLoss(mrnnl);
crossvallloss = kfoldLoss(crossval(mrnnl));
matsumresub_x(i)=matsumresub_x(i)+rloss;
matminresub_x(i)=min(matsumresub_x(i),rloss);
matsumcval_x(i)=matsumcval_x(i)+crossvallloss;
matmincval_x(i)=min(matsumcval_x(i),crossvallloss);
    end
end

matResub=[matsumresub_x matminresub_x]
matCval=[matsumcval_x matmincval_x]
```

Cross-validating model coding:

```
# finding percentage of TB and non TB population predicted to have TB
# from the 1pecent model.

setwd('S:/icv10/private/SecondaryAnalysis/Data/IAMR52SD_HIV')
mr=read.table('MR_HIV_imputed_New.txt',na.strings="NA", header=TRUE)

length(mV191)
xxa=data.frame(rep(seq(1,10,1), 7261))
xx=xxa[1:72607,]

mrdata=data.frame(xx,sm31a=factor(mr$SM31A), mv475=factor(mr$MV475),
     mv024=factor(mr$MV024new), sm410c=factor(mr$SM410CNew),
     mv836=factor(mr$MV836new),  sb69=factor(mr$SB69new),
     mV133=factor(mr$MV133new), mV437=log(mr$MV437), mV438=log(mr$MV438),
     mV456=(mr$MV456)^2, mV191=((mr$MV191)+174001)^(1/3), MV012=mr$MV012)
head(mrdata)

segment=3   #between 1 and 10 to indicate which segment left out of model
valin=data.frame(mrdata[xx!=segment,])
valout=data.frame(mrdata[xx==segment,])

MRglm1pc = glm(sm31a ~ mv475 + mv024  + sm410c + mv836 + sb69 +
       mV133+ mV437 + mV438 + mV456 + mV191 + MV012 +
       sm410c:MV012 + mv024:mV438 + mv024:mV437,
       data=valin, family = binomial)

predMR <- predict(MRglm1pc, newdata=valout, type="response")
MRtruth=data.frame(mrdata[xx==segment,2])

for (i in seq(0.001, 0.03, 0.0005)) {
   predMR2=ifelse(predMR> i, 'TB', 'nTB')
   Compare=data.frame(MRtruth, predMR2)
predTB=(table(Compare)[1,2]/ (table(Compare)[1,1]+table(Compare)[1,2]))
predTB=(table(Compare)[2,2]/ (table(Compare)[2,1]+table(Compare)[2,2]))
result=c(i, predTB, predTB, predTB-predTB)
print(result)
}
```

Creating summaries of variables in R:

```
#creating summary of variables

dat.mrnew<-matrix(ncol=6,nrow=0)
for(i in unique(MV133new))
    {
    for(j in unique(MV026))
        {
        for(k in unique(SMNFHS2))
            {
            for(l in unique(SMSLI))
                {
                for(m in unique(MV159))
                    {
            count<-length(mr[mr[,"MV133new"]==i & mr[,"MV026"]==j &
            mr[,"SMNFHS2"]==k & mr[,"SMSLI"]==l & mr[,"MV159"]==m,1])
        v_a<-i
        v_b<-j
        v_c<-k
        v_d<-l
        v_e<-m
        new.entry<-cbind(v_a,v_b,v_c, v_d, v_e, count)
        dat.mrnew<-rbind(dat.mrnew,new.entry)
                    }
                }
            }
        }
    }
dat.mrnew
```

Creating Directed Acyclic Graphs in Graphviz:

```
## Creating male directed acyclic graph

digraph G {
ratio=1.5
forcelabels=true
size="14";
node [color=lightblue,style=filled];

133  [label="Education level"];
201 [label="#Children born"];
218 [label="#Living children"];
202 [label="#Sons@home"];
203 [label="#Daughters@home"];
136 [label="#Household members"];
511 [label="Age@marriage"];
836 [label="#Sex partners"];
HIV[label="HIV weight"];
157[label="Frequency of \n reading paper"];
024 [label="State"];
026 [label="Country/city"];
smsli [label="Living standard"];
smnfhs2 [label="House type"];
159  [label="Frequency of \n watching tv"];

456 [label="Hemoglobin level"];
437  [label="Weight"];
438  [label="Height"];
191  [label="Wealth Index \n score"];
012  [label="Age"];
152  [label="Age of head \n of household"];


456 -> 024 -> 026 -> smnfhs2 [style = bold];
024 -> smnfhs2 [style = bold];
024 -> HIV[style = bold];
026 -> HIV[style = bold];
456 -> HIV[style = bold];

159 -> 133 -> 511[style = bold];
smnfhs2  -> smsli[style = bold];
smsli  -> 133 [style = bold];
smnfhs2 -> 159 -> smsli [style = bold];
159 -> 157 [style = bold];

191 -> smnfhs2 [style = bold];
191 -> smsli [style = bold];
191 -> 437 [style = bold][label="**"];
191 -> 438 [style = bold][label="**"];
437 -> 438 [style = bold][label="**"];

511 -> 836 [style = bold];
511 -> 218 -> 201 [style = bold];
202 -> 218 [style = bold];
203 -> 218  [style = bold];
218 -> 136 [style = bold][label="**"];
218 -> 012 ->  201 [style = bold];
152 -> 218 [style = bold];
152 -> 201 [style = bold];
152 -> 136 [style = bold];

}
```

[This page intentionally left blank]

References

[1] Aarti Kaulagekar and Anjali Radkar. Social status makes a difference: Tuberculosis scenario during National Family Health Survey - 2. *Indian Journal of Tuberculosis*, 54(542):17–23, 2006. http://www.ncbi.nlm.nih.gov/pubmed/17455419.

[2] WHO. Global Tuberculosis Report 2012. http://www.who.int/tb/publications/global_report/gtbr12_main.pdf, 2012.

[3] Helen A Fletcher, Helen D Donoghue, John Holton, Ildikó Pap, and Mark Spigelman. Widespread occurrence of Mycobacterium tuberculosis DNA from 18th-19th century Hungarians. *American Journal of Physical Anthropology*, 120(2):144–152, 2003. http://discovery.ucl.ac.uk/628/.

[4] Israel Hershkovitz, Helen D Donoghue, David E Minnikin, Gurdyal S Besra, Oona Y-C Lee, Angela M Gernaey, Ehud Galili, Vered Eshed, Charles L Greenblatt, Eshetu Lemma, Gila Kahila Bar-Gal, and Mark Spigelman. Detection and Molecular Characterization of 9000-Year-Old Mycobacterium tuberculosis from a Neolithic Settlement in the Eastern Mediterranean. *PLoS ONE*, 3(10):6, 2008. http://discovery.ucl.ac.uk/14877/.

[5] Albert R Zink, Christophe Sola, Udo Reischl, Nalin Rastogi, Hans Wolf, Andreas G Nerlich, and Waltraud Grabner. Characterization of Mycobacterium tuberculosis Complex DNAs from Egyptian Mummies by Spoligotyping Characterization of Mycobacterium tuberculosis Complex DNAs from Egyptian Mummies by Spoligotyping. 2003. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC149558/.

[6] Core Curriculum on Tuberculosis: What the Clinician Should Know. http://www.cdc.gov/tb/education/corecurr/index.htm, 2011. Published by: Centers for Disease Control and Prevention.

[7] P Hudelson. Gender differentials in tuberculosis: The role of socio-economic and cultural factors. *Tubercle and Lung Disease*, 77(5):391–400, 1996. http://linkinghub.elsevier.com/retrieve/pii/S0962847996901100.

[8] WHO. Global Tuberculosis Report 2012: Annex 2. http://www.who.int/tb/publications/global_report/gtbr12_annex2.pdf, 2012.

[9] Vinod K Diwan and Anna Thorson. Sex, gender, and tuberculosis. *The Lancet*, (3):1000–1001. http://www.thelancet.com/journals/lancet/article/PIIS0140...5/.

[10] Olivier Neyrolles and Lluis Quintana-Murci. Sexual Inequality in Tuberculosis. *PLoS Med*, 6(12), 2009. e1000199.doi:10.1371/journal.pmed.1000199.

[11] Andrew J Codlin, Saira Khowaja, Zhongxue Chen, Mohammad H Rahbar, Ejaz Qadeer, Ismat Ara, Joseph B McCormick, Susan P Fisher-Hoch, and Aamir J Khan. Short report: Gender differences in tuberculosis notification in Pakistan. *American Journal Of Tropical Medicine And Hygiene*, 85(3):514–517, 2011. http://www.ajtmh.org/content/85/3/514.full.pdf+html.

[12] M G Weiss, D Somma, F Karim, A Abouihia, C Auer, J Kemp, and M S Jawahar. Cultural epidemiology of TB with reference to gender in Bangladesh, India and Malawi. *The international journal of tuberculosis and lung disease the official journal of the International Union against Tuberculosis and Lung Disease*, 12(7):837–847, 2008. http://www.ncbi.nlm.nih.gov/pubmed/18544214.

[13] Nguyen Binh Hoa, Dinh Ngoc Sy, Nguyen Viet Nhung, Edine W Tiemersma, Martien W Borgdorff, and Frank G J Cobelens. National survey of tuberculosis prevalence in Viet Nam. *Bulletin of the World Health Organization*, 88(4):273–280, 2010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2855599&tool=pmcentrez&rendertype=abstract.

[14] E Johansson, N H Long, V K Diwan, and A Winkvist. Attitudes to compliance with tuberculosis treatment among women and men in Vietnam. *The international journal of tuberculosis and lung disease the official journal of the International Union against Tuberculosis and Lung Disease*, 3(10):862–868, 1999. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10524582.

[15] WHO. WHO website. http://www.who.int/topics/tuberculosis/en/, 2013.

[16] D Falzon, E Jaramillo, H J Schünemann, M Arentz, M Bauer, J Bayona, L Blanc, J A Caminero, C L Daley, C Duncombe, C Fitzpatrick, A Gebhard, H Getahun, M Henkens, T H Holtz, J Keravec, S Keshavjee, A J Khan, R Kulier, V Leimane, C Lienhardt, C Lu, A Mariandyshev, G B Migliori, F Mirzayev, C D Mitnick, P Nunn, G Nwagboniwe, O Oxlade, D Palmero, P Pavlinac, M I Quelapio, M C Raviglione, M L Rich, S Royce, S Rüsch-Gerdes, A Salakaia, R Sarin, D Sculier, F Varaine, M Vitoria, J L Walson, F Wares, K Weyer, R A White, and M Zignol. WHO guidelines for the programmatic management of drug-resistant tuberculosis: 2011 update. *The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology*, 38(3):516–528, 2011. http://www.ncbi.nlm.nih.gov/pubmed/21828024.

[17] Francis Drobniewski, Yanina Balabanova, Vladyslav Nikolayevsky, Micheal Ruddy, Sergey Kuznetzov, Svetlana Zakharova, Alexander Melentyev, and Ivan Fedorin. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. *Jama The Journal Of The American Medical Association*, 293(22):2726–2731, 2005. http://www.ncbi.nlm.nih.gov/pubmed/15941801.

[18] S Supcharoen. Management of drug resistant tuberculosis. *Bulletin Of The International Union Against Tuberculosis*, 49 suppl 1(August):286–287, 1974. http://whqlibdoc.who.int/hq/1997/who_tb_96.210_(rev.1).pdf.

[19] Amir Khan, John Walley, James Newell, and Imdad Naghma. Tuberculosis in Pakistan: socio-cultural constraints and opportunities in treatment. *Social Science & Medicine*, 50(2):247–254, 2000.

[20] M. S. Al-Hajjaj and I. M. Al-Khatim. High rate of non-compliance with anti-tuberculosis treatment despite a retrieval system : a call for implementation. *Int J Tuberc Lung Dis*, 4(July 1998):345–349, 2000. http://www.ncbi.nlm.nih.gov/pubmed/10777084.

[21] Aliabbas A Husain, Rajpal S Kashyap, Devanand R Kalorey, Shubangi R Warke, Hemant J Purohit, Girdhar M Taori, and Hatim F Daginawal. Effect of repeat dose of BCG vaccination on humoral response in mice model. *Indian Journal of Experimental Biology*, 49(1):7–10, January 2011. http://www.ncbi.nlm.nih.gov/pubmed/21365989.

[22] Alexis Cambanis, Mohammed A Yassin, Andy Ramsay, S Bertel Squire, Isabel Arbide, and Luis E Cuevas. Rural poverty and delayed presentation to tuberculosis services in Ethiopia. *Tropical Medicine & International Health*, 10(4):330–335, April 2005. http://www.ncbi.nlm.nih.gov/pubmed/15807796.

[23] D Vukovic, L Nagorni-Obradovic, and V Bjegovic. Knowledge and misconceptions of tuberculosis in the general population in Serbia. *European journal of clinical microbiology & infectious diseases*, 27(9):761–767, September 2008. http://www.ncbi.nlm.nih.gov/pubmed/18401603.

[24] M Uplekar, S Juvekar, S Morankar, S Rangan, and P Nunn. Tuberculosis patients and practitioners in private clinics in India. *The International Journal of Tuberculosis and Lung Disease*, 2(4):324–329, 1998. http://www.ingentaconnect.com/content/iuatld/ijtld/1998/00000002/00000004/art00010.

[25] Salla A Munro, Simon A Lewin, Helen J Smith, Mark E Engel, Atle Fretheim, and Jimmy Volmink. Patient Adherence to Tuberculosis Treatment: A Systematic Review of Qualitative Research. *PLoS Medicine*, 4(7):16, 2007. http://research-archive.liv.ac.uk/307/.

[26] K Jaggarajamma, G Sudha, V Chandrasekaran, C Nirupa, A Thomas, T Santha, M Muniyandi, and P R Narayanan. Reasons for non-compliance among patients treated under Revised National Tuberculosis Control Programme (RNTCP), Tiruvallur district, south India. *The Indian journal of tuberculosis*, 54(3):130–135, 2007. http://www.ncbi.nlm.nih.gov/pubmed/17886701.

[27] Brandon DL Marshall, Thomas Kerr, Jean A Shoveller, Julio SG Montaner, and Evan Wood. An assessment of factors contributing to treatment adherence and knowledge of TB transmission among patients on TB treatment. *BMC Public Health*, 4(7):68, 2004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2630937&tool=pmcentrez&rendertype=abstract.

[28] R. Liefooghe, N. Michiels, S. Habib, M.B. Moran, and A. De Muynck. Perception and social consequences of tuberculosis: A focus group study of tuberculosis patients in Sialkot, Pakistan. *Social Science & Medicine1*, 41(12):1685–1692, 1995.

[29] A J Rubel and L C Garro. Social and cultural factors in the successful control of tuberculosis. *Public Health Reports*, 107(6):626–636, 1992. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1403712&tool=pmcentrez&rendertype=abstract.

[30] Neel R Gandhi, Anthony Moll, A Willem Sturm, Robert Pawinski, Thiloshini Govender, Umesh Lalloo, Kimberly Zeller, Jason Andrews, and Gerald Friedland. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lance*, 368(9547):1575–1580, 2006. http://www.ncbi.nlm.nih.gov/pubmed/17084757.

[31] Charles-Edward Amory Winslow. The untilled fields of public health. *Science*, 51(1306):23–33, 1920. http://www.jstor.org/stable/1645011?origin=ads.

[32] STOP TB partnership. http://www.stoptb.org/about/, 2013.

[33] WHO. 5 steps of DOTS. http://www.who.int/tb/dots/whatisdots/en/index.html, 2013.

[34] P Arokiasamy, K Karthick, and J Pradhan. Environmental risk factors and prevalence of asthma, tuberculosis and jaundice in India. *International Journal of Environment and Health*, 1(2):221–242, 2007.

[35] N Shetty, M Shemko, M Vaz, and G D'Souza. An epidemiological evaluation of risk factors for tuberculosis in South India: a matched case control study. *The International Journal of Tuberculosis and Lung Disease*, 10(1):80–86, January 2006. http://www.ncbi.nlm.nih.gov/pubmed/16466042.

[36] India's 2005-06 National Family Health Survey, 2005. http://www.measuredhs.com/pubs/pdf/SR128/SR128.pdf.

[37] S I Rajan and K S James. Third National Family Health Survey in India: issues, problems and prospects. *Economic and Political Weekly*, 43(48):33–38, 2008. http://www.epw.org.in.

[38] Indrajit Hazarika. Womens Reproductive Health in Slum Populations in India: Evidence From NFHS-3. *Journal of urban health: bulletin of the New York Academy of Medicine*, 87(2):264–277, 2010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2845837&tool=pmcentrez&rendertype=abstract.

[39] The DHS Wealth Index: Approaches for Rural and Urban Areas. http://www.measuredhs.com/pubs/pdf/SR128/SR128.pdf, 2008.

[40] M Boelaert, F Meheus, A Sanchez, S P Singh, V Vanlerberghe, A Picado, B Meessen, and S Sundar. The poorest of the poor: a poverty appraisal of households affected by visceral leishmaniasis in Bihar, India. *Tropical medicine international health TM IH*, 14(6):639–644, 2009. http://www.ncbi.nlm.nih.gov/pubmed/19392741.

[41] Olivia Oxlade and Megan Murray. Tuberculosis and Poverty: Why Are the Poor at Greater Risk in India? *PLoS ONE*, 7(11):e47533, November 2012. http://dx.plos.org/10.1371/journal.pone.0047533.

[42] Hélène F Delisle. Poverty: the double burden of malnutrition in mothers and the intergenerational impact. *Annals Of The New York Academy Of Sciences*, 1136:172–84, 2008. http://www.ncbi.nlm.nih.gov/pubmed/18579881.

[43] David Arnold. The medicalization of poverty in colonial India*. *Historical Research*, 85(229):488–504, August 2012. http://doi.wiley.com/10.1111/j.1468-2281.2012.00596.x.

[44] V L Pandey and A Chaubal. Comprehending household cooking energy choice in rural India. *Biomass and Bioenergy*, 35(11):4724–4731, 2011. http://www.scopus.com/inward/record.url?eid=2-s2.0-81555219074&partnerID=40&md5=338bfa7e8ed7e5ab8fb4db1135d83bba.

[45] Wesley Foell, Shonali Pachauri, Daniel Spreng, and Hisham Zerriffi. Household cooking fuels and technologies in developing economies. *Energy Policy*, 39(12):7487–7496, 2011. http://linkinghub.elsevier.com/retrieve/pii/S0301421511006136.

[46] Japhet Killewo. Poverty, TB, and HIV infection: a vicious cycle. *Journal of Health, Population, and Nutrition*, 20(4):281–284, December 2002. http://www.ncbi.nlm.nih.gov/pubmed/12659406.

[47] Richard Coker, Martin McKee, Rifat Atun, Boika Dimitrova, Ekaterina Dodonova, Sergei Kuznetsov, and Francis Drobniewski. Risk factors for pulmonary tuberculosis in Russia: case-control study. *BMJ*, 332(7533):85–87, 2006. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1326929&tool=pmcentrez&rendertype=abstract.

[48] Vendhan Gajalakshmi, Richard Peto, Thanjavur Santhanakrishna Kanaka, and Prabhat Jha. Smoking and mortality from tuberculosis and other diseases in India: retrospective study of 43000 adult male deaths and 35000 controls. *Lancet*, 362(9383):507–515, August 2003. http://www.ncbi.nlm.nih.gov/pubmed/12932381.

[49] P Godoy, A Domínguez, J Alcaide, N Camps, J M Jansà, S Minguell, J M Pina, and M Díez. Characteristics of tuberculosis patients with positive sputum smear in Catalonia, Spain. *European Journal of Public Health*, 14(1):71–75, March 2004. http://www.ncbi.nlm.nih.gov/pubmed/15080395.

[50] Monica Das Gupta. Selective Discrimination against Female Children in Rural Punjab, India. *Population and Development Review*, 13(1):77–100, 1987. http://www.jstor.org/stable/1972121.

[51] Per Gustafson, Victor F Gomes, Cesaltina S Vieira, Paulo Rabna, Rémonie Seng, Peter Johansson, Anita Sandström, Renée Norberg, Ida Lisse, Badara Samb, Peter Aaby, and Anders Nauclér. Tuberculosis in Bissau: incidence and risk factors in an urban community in sub-Saharan Africa. *International Journal of Epidemiology*, 33(1):163–172, February 2004. http://www.ncbi.nlm.nih.gov/pubmed/15075165.

[52] Guy Harling, Rodney Ehrlich, and Landon Myer. The social epidemiology of tuberculosis in South Africa: A multilevel analysis. *Social Science & Medicine*, 66(2):492–505, January 2008. http://www.ncbi.nlm.nih.gov/pubmed/17920743.

[53] J I Hawker, S S Bakhshi, S Ali, and C P Farrington. Ecological analysis of ethnic differences in relation between tuberculosis and poverty. *BMJ*, 319(7216):1031–1034, October 1999. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=28253&tool=pmcentrez&rendertype=abstract.

[54] V K Mishra, R D Retherford, and K R Smith. Biomass cooking fuels and prevalence of tuberculosis in India. *International Journal of Infectious Diseases*, 3(3):119–129, January 1999. http://www.ncbi.nlm.nih.gov/pubmed/10460922.

[55] Delia Boccia, James Hargreaves, Bianca Lucia De Stavola, Katherine Fielding, Ab Schaap, Peter Godfrey-Faussett, and Helen Ayles. The association between household socioeconomic position and prevalent tuberculosis in Zambia: A case-control study. *PloS one*, 6(6):e20824, January 2011. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3117783&tool=pmcentrez&rendertype=abstract.

[56] Ward P Myers, Janice L Westenhouse, Jennifer Flood, and Lee W Riley. An ecological study of tuberculosis transmission in California. *American Journal of Public Health*, 96(4):685–690, April 2006. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1470555&tool=pmcentrez&rendertype=abstract.

[57] M Baker, D Das, K Venugopal, and P Howden-Chapman. Tuberculosis associated with household crowding in a developed country. *Journal of Epidemiology and Community Health*, 62(8):715–721, August 2008. http://www.ncbi.nlm.nih.gov/pubmed/18621957.

[58] Stephen D Lawn and Alimuddin I Zumla. Tuberculosis. *Lancet*, 378(9785):57–72, July 2011. http://www.ncbi.nlm.nih.gov/pubmed/21420161.

[59] Christian Lienhardt and Jessica Ann Ogden. Tuberculosis control in resource-poor countries: have we reached the limits of the universal paradigm? *Tropical Medicine & International Health*, 9(7):833–841, July 2004. http://www.ncbi.nlm.nih.gov/pubmed/15228495.

[60] Richard Frothingham, Jason E Stout, and Carol Dukes Hamilton. Current issues in global tuberculosis control. *International Journal of Infectious Diseases*, 9(6):297–311, November 2005. http://www.ncbi.nlm.nih.gov/pubmed/16183319.

[61] D M Nair, a George, and K T Chacko. Tuberculosis in Bombay: new insights from poor urban patients. *Health policy and planning*, 12(1):77–85, March 1997. http://www.ncbi.nlm.nih.gov/pubmed/10166105.

[62] F Barnhoorn and H Adriaanse. In search of factors responsible for noncompliance among tuberculosis patients in Wardha District, India. *Social Science & Medicine*, 34(3):291–306, February 1992. http://www.ncbi.nlm.nih.gov/pubmed/1557670.

[63] Rijo M John. Tobacco consumption patterns and its health implications in India. *Health Policy*, 71(2):213–222, 2005. http://dx.doi.org/10.1016/j.healthpol.2004.08.008.

[64] Valerie Moller, Ida Erstad, and Dalinyebo Zani. Drinking, Smoking, and Morality: Do Drinkers and Smokers Constitute a Stigmatised Stereotype or a Real TB Risk Factor in the Time of HIV/AIDS? *Social Indicators Research*, 98(2):217–238, November 2009. http://www.springerlink.com/index/10.1007/s11205-009-9546-2.

[65] M Harper, F A Ahmadu, J A Ogden, K P McAdam, and C Lienhardt. Identifying the determinants of tuberculosis control in resource-poor countries: insights from a qualitative study in The Gambia. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 97(5):506–510, 2003. http://www.ncbi.nlm.nih.gov/pubmed/15307411.

[66] Javaid Ahmed Khan, Muhammad Irfan, Amna Zaki, Madiha Beg, Syed Fayyaz Hussain, and Nadeem Rizvi. Knowledge, Attitude and Misconceptions regarding tuberculosis in Pakistani patients. *Journal of the Pakistan Medical Association*, 56(5):211–214, 2006. http://www.ncbi.nlm.nih.gov/pubmed/16767946.

[67] Y Balarajan, S Selvaraj, and S V Subramanian. Health care and equity in India. *Lancet*, 377(9764):505–515, 2011. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3093249&tool=pmcentrez&rendertype=abstract.

[68] 2005-06 National Family Health Survey (NFHS-3) Key Findings. http://www.measuredhs.com/pubs/pdf/SR128/SR128.pdf, 2005.

[69] The WHO. Global tuberculosis control: WHO report 2011. *The WHO*, page 258, 2011. http://www.who.int/tb/publications/global_report/en/index.html.

[70] Indira Hirway. Time use survey data for improving Tabulation and Analysis of the Indian Time Use Survey Data for Improving Measurement of Paid and Unpaid Work. Technical Report October.

[71] Nisha Srivastava and Ravi Srivastava. Women, Work, and Employment Outcomes in Rural India. *Economic and Political Weekly*, xlv(28):49–63, 2010. http://www.epw.in/special-articles/women-work-and-employment-outcomes-rural-india.html.

[72] S Ramesh Kumar, Soumya Swaminathan, Timothy Flanigan, K H Mayer, and Raymond Niaura. HIV & smoking in India. *The Indian journal of Medical Research*, 130(1):15–22, 2009. http://www.ncbi.nlm.nih.gov/pubmed/19700796.

[73] T Rooban, P D Madan Kumar, and K Ranganathan. Reach of mass media among tobacco users in India: a preliminary report. *Indian Journal of Cancer*, 47 Suppl 1(5):53–58, 2010. http://www.ncbi.nlm.nih.gov/pubmed/20622415.

[74] Rakesh Munshi and Sang-Hyop Lee. Child Immunization in Madhya Pradesh. https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/3476/NFHSsubjrpt015.pdf;jsessionid=7B9738431800E0AB658C81FD962D5032?sequence=1, 2000.

[75] Anita Raj, Niranjan Saggurti, Donta Balaiah, and Jay G Silverman. Prevalence of child marriage and its effect on fertility and fertility-control outcomes of young women in India: a cross-sectional, observational study. *The journal of family planning and reproductive health care Faculty of Family Planning Reproductive Health Care Royal College of Obstetricians Gynaecologists*, 373(9678):1883–1889Y, 2009. http://www.scopus.com/inward/record.url?eid=2-s2.0-65849120671&partnerID=40&md5=89f0848a43344de779bda385ab7b3a96.

[76] Nawal M Nour. Child Marriage: A Silent Health and Human Rights Issue. *Reviews in obstetrics and gynecology*, 2(1):51–56, 2009. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672998&tool=pmcentrez&rendertype=abstract.

[77] Geeta Gandhi Kingdon. The progress of school education in India. 23(2):168–195, 2007.

[78] Literacy rates from India's Census. http://www.census2011.co.in/literacy.php, 2011.

[79] Chandrakant Lahariya and Jyoti Khandekar. How the findings of national family health survey-3 can act as a trigger for improving the status of anemic mothers and undernourished children in India: a review. *Indian Journal of Medical Sciences*, 61(9):535–544, 2007. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med4&AN=17785892.

[80] Leland K Ackerson and S V Subramanian. Domestic violence and chronic malnutrition among women and children in India. *American Journal of Epidemiology*, 167(10):1188–1196, 2008. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2789268&tool=pmcentrez&rendertype=abstract.

[81] Santosh Mehrotra. Child Malnutrition and Gender Discrimination in South Asia. *Economic and Political Weekly*, 41(10):912–918, 2006.

[82] The WHO. Tuberculosis Control in the South-East Asia Region. Technical report, 2012. http://www.searo.who.int/LinkFiles/Tuberculosis_WHO-TB-Report-2012.pdf.

[83] S V Subramanian, Leland K Ackerson, George Davey Smith, and John A Neetu. Association of maternal height with child mortality, anthropometric failure, and anemia in India. *The Journal Of The American Medical Association*, 301(16):1691–1701, 2009. http://www.ncbi.nlm.nih.gov/pubmed/19383960.

[84] William Joe, U. S Mishra, and K Navaneetham. Health Inequality in India : Evidence from NFHS 3. *Economic and Political Weekly*, (31):41–47.

[85] R Radhakrishna and C Ravi. Malnutrition in India: Trends and Determinants. *Economic and Political Weekly*, 39(7):671–676, 2004. http://www.jstor.org.ezp-prod1.hul.harvard.edu/stable/4414642.

[86] Sheila Isanaka, Ferdinand Mugusi, Willy Urassa, Walter C Willett, Ronald J Bosch, Eduardo Villamor, Donna Spiegelman, Christopher Duggan, and Wafaie W Fawzi. Iron deficiency and anemia predict mortality in patients with tuberculosis. *The Journal of nutrition*, 142(2):350–357, 2012.

[87] E Saathoff, E Villamor, F Mugusi, R J Bosch, W Urassa, and W W Fawzi. Anemia in adults with tuberculosis is associated with HIV and anthropometric status in Dar es Salaam, Tanzania. *The international journal of tuberculosis and lung disease the official journal of the International Union against Tuberculosis and Lung Disease*, 15(7):925–932, 2011.

[88] Arvind Pandey, Dandu C S Reddy, Peter D Ghys, Mariamma Thomas, Damodar Sahu, Madhulekha Bhattacharya, Kanchan D Maiti, Fred Arnold, Shashi Kant, Ajay Khera, and Renu Garg. Improved estimates of India's HIV burden in 2006. *The Indian journal of Medical Research*, 129(1):50–58, 2009.

[89] T Jacob John, Lalit Dandona, Vinod P Sharma, and Manish Kakkar. Continuing challenge of infectious diseases in India. *Lancet*, 377(9761):252–269, January 2011. http://www.ncbi.nlm.nih.gov/pubmed/21227500.

[90] Michele R Decker, George R Seage, David Hemenway, Anita Raj, Niranjan Saggurti, Donta Balaiah, and Jay G Silverman. Intimate partner violence functions as both a risk marker and risk factor for women's HIV infection: findings from Indian husband-wife dyads. *Journal of Acquired Immune Deficiency Syndromes*, 51(5):593–600, 2009. http://www.ncbi.nlm.nih.gov/pubmed/19421070.

[91] Padma Chandrasekaran, Gina Dallabetta, Virginia Loo, Sujata Rao, Helene Gayle, and Ashok Alexander. Containing HIV/AIDS in India: the unfinished agenda. *The Lancet infectious diseases*, 6(8):508–521, 2006. http://www.ncbi.nlm.nih.gov/pubmed/16870529.

[92] Lakshmi Ramakrishnan, Abhishek Gautam, Prabuddhagopal Goswami, Srinivasan Kallam, Rajatashuvra Adhikary, Mandar K Mainkar, Banadakoppa M Ramesh, Guy Morineau, Bitra George, and Ramesh S Paranjape. Programme coverage, condom use and STI treatment among FSWs in a large-scale HIV prevention programme: results from cross-sectional surveys in 22 districts

in southern India. *Sexually Transmitted Infections*, 86 Suppl 1(Suppl 1):62–68, 2010. http://www.ncbi.nlm.nih.gov/pubmed/20167734.

[93] Frank J Jr Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. http://www.jstor.org/stable/2280095.

[94] Tyler J VanderWeele and James M Robins. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *American Journal of Epidemiology*, 166(9):1096–1104, 2007. http://www.ncbi.nlm.nih.gov/pubmed/17702973.

[95] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(September):96–146, 2009. http://projecteuclid.org/euclid.ssu/1255440554.

[96] R.W Robinson. *Counting Labeled Acyclic Digraphs*. 1973.

[97] J Bartlett. Tuberculosis and HIV Infection: Partners in Human Tragedy. (5):1–2, 2007. http://jid.oxfordjournals.org/content/196/Supplement_1/S124.full.pdf+html.

[98] Measured DHS surveys. http://www.measuredhs.com/What-We-Do/Survey-Search.cfm, 2005.