International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015
ISSN 2091-2730

# A data mining perspective on prevalence of Tuberculosis in India

Natalya Kumar

B.Tech - Mody University, bestalya@gmail.com, +919572767969

**Abstract**— Abstract Tuberculosis (TB) is a widespread infectious disease caused by various strains of mycobacteria, usually Mycobacterium tuberculosis. It is spread through the air by a person suffering from TB and typically attacks the lungs. A single patient can infect 10 or more people in a year. More people in the developing world contract tuberculosis because of a poor immune system. India has the highest burden of TB in the world, an estimated 2 million cases annually, and accounting for approximately one fifth of the global incidence. It is also estimated by the World Health Organization (WHO) that 300,000 people die from TB each year in India. Eighty per cent of TB patients are in the economically most productive years of their lives. The purpose of the current work is to analyze the TB case notification statistics for India and the Treatment Outcomes of Notified TB Cases to make future predictions. Estimating Tuberculosis disease burden is important for planning, monitoring and evaluating the TB control programme.

**Keywords**— Tuberculosis, India, RNTCP, Diagnosis, Notification, Prevalence, Incidence, Mortality, Success, Eradication, Data Mining.

## INTRODUCTION

India bears a disproportionately large burden of the world's tuberculosis rates, as it resides to be the biggest health problem in India. It remains one of the largest on India's health and wellness scale. India has a large burden of the world's TB, one that this developing country can ill afford, with an estimated economic loss of US $43 billion and 100 million lost annually directly due to this disease. Treatment in India is on the rise just as the disease itself is on the rise. To prevent spreading TB, it's important to get treatment quickly and to follow it through to completion. The Indian government's Revised National TB Control Programme (RNTCP) started in India during 1997. The program uses the WHO recommended "Directly Observed Treatment Short" Course (DOTS) strategy to develop ideas and data on TB treatment. This group's initial objective is to achieve and maintain a TB treatment success rate of at least 85% in India among new patients. In 2010 the RNTCP made a major policy decision that it would change focus and adopt the concept of Universal Access to quality diagnosis and TB treatment for all TB patients. By doing so, they extend out a helping hand to all people diagnosed with TB, and in addition, provide better quality services and improve on therapy for these patients.

In this paper, Statistics from the past years regarding the TB case notification and treatment outcomes in India are used for making estimations and future predictions. Estimating tuberculosis disease burden is important for planning, monitoring and evaluating the TB control program. Weka tool is used for the evaluation and forecast using regression algorithms.

## METHODOLOGY

Over the 11 year analysis period, the population covered increased from 139 million to 1.21 billion populations. Smear microscopy services are reported independently of case notification results. As expected from service expansion, the absolute number of TB suspects examined by smear microscopy annually has increased manifold, from 0.96 million to 7.8 million. Over the same time period, the rate of TB suspect examination also increased by 50%, from 421 per 100,000 population covered by RNTCP services to 651 per 100,000 population covered. Similarly, the rate of sputum smear positive cases diagnosed by microscopy has increased by 20%, from 62 to 79 per 100,000 population [Figure 1]. The average number of suspects examined for every sputum smear positive case diagnosed has gradually increased about 1.3% per year, from 2001 to 2011, the number of suspects examined per smear positive

case diagnosed has increased by 28% from 6.4 to 8.3 suspects (Figure 2). Total and sputum smear positive case notification is also shown in Table 1. An average difference of 11.3% [Range 8– 15%] was observed between the rate of sputum-positive cases diagnosed and the sputum-positive case notification rate.

## Effectiveness of RNTCP and its future

Though the programme RNTCP was formally initiated in the year 1997 and the quarterly reporting mechanism was in place since inception, the data presented below extend from the year 1999, when approximately about 10% of the country's population was covered onwards and makes a prediction till the year xxxx. The rapid pace of DOTS expansion over the past decade complicates longitudinal data analysis in a number of ways. District-bydistrict scale-up of RNTCP services over several years changes the denominator of population covered every quarter. Basic demographic characteristics of implementing districts differed over the expansion years, as well as the expected evolution of services and TB epidemiology in areas implementing RNTCP over longer time periods. The future trend is based on the past records to check the effectiveness of the programme.

For the purposes of this analysis, districts implementing RNTCP less than one year during the initial year of implementation were attributed to cover a population proportionate to the number of days in the first year that services were available in each district. The rates presented in this section are all per 100,000 populations after adjusting for the number of days of implementation by individual districts till year 2006. Also the population of the districts is based on 2001 census and 2011 Census India for these two years and estimated for the rest of the years based on these two Censuses. Though the population in the tables is complete population of services covered as on 31st December of that year.

## Sputum Microscopy Services and TB Suspect Examination

Over the 11 year from 1999 to 2011 analysis period, the population covered increased from 139 million to 1.21 billion populations. Smear microscopy services are reported independently of case notification results. As expected from service expansion, the absolute number of TB suspects examined by smear microscopy annually has increased manifold, from 0.96 million to 7.8 million. Over the same time period, the rate of TB suspect examination also increased by 50%, from 421 per 100,000 population covered by RNTCP services to 651 per 100,000 population covered. Similarly, the rate of sputum smear positive cases diagnosed by microscopy has increased by 20%, from 62 to 79 per 100,000 population. The average number of suspects examined for every sputum smear positive case diagnosed has gradually increased about 1.3% per year, from 2001 to 2011, the number of suspects examined per smear positive case diagnosed has increased by 28% from 6.4 to 8.3 suspects. An average difference of 11.3% [Range 8– 15%] was observed between the rate of sputum-positive cases diagnosed and the sputum-positive case notification rate and so the predictions are made with reference to the findings.

**Figure 1: rate of TB suspect examined and smear positive TB cases diagnosed per 100,000 population**

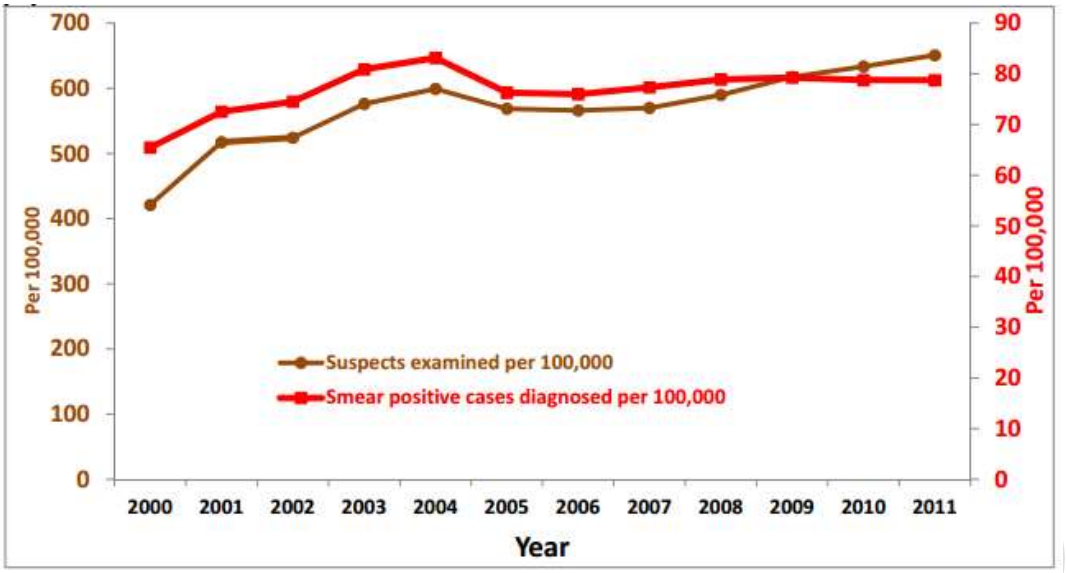
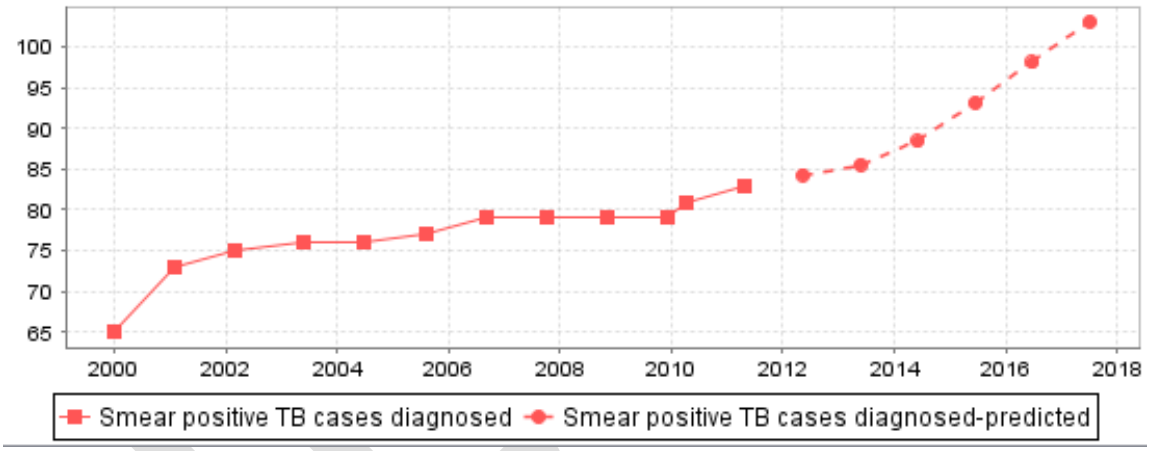



**Figure 2: future forecast for: Smear positive TB cases diagnosed**




## Notification Rates of TB Cases

Overall, case notification has increased over the 12 year analysis period, and the notification rates of most types of TB cases has steadily increased or remained stable, with the exceptions of new smear-negative and "treatment after default" . The total case notification rate has increased from 101 cases per 100,000 populations in 1999 to 125 per 100,000 population in 2011, though the last 4 years case notification has been effectively flat or rather decreasing. The NSP case notification rate has increased from 39 cases per 100,000 populations in 1999 to 53 per 100,000 populations in the year 2008, and has remained at 53/100,000 for the past 4 years. The NSN notification rates have shown a decreasing trend from 45 per 100,000 populations in 2004 to 28 per 100,000 population in 2011, and continues to fall without clear explanation. Some of the arguments for this are increased efforts to get the sputum examined and bacilli demonstrated with increasing availability and application of quality sputum smear microscopy services expanded under the programme. The notification rate of re-treatment cases has increased by 40% over the past 12 years, from 18 per 100,000 populations in 1999 to 25 per 100,000 populations in 2011. The increase in retreatment notification rates appears to be driven largely by increases in the notification rates of the 'relapse' and 'others' types of re-treatment cases. The 're-treatment others' notification rate has almost

doubled from 4 per 100,000 populations in 1999 to 8 per 100,000 populations in 2011. The notification rate of failure-type re-treatment cases has remained almost stable from 2002 onwards at the rate of 2 cases per 100,000 populations. The "Treatment after default" notification rates have declined from 10/100,000 population in 2001 to 6/100,000 population in 2011.

**Figure 3: Trends in type of TB case notification rate (199-2011)**

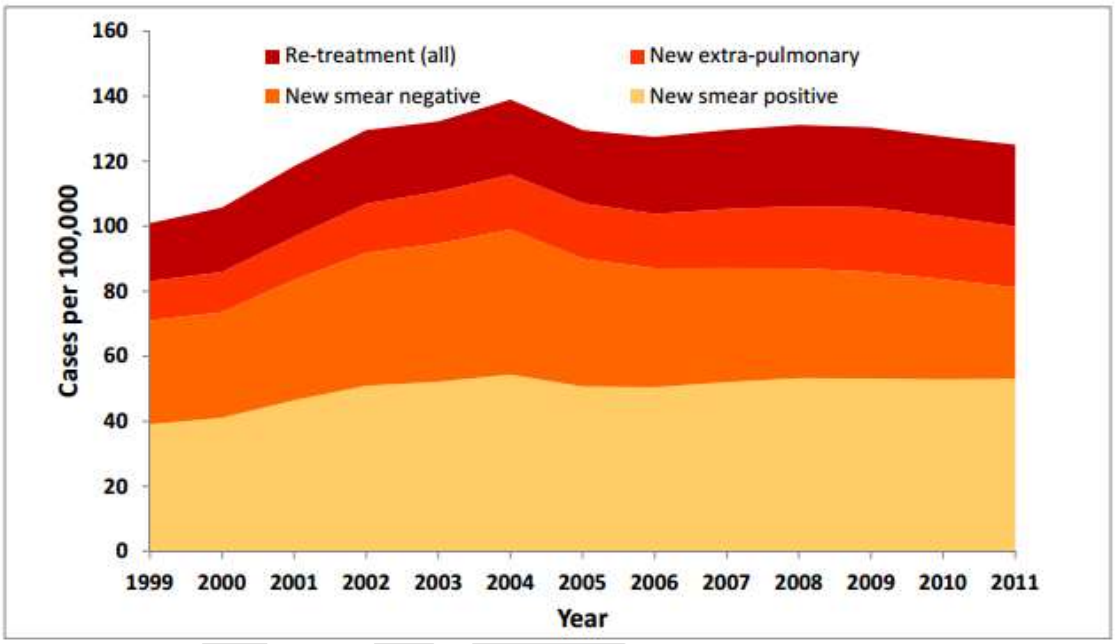
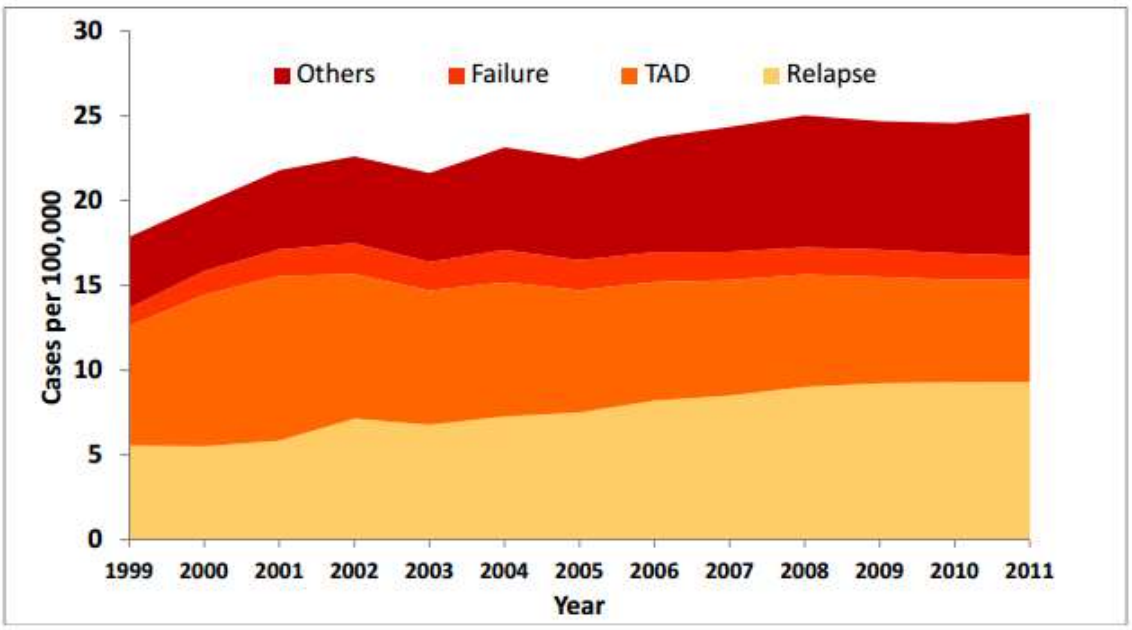



**Figure 4: Trends in type of re-treatment TB case notification rate (199-2011)**



## All New (incident) TB Case Notification

The number and rate of all new (incident) cases notified in the country has steadily increased at the rate of 7% annually for several years initially in the implementation of the programme starting from 83 per 100,000 population in 1999 to 116 per 100,000 population in 2004, with almost 40% increase in half a decade. The decline began after complete coverage in the country, and the all new (incident) TB case notification rate has decreased from 116 per 100,000 population in 2004 to 96 per 100,000 population in year 2012 showing a decline of 20%, almost 2% annually which is estimated to continue with the progressing years.

## Treatment Outcomes of Notified TB Cases

The treatment success rate has been > 85% since the year 2001. The death rate and failure rate has been about 5% and 2% respectively. The default rates has decreased from 9% for the cohort of TB patients registered in 1999 to 6% for the cohort of patients registered in 2010

**Table 1: Treatment outcomes among notified new TB cases, 1999–2010**

| | New smear positive | | | | New smear negative | | | | New Extra Pulmonary | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Success | Death | Failure | Default | Success | Death | Failure | Default | Success | Death | Failure | Default |
| 1999 | 82% | 5% | 3% | 9% | 85% | 4% | 1% | 9% | 91% | 2% | 0% | 6% |
| 2000 | 84% | 4% | 3% | 8% | 86% | 3% | 1% | 9% | 91% | 2% | 0% | 7% |
| 2001 | 85% | 5% | 3% | 7% | 86% | 4% | 1% | 8% | 91% | 2% | 0% | 6% |
| 2002 | 87% | 4% | 3% | 6% | 87% | 4% | 1% | 7% | 92% | 2% | 0% | 5% |
| 2003 | 86% | 5% | 2% | 6% | 87% | 4% | 1% | 7% | 92% | 2% | 0% | 5% |
| 2004 | 86% | 4% | 2% | 7% | 87% | 4% | 1% | 8% | 92% | 2% | 0% | 5% |
| 2005 | 86% | 5% | 2% | 7% | 87% | 4% | 1% | 8% | 91% | 2% | 0% | 6% |
| 2006 | 86% | 5% | 2% | 6% | 87% | 4% | 1% | 8% | 90% | 3% | 0% | 5% |
| 2007 | 87% | 5% | 2% | 6% | 87% | 3% | 1% | 8% | 91% | 2% | 0% | 5% |
| 2008 | 87% | 4% | 2% | 6% | 88% | 3% | 1% | 7% | 92% | 3% | 0% | 4% |
| 2009 | 87% | 4% | 2% | 6% | 88% | 3% | 1% | 7% | 92% | 2% | 0% | 4% |
| 2010 | 88% | 4% | 2% | 6% | 89% | 3% | 1% | 7% | 93% | 3% | 0% | 4% |

As a result of rapid expansion in diagnostic facilities, the proportion of sputum- positive cases confirmed in the laboratory have doubled than that of the previous programme and is on par with international standards. Despite the rapid expansion, overall performance remains good and in many areas is excellent. Treatment success rates have tripled from 25%in the earlier programme to 86%in RNTCP. By the year 2025, a success rate of 99.35% is forecasted with a confidence level of 95% using regression algorithm. Weka tool is used for the purpose of prediction.

**Figure 5: Treatment Success Rate**

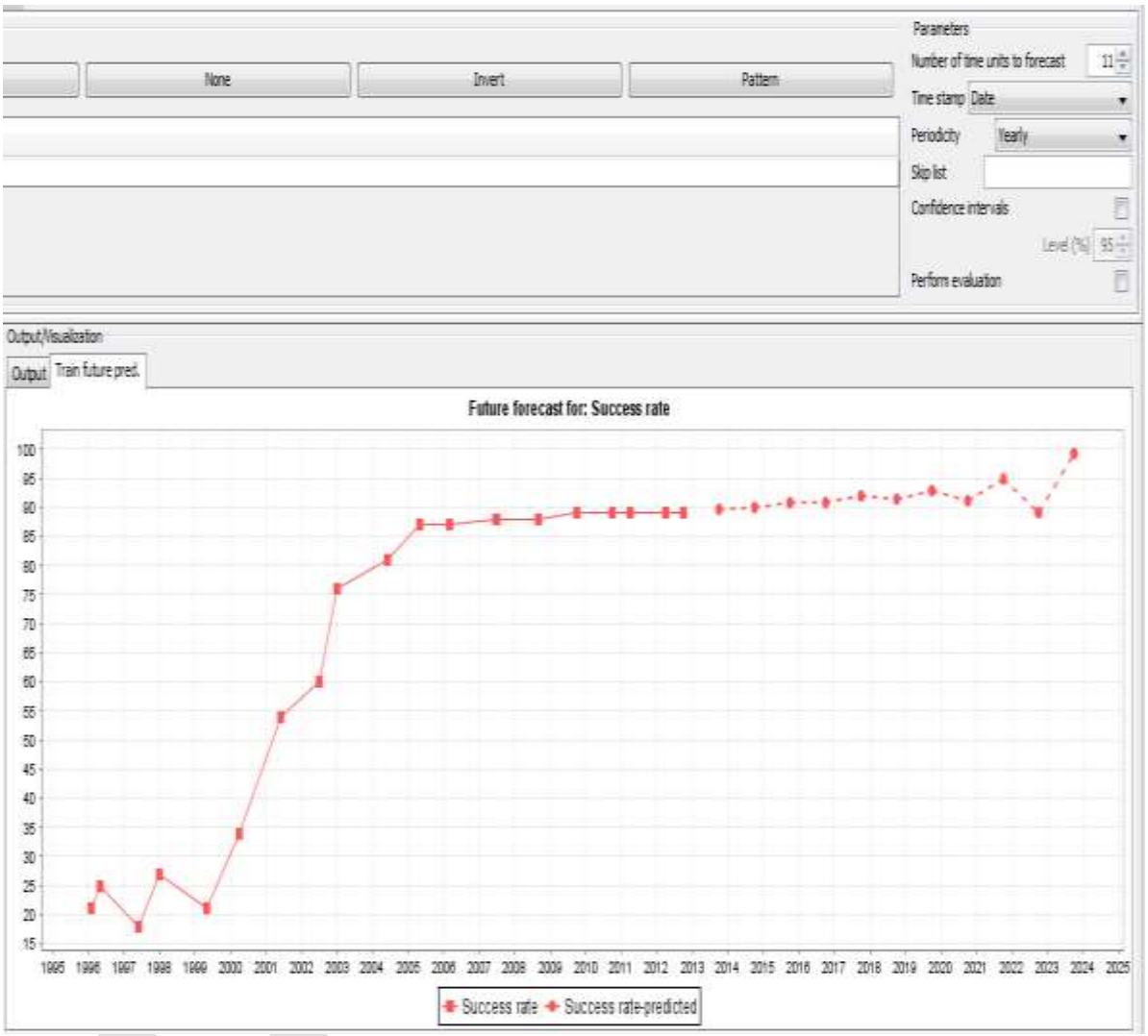



**Figure 6: Reporting progress towards set targets**

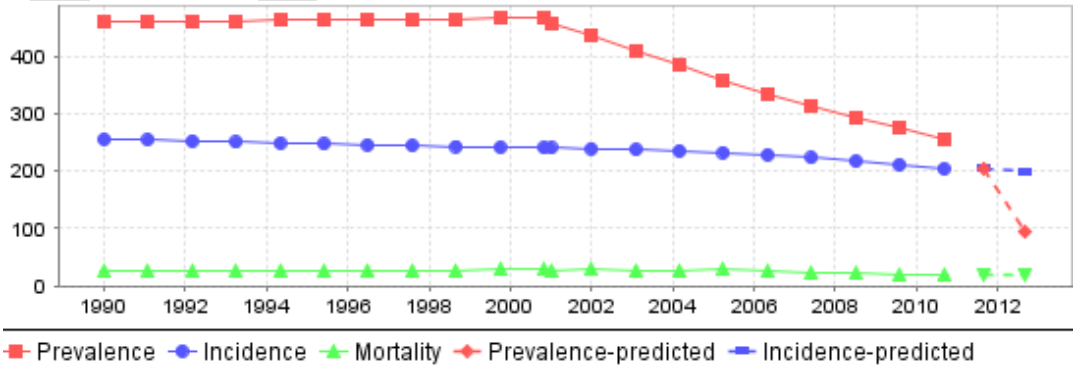
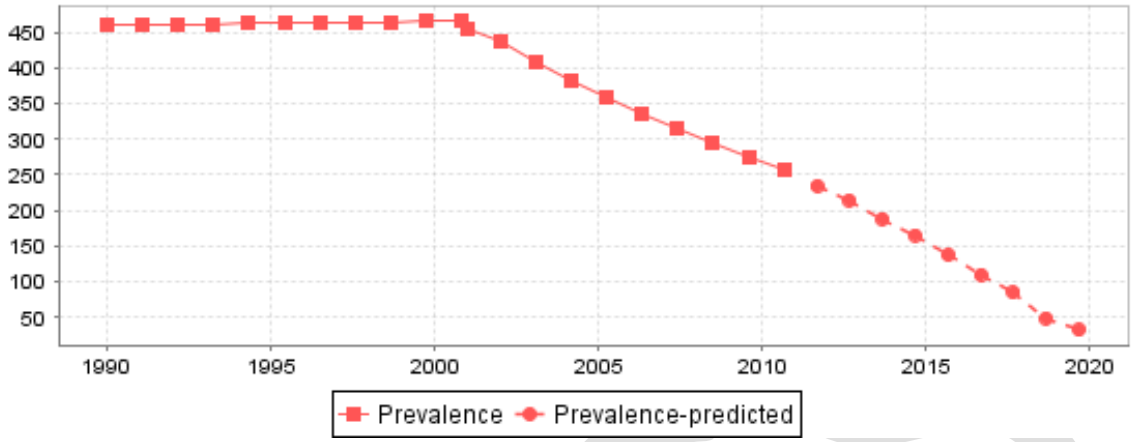


**Figure 7: Future forecast for: Prevalence**




Regarding the target of halving the mortality rate compared to the 1990 baseline, the Region had reached the target in 2013. In fact, considering only the best estimate, in 2013, the mortality rate decreased by 53%; according to the projections based on the assumption that the current trend will not change, the Region would sustain the achievement and even the upper uncertainly bound is expected to be almost entirely below the target.

## CONCLUSION

RNTCP is performing well in terms of reduction of TB burden in India. Analysis of progress with regard to tuberculosis control shows that the Region has achieved or is well on track to halt and begin to reverse the incidence of tuberculosis by 2015, and halve the TB death and prevalence rates by 2015, compared with 1990 levels. By the end of 2020, the prevalence of the disease is forcasted to tend towards zero. Regarding the targets of halving the prevalence rates compared to the 1990 baseline, the Region is on track to reach the targets. In fact, considering only the best estimate, in 2013, the prevalence rate decreased by 47%; according to the projections based on the assumption that the current trend will not change, the Region would reach 50% reduction of baseline data. However, almost the entire upper uncertainity bound would be over the target; more accurate estimates resulting from completed or planned prevalence surveys will be useful to confirm achievements in the Region beyond any doubt.

**Appendix**

**Smear positive TB cases diagnosed**

=== Run information ===

Scheme:

LinearRegression -S 0 -R 1.0E-8

Lagged and derived variable options:

-F "[Smear positive TB cases diagnosed]" -L 1 -M 6 -G Date

Relation: wine2

Instances: 12

Attributes: 3

Rate of TB suspects examined

Smear positive TB cases diagnosed

Date

Transformed training data:

Smear positive TB cases diagnosed

Date-remapped

Lag_Smear positive TB cases diagnosed-1

Lag_Smear positive TB cases diagnosed-2

Lag_Smear positive TB cases diagnosed-3

Lag_Smear positive TB cases diagnosed-4

Lag_Smear positive TB cases diagnosed-5

Lag_Smear positive TB cases diagnosed-6

Date-remapped^2

Date-remapped^3

Date-remapped*Lag_Smear positive TB cases diagnosed-1

Date-remapped*Lag_Smear positive TB cases diagnosed-2

Date-remapped*Lag_Smear positive TB cases diagnosed-3

Date-remapped*Lag_Smear positive TB cases diagnosed-4

Date-remapped*Lag_Smear positive TB cases diagnosed-5

Date-remapped*Lag_Smear positive TB cases diagnosed-6

Smear positive TB cases diagnosed:

Linear Regression Model

Smear positive TB cases diagnosed =

3.8772 * Date-remapped +
-0.9146 * Lag_Smear positive TB cases diagnosed-3 +
0.9833 * Lag_Smear positive TB cases diagnosed-4 +
0.5973 * Lag_Smear positive TB cases diagnosed-5 +
-0.4175 * Lag_Smear positive TB cases diagnosed-6 +
-0.4089 * Date-remapped^2 +
0.0202 * Date-remapped^3 +
-0.0102 * Date-remapped*Lag_Smear positive TB cases diagnosed-1 +
0.0291 * Date-remapped*Lag_Smear positive TB cases diagnosed-3 +
-0.0422 * Date-remapped*Lag_Smear positive TB cases diagnosed-4 +
-0.0079 * Date-remapped*Lag_Smear positive TB cases diagnosed-5 +
0.0233 * Date-remapped*Lag_Smear positive TB cases diagnosed-6 +

50.041

=== Future predictions from end of training data ===

| Time | Smear positive TB cases diagnosed |
|---|---|
| 2000-01-01 | 65 |
| 2001-01-11 | 73 |
| 2002-01-22 | 75 |
| 2003-02-02 | 76 |
| 2004-02-13 | 76 |
| 2005-02-23 | 77 |
| 2006-03-07 | 79 |
| 2007-03-18 | 79 |
| 2008-03-28 | 79 |
| 2009-04-08 | 79 |
| 2010-04-19 | 81 |
| 2011-04-30 | 83 |
| 2012-05-11* | 84.2948 |
| 2013-05-22* | 85.5301 |
| 2014-06-02* | 88.4084 |
| 2015-06-13* | 93.1353 |
| 2016-06-23* | 98.1013 |
| 2017-07-05* | 103.0451 |

**Trends in suspects examined per smear positive TB case diagnosed**

=== Run information ===

Scheme:
LinearRegression -S 0 -R 1.0E-8

Lagged and derived variable options:
-F [trend] -L 1 -M 6 -G Date

Relation: wine2
Instances: 12
Attributes: 2
trend
Date

Transformed training data:

trend
Date-remapped
Lag_trend-1
Lag_trend-2
Lag_trend-3
Lag_trend-4
Lag_trend-5
Lag_trend-6
Date-remapped^2
Date-remapped^3
Date-remapped*Lag_trend-1
Date-remapped*Lag_trend-2
Date-remapped*Lag_trend-3
Date-remapped*Lag_trend-4
Date-remapped*Lag_trend-5
Date-remapped*Lag_trend-6

trend:

Linear Regression Model

trend =

0.0385 * Date-remapped +
-0.0068 * Lag_trend-2 +
-0.746 * Lag_trend-3 +
0.7686 * Lag_trend-4 +
-0.2179 * Lag_trend-5 +
-0.3186 * Lag_trend-6 +
0.0064 * Date-remapped^2 +
0.0012 * Date-remapped^3 +
-0.0144 * Date-remapped*Lag_trend-1 +
0.0014 * Date-remapped*Lag_trend-2 +
0.0131 * Date-remapped*Lag_trend-3 +
-0.0332 * Date-remapped*Lag_trend-4 +
0.0097 * Date-remapped*Lag_trend-5 +
11.3399

=== Future predictions from end of training data ===

| Time | trend |
|---|---|
| 2000-01-01 | 6.48 |
| 2001-01-30 | 7.08 |
| 2002-03-02 | 6.99 |
| 2003-04-01 | 7.11 |
| 2004-05-01 | 7.22 |
| 2005-06-01 | 7.49 |
| 2006-07-01 | 7.44 |
| 2007-08-01 | 7.4 |
| 2008-08-31 | 7.47 |
| 2009-09-30 | 7.81 |
| 2010-10-31 | 8.01 |
| 2011-11-30 | 8.24 |
| 2012-12-30* | 8.5391 |
| 2014-01-30* | 9.0823 |
| 2015-03-01* | 9.5447 |
| 2016-03-31* | 10.0223 |
| 2017-05-01* | 10.5312 |

**Forecast for prevalence**

=== Run information ===

Scheme:

LinearRegression -S 0 -R 1.0E-8

Lagged and derived variable options:

-F [Prevalence] -L 1 -M 5 -G Year

Relation: wine2

Instances: 21

Attributes: 4

Prevalence

Incidence

Mortality

Year

Transformed training data:

Prevalence

Year-remapped

Lag_Prevalence-1
Lag_Prevalence-2
Lag_Prevalence-3
Lag_Prevalence-4
Lag_Prevalence-5
Year-remapped^2
Year-remapped^3
Year-remapped*Lag_Prevalence-1
Year-remapped*Lag_Prevalence-2
Year-remapped*Lag_Prevalence-3
Year-remapped*Lag_Prevalence-4
Year-remapped*Lag_Prevalence-5

Prevalence:

Linear Regression Model

Prevalence =

-5.0562 * Year-remapped +
1.5441 * Lag_Prevalence-2 +
-1.8314 * Lag_Prevalence-4 +
1.0526 * Lag_Prevalence-5 +
0.0124 * Year-remapped*Lag_Prevalence-2 +
-0.0107 * Year-remapped*Lag_Prevalence-4 +
141.305

=== Future predictions from end of training data ===

| Time | Prevalence |
|---|---|
| 1990-01-01 | 459 |
| 1991-01-01 | 460 |
| 1992-01-01 | 460 |
| 1993-01-01 | 461 |
| 1994-01-01 | 462 |
| 1995-01-01 | 462 |
| 1996-01-01 | 463 |
| 1997-01-01 | 464 |
| 1998-01-01 | 464 |
| 1999-01-01 | 465 |
| 2000-01-01 | 466 |

| | |
|---|---|
| 2001-01-01 | 456 |
| 2002-01-01 | 436 |
| 2003-01-01 | 409 |
| 2004-01-01 | 383 |
| 2005-01-01 | 358 |
| 2006-01-01 | 335 |
| 2007-01-01 | 314 |
| 2008-01-01 | 294 |
| 2009-01-01 | 275 |
| 2010-01-01 | 256 |
| 2011-01-01* | 234.6434 |
| 2012-01-01* | 214.3271 |
| 2013-01-01* | 188.6203 |
| 2014-01-01* | 165.769 |
| 2015-01-01* | 137.7494 |
| 2016-01-01* | 110.2066 |
| 2017-01-01* | 85.3645 |
| 2018-01-01* | 49.5028 |
| 2019-01-01* | 32.7302 |

**REFERENCES:**

[1] Udwadia, Z "This should not exist", Hindustan Times, November 27, 2012 www.hindustantimes.com/ - See more at: http://www.tbfacts.org/tb-india/#sthash.llxxjViE.dpuf
[2] Pai, M "Formidable killer: drug-resistant tuberculosis", The Tribune, India, August 6, 2013www.tribuneindia.com/2013/ - See more at: http://www.tbfacts.org/tb-india/#sthash.llxxjViE.dpuf
[3] Managing the Revised National Tuberculosis Control Programme in Your Area – A training course, E1- Exercise workbook, Central TB Division Directorate General of Health Services, Ministry of Health and Family Welfare Nirman Bhavan, New Delhi 110011.
[4] Predicting School Ranks Through Data Mining Micah Oppenheim—Boston University CS 105—Professor David G. Sullivan, Ph.D
[5] A Data Mining Perspective on the Prevalence of Polio in India Neera Singh Indian Institute of Information Technology, Allahabad, India, Neera Singh et al. / International Journal on Computer Science and Engineering (IJCSE).
[6] http://infochangeindia.org/public-health/books-a-reports/eradicating-tuberculosis-the-unfinished-agenda.html
[7] Srivastava, K, "TB epidemic looms large with Rs 2,000 crore fund cut, erred policy", dna, 10 January, 2015 www.dnaindia.com/ - See more at: http://www.tbfacts.org/tb-india/#sthash.llxxjViE.dpuf
[8] "Private hospitals report 700 TB cases", The Times of India, August 19, 2013http://articles.timesofindia.indiatimes.com/2013-08-19/pune/ - See more at: http://www.tbfacts.org/tb-india/#sthash.llxxjViE.dpuf
[9] Bhalchandra Chorghade "To fight MDR-TB, act on time", http://dnasyndication.com/dna/MUMBAI/ - See more at: http://www.tbfacts.org/tb-india/#sthash.llxxjViE.dpuf
[10] Tuberculosis control in the South-East Asia Region Annual TB report 2015 World Health Organization Regional Office for south east Asia - http://www.searo.who.int/tb/annual-tb-report-2015.pdf.
[11] Government of India TB India 2013 Revised National TB Control Programme Annual Status Report - http://www.tbcindia.nic.in/pdfs/tb%20india%202013.pdf
[12] Sinha, K "Finally, tuberculosis declared a notifiable disease", The Times of India, May 9, 2012 //articles.timesofindia.indiatimes.com/2012-05-09/india/31640562_1_mdr-tb-tb-cases