Mortality Analysis in Bihar (India) by analysing the deaths from year 2007 to 2011

Sandeep Nannamu-0882000, Sanhith Reddy-0889854, Shreya Dasari-0888189, Sai Deepika Surabhi-0884203

Abstract— India contributes the highest global share of deaths. Continuous monitoring of the reduction in the mortality rate at local level is thus essential to set priorities for policy-makers and health professionals. Although death mortality have progressively declined in most countries during the first 90 years of this century, there are still substantial differences between these countries with respect to these important health parameters. It provides relevant information on declines in mortality. In this study, we aim to provide an update on district-level disparities on preventable deaths and it will not only show trends in these health measures in Bihar, but also discuss relevant family characteristics and summarize the social support system, including health care services. A lot of analaysis on the mortality rate with respect to the patientâĂŹs symptoms can be done by collecting data regarding various deaths which is easier if the deaths are registered. We have anaylysed a survey level dataset containing information regarding deaths in Bihar from 2007 to 2011, we go through the various related satistics and build a prediction in the end regarding whether the deaths have been registered or not.

I. INTRODUCTION

Majority of the scientific studies in biomedical and health-care spaces ad- dress issues whose final objective is to anticipate or prevent death and dis- eased. Since the rise of the big data science, various machine learning based methods and advances have been proposed and applied to improve human well being by figuring out the solutions to various computational challenges that we confront nowadays. A less self-evident issue, that remains to be broadly investigated by analysts, is whether Big Data science can contribute to our understanding of components driving to death or infections, through examination of multiple-cause mortality information. In reality, it is broadly accepted that tallying the dead may be a critical speculation to diminish the premature mortality. There have been a few studies that have demonstrated to offer significant impacts on our understanding of the major causes of death utilizing the measurable examination of recorded death information. Considering these studies, we were interested to explore the achievability of this developing field in learning hidden complex designs. Whereas the sheer sum of accessible information collected from the enlisted death certificates makes it agreeable for big data analysis, it poses a few key challenges at the same time. In specific, the multiple-cause of-death information are unstructured and are regularly inaccurate and noisy. Additionally, the high number of mortality codes makes examination of multiple-cause affiliations indeed more challenging. Through and through, these call for advanced procedures for mining huge datasets. In this project we have collected the information relating to the deaths caused in Bihar(INDIA) from the year 2007 to 2011. The aim is to analyze the various attributes, go through all the statistics and and build prediction model regarding registration of deaths in the state of BIHAR.

II. RELATED WORK

A. Summary

The paper composed by Chandan Kumar(Professor in IIT Roorkee), Prashant Kumar Singh (Teacher in IIPS) Rajesh Kumar Rai(Professor in TISS) examines that when controlling for biophysical and geographical variables (counting precipitation, efficiency of agrarian lands, topography/temperature, and advertise get to through street systems), financial and wellbeing care markers offer assistance to clarify varieties of under-five mortality rate over areas from nine high focus states in India. The writing on this subject is uncertain since the overview information, upon which most studies of child mortality depend, once in a while incorporate factors that degree these components. This paper presents these factors into an examination of 284 areas from nine center states in India. India has the worldâĂŹs most elevated rate (21 percent) of under-five deaths, evaluated at 1726000 in 2009. The nation overseen to diminish the Under-Five Mortality Rate (U5MR) from 118 per 1000 live births in 1990 to 66 per 1000 live births in 2009. The average annual rate of decline at 3.1 percent was considered insufficient to achieve Millennium Development Goal (MDG) 4 that targets minimizing under-five mortality to 39 per 1000 live births by 2015 . The north-south variation in child mortality in India is reflected in literature , where some of the north Indian states such as Rajasthan, Uttar Pradesh, Bihar, Orissa, Chhattisgarh and Madhya Pradesh persistently performed poorly in health care , because of the unacceptably high fertility and mortality indicators

The Eight Empowered Action Group (EAG) states (Bihar, Chhattisgarh, Jharkhand, Madhya Pradesh, Orissa, Rajasthan, Uttarakhand, Uttar Pradesh and Assam), which account for approximately 48 percent of IndiaâĂŹs populace, are assigned as High Center States by the Government of India. The U5MR in Uttar Pradesh (94 per 1000 live births), Madhya Pradesh (89 per 1000 live births), Orissa (82 per 1000 live births), Assam and Bihar (77 and 78 per 1000 live births) are nearly comparative to the U5MR in a few African nations Djibouti (94 per 1000 live births), Zimbabwe (90 per 1000 live births), Kenya (84 per 1000 live births), Sao Tome and Principe (78 per1000 live births) respectively, based on the area level U5MR that has been made accessible as of late by the Yearly Wellbeing Overview (AHS) 2011, we evaluate the levels of under-five mortality and its spatial design in these nine states in India. Utilizing Exploratory Spatial Information Examination (ESDA) and spatial econometric methods, this paper examines if, when controlling for biophysical and geographical variables, socioeconomic and health programs related indicators help to explain variation in U5MR across 284 districts in 9 high focus group states. They also intend to identify some of the critical districts with high under-five mortality in order to prioritize the implementation of program initiatives.

The paper on Trend analysis of mortality rates and causes of death in children under 5 years old in Beijing, China from 1992 to 2015 analyzed break down patterns in mortality and reasons for death among kids under 5 years in Beijing, China somewhere in the range of 1992 and 2015 and to figure out Under-5 Death Mortality Rates (U5MRs) for the period 2016 2020. The paper uncovered the long patterns and attributes in death rates and driving reasons for death in less than 5 year youngsters through a 24-year examination. The paper was about the wellbeing states of kids matured under 5 years following the usage of the âĂŹSelective Two-Child PolicyâĂŹ in Beijing in 2014. The exactness of the information is solid in light of the fact that the example involves the whole population of all under-5 children in all areas of Beijing dependent on the observation organize. Inferable from the one of a kind territorial divisions in Beijing and absence of particular information on each region, the paper didnâĂŹt give the contrasts among provincial and urban regions. The autoregressive incorporated moving normal (ARIMA) (1,1,1) demonstrate estimated future U5MRs dependably for just a brief timeframe, on the grounds that the model just considered varieties in mortality with time rather than other conceivable affecting components, and information must be constantly refreshed to foresee promote death rates. Different models which can anticipate U5MRs all the more precisely and catch its examples all the more particularly should be investigated. Since they just gathered 2-year information (2014 and 2015) after the âĂŹParticular Two-tyke PolicyâĂŹ, the representativeness of the changing pattern and strong measurable examination were constrained. The connection between the changing pattern in U5MRs and arrangement control is as yet misty and should be watched for further years.

According to the paper written by Richard Doll, emeritus professor of medicine, Richard Peto, professor of medical statistics and epidemiology, Jillian Boreham, senior research fellow and Isabelle Sutherland, research assistant on Mortality in relation to smoking: 50 yearsâĂŹ observations on male British doctors is to analyze the hazards of cigarette smoking in men who shaped their propensities at various periods, and the degree of the decrease in hazard when cigarette smoking is ceased at various ages. The abundance mortality related to smoking primarily included vascular, neoplastic, and respiratory illnesses that can be caused by smoking. Men conceived in 1900-1930 who smoked just cigarettes and kept smoking kicked the bucket overall around 10 years more youthful than long lasting non-smokers. End at age60, 50, 40, or 30 years picked up, individually, around 3, 6, 9, or 10 years of future. The abundance mortality related with cigarette smoking was less for men conceived in the nineteenth century and was most noteworthy for men conceived during the 1920s. The cigarette smoker versus non-smoker probabilities of passing on in middle age (35-69) were 42 percent 24 percent (a twofold demise rate proportion) for those conceived in 1900-1909, yet were 43 percent 15 percent (a triple demise rate proportion) for those conceived during the 1920s. At more seasoned ages, the cigarette smoker versus non-smoker probabilities of making due from age 70 to 90 were 10 percent 12 percent at the demise rates of the 1950s (that is, among men conceived around the 1870s) however were 7 percent 33 percent (again a triple passing rate proportion) at the demise rates of the 1990s (that is, among men conceived around the 1910s).This examination reported the quantity of passings amiable to medicinal services in LMICs and is the first to gauge the extent of these passings because of low quality of consideration versus non-use of consideration. This finding has imperative approach suggestions for nations seeking after all inclusive wellbeing inclusion as expanded access to low quality of consideration is probably not going to enhance wellbeing results. Our examination found that about 8 million individuals bite the dust each year in view of an absence of access to superb consideration. We found a higher extent of amiable passings are among wellbeing framework clients than non-clients in LMICs. Passings caused by low quality human services crossed the conditions incorporated into the Sustainable Development Goals, including cardiovascular ailments, neonatal conditions and street auto collisions. In spite of the fact that the 2016 GBD think about did not report quantities of manageable passings or segment these passings into the different commitments of nature of consideration and usage, it observed generous inconsistencies in agreeable mortality crosswise over districts and identified with levels of advancement. The proof of low quality social insurance challenges the supposition that expanding usage of wellbeing administrations will be adequate to decrease mortality in lower-pay nations. Be that as it may, to date, there have not been any investigations evaluating the potential job of better-quality administrations versus more prominent inclusion in decreasing mortality for conditions agreeable to restorative consideration. This report will assess the overabundance passings agreeable to social insurance in LMICs and the overall commitments of non-usage of human services administrations and receipt of low quality consideration to these passings.

According to the paper written by Craig S Knott, research associate, Ngaire Coombs, research associate, Emmanuel Stamatakis, associate professor, Jane P Biddulph, lecturer on All cause mortality and the case for age specific alcohol consumption guidelines: pooled analyses of up to 10 population based cohorts is to examine the reasonableness of age particular breaking points for alcohol consumption and to investigate the relationship between alcohol consumption and mortality in various age groups. This paper uses two proportions of alcohol use trying to catch distinctive drinking practices related with the danger of all reason mortality:

normal week by week utilization and use on the heaviest savoring day the prior week meet. Results from these two measures were observed to be correlative, both showing that defensive relationship between alcohol consumption and mortality were to a great extent particular to ladies matured 65 years or more (table 4). Limiting information to years for which both introduction factors were accessible (1999-2002), relationships between the factors barring never and previous drinking classes were solid inside age-sex strata (r=0.57 to 0.65, PÂą0.001).This paper uses two proportions of alcohol use trying to catch distinctive drinking practices related with the danger of all reason mortality: normal week by week utilization and use on the heaviest savoring day the prior week meet. Results from these two measures were observed to be correlative, both show- ing that defensive relationship between alcohol consumption and mortality were to a great extent particular to ladies matured 65 years or more. Limiting information to years for which both introduction factors were accessible (1999-2002), relationships betweenâĂŹs the factors barring never and previous drinking classes were solid inside age-sex strata (r=0.57 to 0.65, PÂą0.001).

According to paper Increased mortality related to heavy alcohol intake pattern written by Dr T Laatikainen, this study talks about how consumptions of large amount of alcohol will be harmful even though mild alcohol intake is related to decreased all cause and ischaemic heart disease. Their biggest strength was to find that mortality among men drinking six or more bottles of beer has been has been found to be higher than among those restricting their intake to less than three bottles of beer per occasion. They dint take the beverage specific analysis into consideration as in their statistics depicted that people consuming different types of drinks and people consuming more than one type of drink had a higher risk of heart disease.

According to paper The effect of Peer Review on Mortality Rates written by Friedrich-Naumann, this study talks about how mortality rate scan be lowered in hospitals. The main finding of our study is that a comprehensive quality management system, including a mandatory process of âĂŸpeer reviewingâĂŹ, was associated with a significant decrease in mortality rates when prompted by specific quality indica- tors. The peer review process and detailed action plans for improvement, which were triggered by higher than expected mortality rates, decreased the observed mortality rates down to expected rates after the peer review. The presented data cannot rigorously prove that the process of peer reviewing decreases mortality rates. Rather, the results suggest that the quality management system including use of a defined set of indicators that are transparently reported and used as a mandatory trigger for peer reviewing is effective at improving mortality rates that are worse than expected. A further limitation is the lack of data from extended time periods pre and post-peer review. Most of the hospitals with suboptimal mortality rates were recently acquired, and did not have mortality rates from several years prior to the review process.

According to paper Chronic Disease Mortality in a Cohort of Smokeless Tobacco Users written by Howard G, this study talks about hazard ratios from several specific chronic diseases. Male exclusive smokeless tobacco users did not experience significant increases in mortality for any type of cancer considered. The increased mortality from lung cancer among female smokeless tobacco users (never or ever smokers), although statistically significant, was based on only three deaths and four deaths, respectively. Smokeless tobacco use was not associated with significant increases in mortality for ischemic heart disease or stroke in either gender. The analyses investigating the combined effects of smokeless tobacco use and smoking on specific outcomes were restricted to male subjects and not females. Another limitation of this analysis is that the exposure category is based only on ever use of smokeless tobacco. Therefore, potential increases in mortality associated with current versus former use could not be determined

According to paper Premature Death Among Primary Care Patients With a History of Self-Harm written by Mathew J. Carr, this data examines premature mortality in a nationally representative cohort of primary care patients who had harmed themselves. They examined risks of all-cause and cause-specific premature death in a nationally representative primary care cohort, with complete case ascertainment via linkage to national mortality records. The utilized an ideal examination configuration by looking at dangers specifically at the individual patient dimension between an occurrence self-hurt accomplice and an unaffected correlation companion inspected from a similar populace. This is a more robust methodology than looking in danger in a roundabout way by means of age-and sex-institutionalized mortality proportions computed utilizing broadly amassed information, as was accounted for in past examinations. By portraying an occurrence partner structure, blocked predominant companion inclination which belittles the quality of presentation result affiliations, and which could have affected past examinations of this theme. At last, the structure was additionally improved by having up to 20 coordinated correlation people for everyone in oneself damage companion to empower examination of mortality results that are especially uncommon in the overall public, for example, suicide. Major limitations include lacking the ability to examine confounding or effect modification by ethnicity and individual-level socioeconomic status (beyond a score allocated at the patient postcode level) and mortality record linkage scheme may not be generalizable to the entire UK population

According to the paper The relation between different dimensions of alcohol consumption and burden of disease: an overview written by JÃijrgen Rehm, Dolly Baliunas, Guilherme L. G. Borges, Kathryn Graham, Hyacinth Irving, Tara Kehoe, Charles D. Parry, Jayadeep Patra, Svetlana Popova, Vladimir Poznyak, Michael Roerecke, Robin Room, Andriy V. Samokhvalov, Benjamin Taylor, this study talks about the estimation related to the global burden of disease and injury caused by alcohol to evaluate the evidence for a causal impact of average amount of consumption of alcohol and pattern of drinking on diseases and injuries. Their biggest strength was to find the impact of alcohol consumption for specifically these diseases: tuberculosis, mouth, nasopharynx, other pharynx and oropharynx cancer, oesophageal cancer, colon and rectum cancer, liver cancer, female breast cancer, diabetes mellitus, alcohol use disorders, unipolar depressive disorders, epilepsy, hypertensive heart disease, ischaemic heart disease (IHD), ischaemic and haemorrhagic stroke, conduction disorders and other dysrhythmias, lower respiratory infections (pneumonia), cirrhosis of the liver, preterm birth complications and fetal alcohol syndrome. They failed to show how the depressive disorders made an impact on alcohol consumption. Our study does not include as many diseases but has few depressing disorders. This study would have been impeccable if they could have shown the differential role of drinking pattern and the amount of alcohol consumed which will be helpful in estimating a more appropriate alcohol and disease correspondence.

According to the paper 50-Year Trends in Smoking-Related Mortality in the United States written by MJ Thun, this study talks about the infectious dangers from cigarette smoking expanded in the United States over a large portion of intervals till the twentieth century, first among male smokers and later among female smokers. They estimated the changes in mortality across the following periods 1959âĂŞ1965, 1982âĂŞ1988, and 2000âĂŞ2010 among individuals aged 50 years or older comparing based upon gender, self-reported smoking status, and previous literature survey. The strengths of their study include its size, prospective design, national scope, and 50-year time span. The results show estimates of temporal changes in cause-specific mortality and the contemporary risks from smoking in the United States. A disadvantage was the study only focuses on whites, 50 years of age or older, who were born between 1870 and 1954. They could not assess risks among younger contemporary smokers. Most current smokers in the modern regions had smoked for at least 30 years, limiting the range over which they could alter the influence of the duration of smoking. Our study concentrates not only on people aged 50 or older but also on younger aged individuals.

According to the paper Work-related stress in midlife and all-cause mortality: can sense of coherence modify this association written by Charlotta Nilsen, Ross Andel, Johan Fritzell and Ingemar KÃěreholt, this study examines three points. One, the relationship between work-related stress in midlife and mortality. Two, if the midlife sense of coherence altered any observed relationships. Three, if the results for men and women are similar or not. The main considerable distinction they found was that in ladies, self-announced jobs were related with less mortality, while in men, the mixture of self-detailed inactive occupations and a powerless SOC was related with expanded mortality. Maybe men's character is bound to be related to their activity, with passive jobs negatively affecting mental prosperity, which may thus unfavorably influence mortality. This study did not take the health of the individual into consideration as health plays a major factor in a work-stress area. The sense of coherence was estimated based on upon fewer parameters. Chronic exposure, i.e., whether the person is working for an extended period or short period will have an impact on him. Our work relates the occupational status of various other attributes like illness type, disability status which will generate a better result.

According to the paper Job strain among blue-collar and white-collar employees as a determinant of total mortality: a 28-year population-based follow-up written by Dr. Mikaela von Bonsdorff, this paper examined the effects of job demand, job control and job strain on total mortality among middle-aged white and blue-collar public-sector employees in a representative cohort with a 28-year follow-up. As per the model described by Karasek, the job strain was categorized to low job strain (low job demand and high job control), high job strain (high job demand and low job control) and passive job strain (low job demand and low job control). An advantage was illustrative with the huge example of open part representatives working both in salaried and hands-on callings and the long follow-up time on mortality gathered from the national mortality enlist. Also, by categorizing job strain, appropriate levels of strain were estimated. A limitation is the self-reported job strain. However, high correlations between subjective and expert ratings on work conditions have been reported. The assessment of job strain was measured only once at a point in midlife which may not entirely reflect long-term job strain; however, the municipal employees in the cohort had constant work histories indicating stability, probably also for job strain because of working early in their life.

According to the paper Incidence of Adverse Events and Negligence in Hospitalized Patients âĂŤ Results of the Harvard Medical Practice Study I written by Troyen A. Brennan, Lucian L. Leape, Nan M. Laird, Liesi Hebert, A. Russell Localio, Ann G. Lawthers, Joseph P. Newhouse, Paul C. Weiler and Howard H. Hiatt, this study estimated the occurrence of adverse events determined as damage caused by medical management and injuries which were resulted due to the hospitals negligence and low quality care. They took their data from 51 hospitals selected randomly in New York. Their study found that most of the adverse events occurred in the disability department due to non-responsive action by nurses. However, their paper did not consider the time factor. They stated many risk factors and occurrence of adverse events to be true in most the case but considering their gathering of data which is limited, their prediction may or may not be fitting.

According to the paper Assessment of Tobacco Consumption and Control in India written by Priya Mohan, Harry A Lando, Sigamani Panneer, the point of this article was to integrate the accessible logical information on tobacco use in India, with a view to evaluating the extent of the issue, surveying the tobacco control enactment and its effect at the miniaturized scale and large scale dimension of tobacco control in India. The requirement for this exhaustive assessment was to build up a superior comprehension of tobacco âĂŞ utilization design, control arrangements, and the holes that should be tended to will fill in as a reference for creating pragmatic control approach. The data was gathered from National Family Health Surveys (NFHS) that is from one of the households in both rural and urban areas. They found that Beedis, a tobacco product which has more nicotine and tar than a regular cigarette was smoked by the lower working class. They also evaluated how passive smoking affects people of different ages. Among children may cause, middle ear and respiratory infections and exposure at an extreme level may cause sudden death. For pregnant women, might harm the child the motherâĂŹs womb. In the case of the elderly population, the probability of occurrence of the diseases: hearing impairment, cataract, dementia is high. A limitation to this study was asking the household instead of the participant. Their research was focused more on rural areas. Our study emphasizes urban areas as well, concluding based on various factors like health record, occupation status and age group.

According to the paper The State of US Health, 1990-2016, Burden of Diseases, Injuries, and Risk Factors Among US States written by AH Mokdad, this study reports progression in the burden of diseases, injuries and risk factors at the state level during 1990 to 2016 of Global Burden of Disease Study. Their analysis indicates patterns in various age groups and shows that enhancement in some health results, for example, ischemic heart disease, lung malignancy, and neonatal preterm entanglements, are adjusted by high death rates from drug use disorders, chronic pulmonary disease, self-harm, chronic kidney disease, cirrhosis, and hypertensive heart disease. Outline measures, for example, future, that don't separate the patterns in various age bunches veil the heterogeneous headings for US health status by age and state. Well beyond the drivers of disparate patterns, the study uncovers that there have been far more noteworthy advancement in decreasing the weight of some significant reasons for years of life lost, for example, ischemic heart disease and lung malignancy, however no progress intending to a portion of the primary sources of years lived with disability, for example, emotional health issue and musculoskeletal issue. The major limitation is the limited availability of data which resulted in less accuracy. Diet was not considered because of the restriction to the global burden of disease study. Our research also doesnâĂŹt include any nutritional analysis, but the usage of demographic parameters for an individual will be good enough to enhance the outcome.

According to the paper Patterns of Mortality by Occupation in the UK, 1991-2011: a comparative analysis of linked census and mortality records written by Dr. Srinivasa Vittal Katikireddi, Prof. Alastair H Leyland, Prof. Martin McKee, Kevin Ralston, Prof David Stuckler, this paper evaluated mortality by comprehensive work-related groups in each part of the UK (England and Wales, Scotland, and Northern Ireland), contrasts in rates among England and Wales and Scotland, and changes after some time in Scotland. They found that the men belonging to medical proficiency, business and public services comparatively had a very low death rate. Men working in factories, construction sites, delivery executive and heavy machinery operative were found to have a high mortality rate. For women, there were different categories for each country in the UK. In England, Scotland and Northern Ireland women working in teaching and business areas were found to have a low mortality rate. The high death rate was found with women working in factories and plantation sectors. The biggest advantage to this study was the availability of huge samples of datasets which allowed them to make a thorough analysis. This also helped them in avoiding the confusion of finding the occupation from death certificate and self-report. However, their study has a major drawback; they did not perform analysis on individuals having another job instead of their main job. They also did not take the persons health status into account. Or paper talks about each job specifically instead of a domain. We make an analysis based on how people from different age groups will react to their respective job.

According to the paper Medical Certificates of Cause of Death for people with intellectual disabilities: A systematic literature review written by Fred Dunwoodie Stirton, Pauline Heslop, this paper strives show the existence of an orderly survey of research relating to the accuracy of medical certificates for the cause of death for distinguishing reasons for the death of individuals with intellectual disability. The survey condenses investigate that recognizes potential troubles in depending on medical certificates for the cause of death to help comprehend the reasons for the death of individuals with intellectual disability, why these challenges may happen and the effect they have. The paper finishes up with proposals for announcing the reasons for the death of individuals with an intellectual disability on medical certificates for the cause of death. There were many misleading entries in the data of the which was difficult for them to adapt. This happened since the entire data is completely taken from the medical certificates for the cause of death. The report did not have many details which limited their research.

According to the paper Disability Status, Mortality, and Leading Causes of Death in the United States Community Population written by Valerie L. Forman-Hoffman, MPH, Kimberly L. Ault, Wayne L. Anderson, Joshua M. Weiner, Alissa Stevens, Vincent A. Campbell, and Brian S. Armour, they utilized information on US adults to evaluate the relationship among disability on all-cause mortality and on cause-explicit death. Disability types include impairment, cognitive, movement, and employment disability. Cause-explicit passingâĂŹs included coronary illness, dangerous neoplasms, perpetual lower respiratory ailments, cerebrovascular infections, inadvertent mishaps, and suicides/ambushes. They utilized Cox corresponding risk relapse models to gauge the probability of mortality, consolidating both times to death and also alteration for several useful disadvantages and statistic and financial attributes. Also, they inspected the rank-requested reasons for death for adults with and without an incapacity at a standard which later died. The usage of Cox proportional models instead of the regression model to estimate the death until it occurred was a big asset to their research. They did not take behavioral characteristics like smoking, drinking, and any chronic health issues into consideration.

III. Analysing the Mortality dataset of Bihar

A. Data Source

This is a unit level survey dataset containing the details relating to deaths of various residents in the households of Bihar during the period 2007 to 2011 and it includes information on sex of deceased, date of death, age at death, registration of death and source of medical attention received before death. It has more than 70 features representing individuals, their living conditions and factors leading to death. It also includes information on various crucial factors affecting maternal mortality i.e case of deaths associated with pregnancy, information on situa- tions leading or contributing to death, symptoms preceding death etc. Some examples of attributes present in the data set are deceased race, place of death, treatment source, place of death, death symptoms, sex, date of birth, disability status, injury treatment type, diagnosis source etc.

B. Basic Statistics from the data

Let's have a look at the basic statistics relating to the various deaths summarized from this dataset. All these statistics have been represented with the help of a visualization tool Infogram.

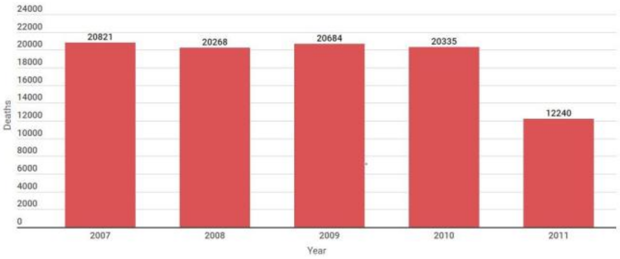


Fig. 1. Comparing the number of deaths in the 5 years. The number of deaths is in the same range from 2001 to 2010, however the deaths have decreased in 2011

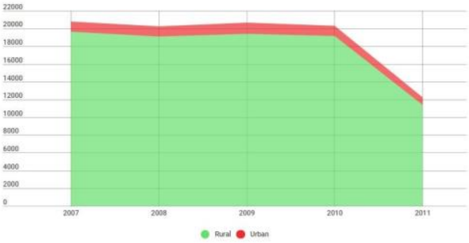


Fig. 2. An area chart Comparing the number of deaths in urban vs rural areas

We have extracted summary tables from R studio by aggregating information in all records in the dataset. Even though all the information has been summarized in the form of simple tables as shown above, it is still difficult to keep track of all this. This is where data visualization comes into picture. You will be surprised how easy it is to grasp all the information once they are all represented in the form of graphs and charts.

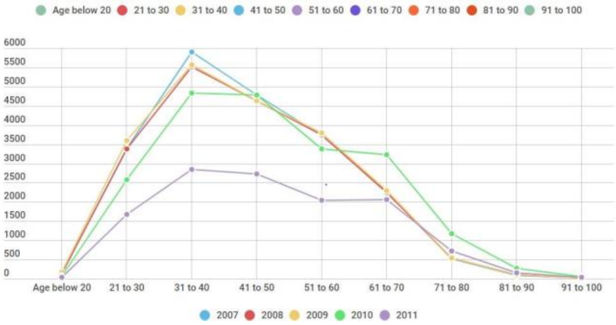


Fig. 3. Comparing the number of deaths in different age groups. The numerical variable AGE has been converted into a categorical variable AGE GROUP dividing the age into 9 categories. The majority of deaths appear to be in the age group 30-50

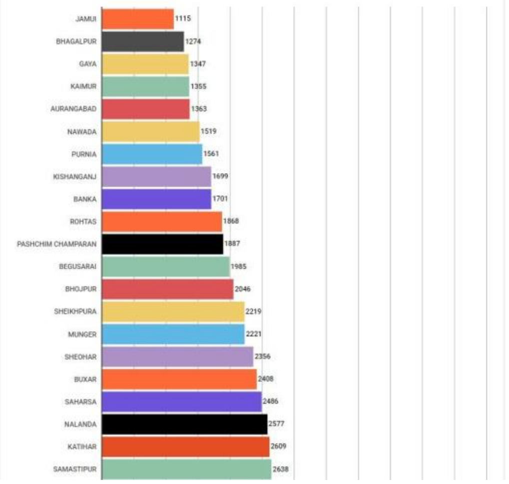


Fig. 4. Comparing the number of deaths in different districts-part 1

Here is the link for the visuals created in Infogram for the above discussed dataset

https://infogram.com/1pmn6wd76d5y0vs3yxj7nn7p77fzldrg21y?live

This link contains a series of simple interactive graphs and charts representing all the important information from the dataset. Visual models like bar charts, pie charts, line plots, stacked bar charts, area charts, donut charts, and grouped bar charts have been used for this. Some of them have been mentioned below for illustration

Once you go through all these visuals it is very easy to summarize the entire dataset of almost 1 lakh records in a few sentences. It becomes a simple task for the user to come to conclusions based on the above statistics which are from 2007 to 2011, the number of deaths were in the same range, somewhere around 20k, but the number of deaths dropped to 2011 in 12k. In all those 5 years, majority of the deaths are in the age group of 30 to 50. Jamui is the district with lowest number of deaths (1115) and in these 5 years and Khagaria is the district with highest number of deaths (5528). Deaths in rural area are significantly higher than deaths in rural area. Male deaths are very much higher than female deaths.

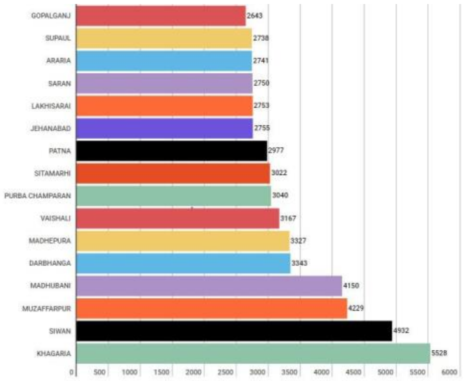


Fig. 5. Comparing the number of deaths in different districts-part 2. - Jamui is the district with lowest number of deaths (1115) and in these 5 years and Khagaria is the district with highest number of deaths (5528)

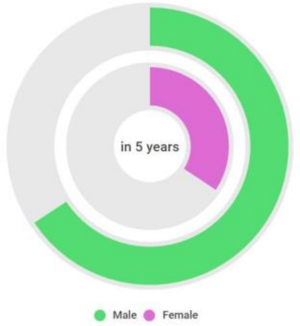


Fig. 6. The number of male deaths is significantly higher than the number of female deaths

Majority of deaths are unregistered although the registration of deaths relative to unregistered deaths is slowly increasing year by year.

IV. Prediction model for registration of deaths

A. Variable Selection

Variable selections were done based on All Possible Regression method which tests all possible subsets of a set of independent variables. If there are K potential independent variables, then there are 2k2k distinct subsets of them to be tested. For example, if you have 10 candidate independent variables, the number of subsets to be tested is 210210, which is 1024, and if you have 20 candidate variables, the number is 220220, which is more than one million. This can be implemented by installing Olsrr package in Rstudio The

âĂŹplotâĂŹ method shows the panel of fit criteria for all

possible regression methods.

model1 <- lm(is death reg agegroup+is death associated with pregnan+treatment source+death symptoms+occupation status, data=mortality)

kmodel<-olsstepallpossible(model)

plot(kmodel)

Please find below a sample of kmodel plot and the exported kmodel data which contains all the combination of variables from which the combinations suggested by the model were selected

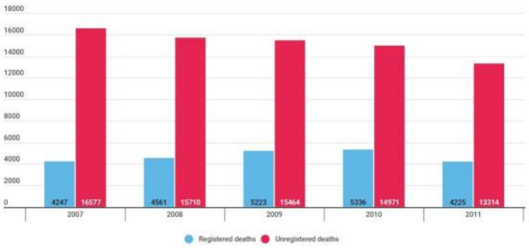


Fig. 7. The number of registered deaths relative to the number of unregistered deaths is very low but its slowly increasing over the years

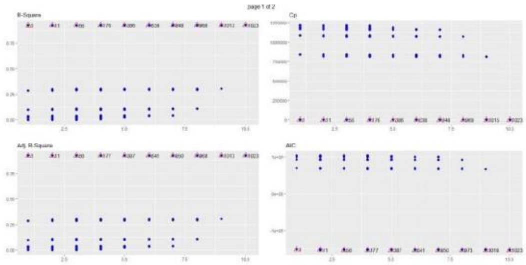
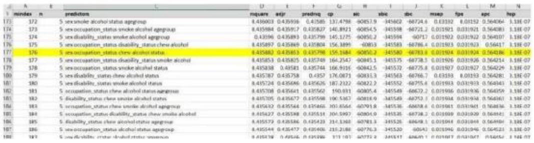


Fig. 8. Kmodel plot



	mindex	n	predictors	rsquare	adjr	predrsq	cp	aic	sbic	sbc	msep	fpe	apc	hsp
173	172	5	sex smoke alcohol status agegroup	0.436003	0.435936	0.43585	137.4798	-60857.9	-345602	-68724.6	0.03192	0.03192	0.564064	1.18E-07
174	173	5	sex occupation_status smoke alcohol agegroup	0.435984	0.435917	0.435827	140.8921	-60854.5	-345598	-68721.2	0.031921	0.031921	0.564083	1.18E-07
175	174	5	sex disability_status smoke alcohol agegroup	0.43596	0.435893	0.435799	145.1275	-60850.2	-345594	-60717	0.031922	0.031922	0.564107	1.18E-07
176	175	5	sex occupation_status disability_status chew alcohol	0.435897	0.435869	0.435804	156.3899	-60853	-345583	-68786.4	0.031923	0.031923	0.56417	1.18E-07
177	176	5	sex occupation_status chew alcohol status	0.435882	0.435853	0.435798	159.1684	-60850.2	-345580	-68783.6	0.031924	0.031924	0.564186	1.18E-07
178	177	5	sex occupation_status disability_status smoke alcohol	0.435853	0.435825	0.435749	164.2547	-60845.1	-345575	-68738.5	0.031926	0.031926	0.564214	1.18E-07
179	178	5	sex occupation_status smoke alcohol status	0.435838	0.43581	0.435744	166.9316	-60842.5	-345572	-68775.8	0.031927	0.031927	0.564229	1.18E-07
180	179	5	sex disability_status chew alcohol status	0.435787	0.435758	0.4357	176.0871	-60833.3	-345563	-68766.7	0.03193	0.03193	0.564281	1.18E-07
181	180	5	sex disability_status smoke alcohol status	0.435724	0.435696	0.435626	187.2322	-60822.2	-345552	-68755.4	0.031933	0.031933	0.564341	1.18E-07
182	181	5	occupation_status chew alcohol status agegroup	0.435708	0.435641	0.435562	190.031	-60805.4	-345549	-68672.2	0.031936	0.031936	0.564359	1.18E-07
183	182	5	disability_status chew smoke alcohol status	0.435705	0.435677	0.435598	190.5367	-60818.9	-345549	-68752.1	0.031934	0.031934	0.564361	1.18E-07
184	183	5	occupation_status chew smoke alcohol agegroup	0.435632	0.435564	0.435466	203.6564	-60791.8	-345536	-68658.6	0.031941	0.031941	0.564436	1.18E-07
185	184	5	occupation_status disability_status chew smoke alcohol	0.435627	0.435598	0.435514	204.5997	-60804.0	-345535	-68738.2	0.031939	0.031939	0.564441	1.18E-07
186	185	5	disability_status chew alcohol status agegroup	0.435573	0.435506	0.435423	214.1368	-60781.3	-345526	-68648.1	0.031944	0.031944	0.564494	1.18E-07
187	186	5	sex occupation_status alcohol status agegroup	0.435544	0.435477	0.435406	219.2188	-60776.3	-345520	-60643	0.031946	0.031946	0.564523	1.18E-07
188	187	5	sex disability_status alcohol status agegroup	0.435438	0.43546	0.435396	222.181	-60773.3	-345517	-68640.1	0.031947	0.031947	0.56454	1.18E-07

Fig. 9. All the possible combinations generated by kmodel

B. Building a regression model for prediction

Linear Regression could be a straight approach to demonstrating the relationship between a dependent variable and one or more independent variables which can be used as a basis for prediction. In our data set, a sample prediction test was done by plotting a regression line for the variable age against the variables year of marriage and year of birth. Please find below the sample code, linear regression plot and the predicted plot for the test variables used.

plot(age year-of-marriage, data = mortality4)

m<-mean(mortality4age)

abline(h=m)

lm1<-lm(age year-of-marriage, data = mortality4)

abline(lm1, col=âĂIredâĂI)

lm2<-lm(age year-of-marriage âĹŮ year-of-birth, mortality4) plot(lm2)

termplot(lm2)

summary(lm2)

m<-mean(year-of-marriage)

m2<-1969

p1<-predict(lm2, data.frame(year-of-marriage = m, year-of-birth=1920:1993)) p2<-predict(lm2, data.frame(year-of-marriage = m2, year-of-birth=1920:1993))

plot(age year-of-marriage, mortality4) lines(1920:1993, p1, col=âĂIredâĂI) lines(1920:1993, p2, col=âĂIblueâĂI)

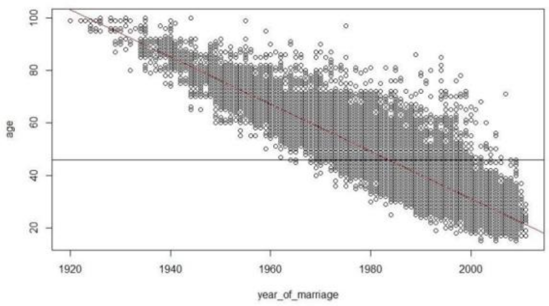


Fig. 10. Base Regression line

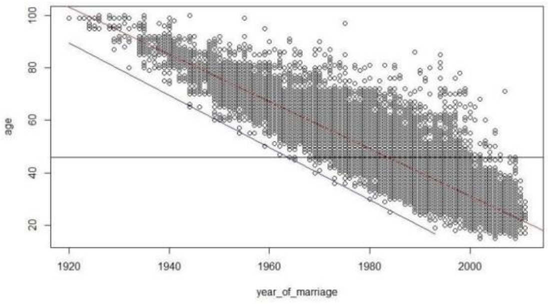


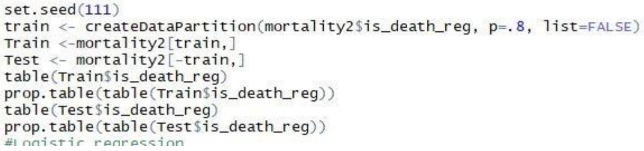
Fig. 11. Regression line of prediction model

Red line is the base regression line and Blue line is the prediction. It is evident that the prediction is in tune with the regression line. However this model is only applicable for continuous variables. When your target variable i.e. the variable you are trying to predict is a categorical one, all the values wont fit in a simple straight line. We will be needing something more sophisticated than this. This is where logistic regression comes into picture.

Similar to the regression analyses, the logistic regression is also a predictive analysis. Logistic regression is usually used to explain the relationship between one dependent binary variable and one or more nominal, ordinal independent variables. When the response variable is categorical in nature, logical regression analyses is best suited for predictions. In our case, the target variable is is-death-reg(death registered or not) which is binomial in nature consisting of two categories (0 and 1 for not registered and registered). This variable is-death-reg is plotted against other independent variables like agegroup, treatment source, death-symptoms, occupation status, death related to pregnancy.

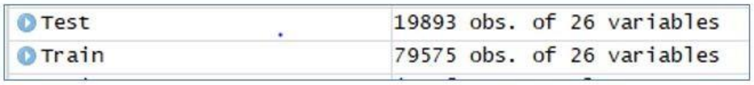
First we partition our data in training set and test set. 80 percent of our data is training set and 20 percent is test set.

Once you look at the summary of initial model, you will notice the p values of categories in age group variable in high, which means the confidence interval (1- P) is low. Consider agegroup 31-40 with p vale of 0.673. The confidence interval 1-0.673 = 0.327 i.e.32.7 percent is very low. This implies that we can drop this variable from the model as it does not have any predictive power The more stars a category has, the more important is that for prediction, lesser the stars , less is it usefulness in prediction.



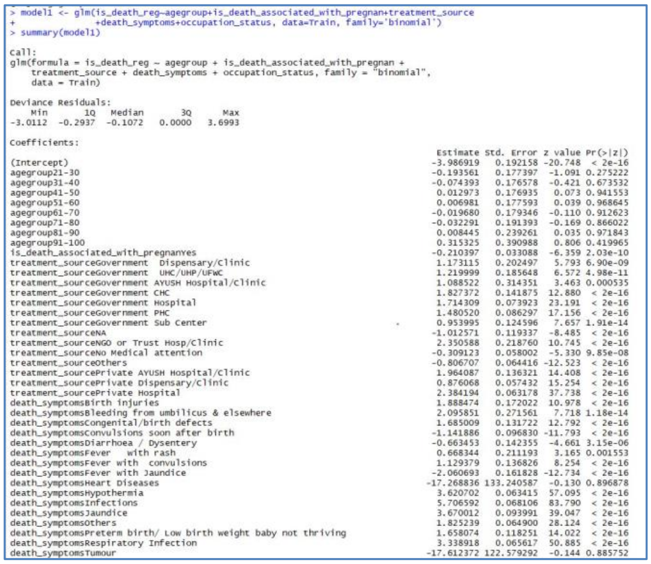
```
set.seed(111)
train <- createDataPartition(mortality2$is_death_reg, p=.8, list=FALSE)
Train <-mortality2[train,]
Test <- mortality2[-train,]
table(Train$is_death_reg)
prop.table(table(Train$is_death_reg))
table(Test$is_death_reg)
prop.table(table(Test$is_death_reg))
#Logistic regression
```

Fig. 12. Code snippet of data partition



```
Test     19893 obs. of 26 variables
Train    79575 obs. of 26 variables
```

Fig. 13. Data sets partitioned into training and test sets



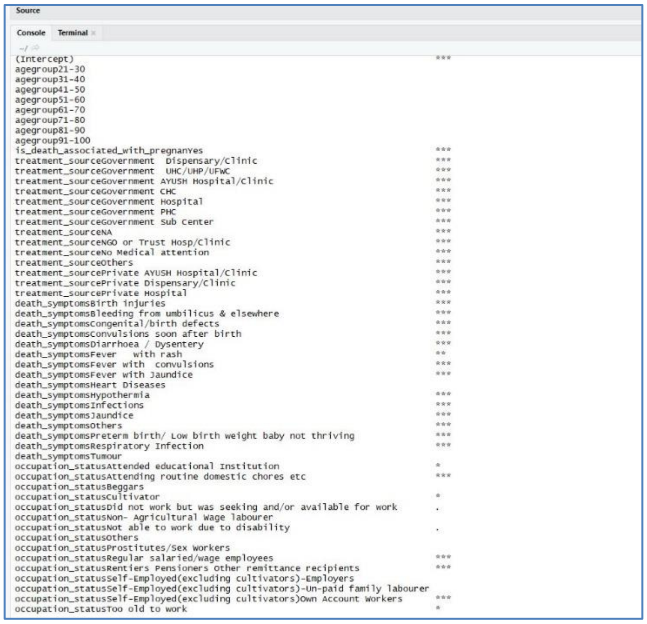
```
> model1 <- glm(is_death_reg~agegroup+is_death_associated_with_pregnan+treatment_source
+             +death_symptoms+occupation_status, data=Train, family='binomial')
> summary(model1)

Call:
glm(formula = is_death_reg ~ agegroup + is_death_associated_with_pregnan + 
    treatment_source + death_symptoms + occupation_status, family = "binomial", 
    data = Train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0112  -0.2937  -0.1072   0.0000   3.6993  

Coefficients:
                                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                    -3.986919   0.192158 -20.748  < 2e-16
agegroup21-30                                                  -0.193561   0.177397  -1.091 0.275222
agegroup31-40                                                  -0.074393   0.176578  -0.421 0.673532
agegroup41-50                                                   0.012973   0.176935   0.073 0.941553
agegroup51-60                                                   0.006981   0.177593   0.039 0.968645
agegroup61-70                                                  -0.019680   0.179346  -0.110 0.912623
agegroup71-80                                                  -0.032291   0.191393  -0.169 0.866022
agegroup81-90                                                   0.008445   0.239261   0.035 0.971843
agegroup91-100                                                  0.315325   0.390988   0.806 0.419965
is_death_associated_with_pregnanYes                            -0.210397   0.033088  -6.359 2.03e-10
treatment_sourceGovernment  Dispensary/Clinic                   1.173115   0.202497   5.793 6.90e-09
treatment_sourceGovernment  UHC/UHP/UFWC                        1.219999   0.185648   6.572 4.98e-11
treatment_sourceGovernment AYUSH Hospital/Clinic                1.088522   0.314351   3.463 0.000535
treatment_sourceGovernment CHC                                  1.827372   0.141875  12.880  < 2e-16
treatment_sourceGovernment Hospital                             1.714309   0.073923  23.191  < 2e-16
treatment_sourceGovernment PHC                                  1.480520   0.086297  17.156  < 2e-16
treatment_sourceGovernment Sub Center                           0.953995   0.124596   7.657 1.91e-14
treatment_sourceNA                                             -1.012571   0.119337  -8.485  < 2e-16
treatment_sourceNGO or Trust Hosp/Clinic                        2.350588   0.218760  10.745  < 2e-16
treatment_sourceNo Medical attention                           -0.309123   0.058002  -5.330 9.85e-08
treatment_sourceOthers                                         -0.806707   0.064416 -12.523  < 2e-16
treatment_sourcePrivate AYUSH Hospital/Clinic                   1.964087   0.136321  14.408  < 2e-16
treatment_sourcePrivate Dispensary/Clinic                       0.876068   0.057432  15.254  < 2e-16
treatment_sourcePrivate Hospital                                2.384194   0.063178  37.738  < 2e-16
death_symptomsBirth injuries                                    1.888474   0.172022  10.978  < 2e-16
death_symptomsBleeding from umbilicus & elsewhere               2.095851   0.271561   7.718 1.18e-14
death_symptomsCongenital/birth defects                          1.685009   0.131722  12.792  < 2e-16
death_symptomsConvulsions soon after birth                     -1.141886   0.096830 -11.793  < 2e-16
death_symptomsDiarrhoea / Dysentery                            -0.663453   0.142355  -4.661 3.15e-06
death_symptomsFever   with rash                                 0.668344   0.211193   3.165 0.001553
death_symptomsFever with  convulsions                           1.129379   0.136826   8.254  < 2e-16
death_symptomsFever with Jaundice                              -2.060693   0.161828 -12.734  < 2e-16
death_symptomsHeart Diseases                                  -17.268836 133.240587  -0.130 0.896878
death_symptomsHypothermia                                       3.620702   0.063415  57.095  < 2e-16
death_symptomsInfections                                        5.706592   0.068106  83.790  < 2e-16
death_symptomsJaundice                                          3.670012   0.093991  39.047  < 2e-16
death_symptomsOthers                                            1.825239   0.064900  28.124  < 2e-16
death_symptomsPreterm birth/ Low birth weight baby not thriving 1.658074   0.118251  14.022  < 2e-16
death_symptomsRespiratory Infection                             3.338918   0.065617  50.885  < 2e-16
death_symptomsTumour                                          -17.612372 122.579292  -0.144 0.885752
```

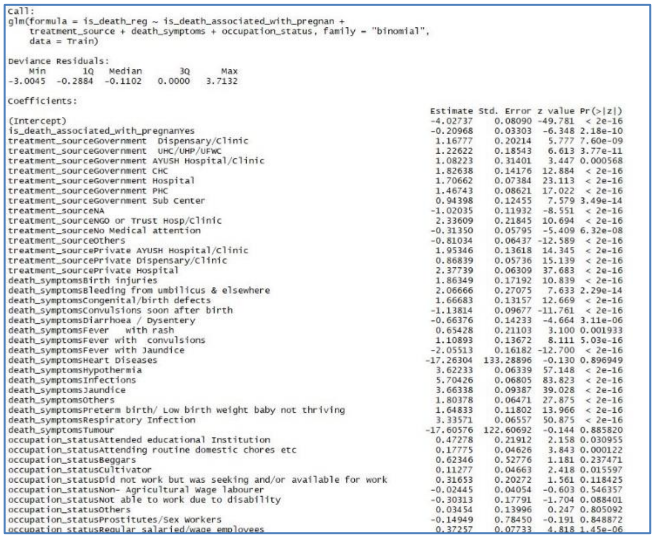
Fig. 14. Summary of Initial model part 1



```
(Intercept)                                                                ***
agegroup21-30
agegroup31-40
agegroup41-50
agegroup51-60
agegroup61-70
agegroup71-80
agegroup81-90
agegroup91-100
is_death_associated_with_pregnanYes                                        ***
treatment_sourceGovernment  Dispensary/Clinic                              ***
treatment_sourceGovernment  UHC/UHP/UFWC                                   ***
treatment_sourceGovernment AYUSH Hospital/Clinic                           ***
treatment_sourceGovernment CHC                                             ***
treatment_sourceGovernment Hospital                                        ***
treatment_sourceGovernment PHC                                             ***
treatment_sourceGovernment Sub Center                                      ***
treatment_sourceNA                                                         ***
treatment_sourceNGO or Trust Hosp/Clinic                                   ***
treatment_sourceNo Medical attention                                       ***
treatment_sourceOthers                                                     ***
treatment_sourcePrivate AYUSH Hospital/Clinic                              ***
treatment_sourcePrivate Dispensary/Clinic                                  ***
treatment_sourcePrivate Hospital                                           ***
death_symptomsBirth injuries                                               ***
death_symptomsBleeding from umbilicus & elsewhere                          ***
death_symptomsCongenital/birth defects                                     ***
death_symptomsConvulsions soon after birth                                 ***
death_symptomsDiarrhoea / Dysentery                                        ***
death_symptomsFever   with rash                                            **
death_symptomsFever with  convulsions                                      ***
death_symptomsFever with Jaundice                                          ***
death_symptomsHeart Diseases
death_symptomsHypothermia                                                  ***
death_symptomsInfections                                                   ***
death_symptomsJaundice                                                     ***
death_symptomsOthers                                                       ***
death_symptomsPreterm birth/ Low birth weight baby not thriving            ***
death_symptomsRespiratory Infection                                        ***
death_symptomsTumour
occupation_statusAttended educational Institution                          *
occupation_statusAttending routine domestic chores etc                     ***
occupation_statusBeggars
occupation_statusCultivator                                                *
occupation_statusDid not work but was seeking and/or available for work    .
occupation_statusNon- Agricultural Wage labourer
occupation_statusNot able to work due to disability                        .
occupation_statusOthers
occupation_statusProstitutes/Sex Workers
occupation_statusRegular salaried/wage employees                           ***
occupation_statusRentiers Pensioners Other remittance recipients           ***
occupation_statusSelf-Employed(excluding cultivators)-Employers
occupation_statusSelf-Employed(excluding cultivators)-Un-paid family labourer
occupation_statusSelf-Employed(excluding cultivators)Own Account Workers   ***
occupation_statusToo old to work                                           *
```

Fig. 15. Summary of Initial model part 2

Generating a better model using the training dataset after removing the variable age group. Please find below the summary of the improved model below.

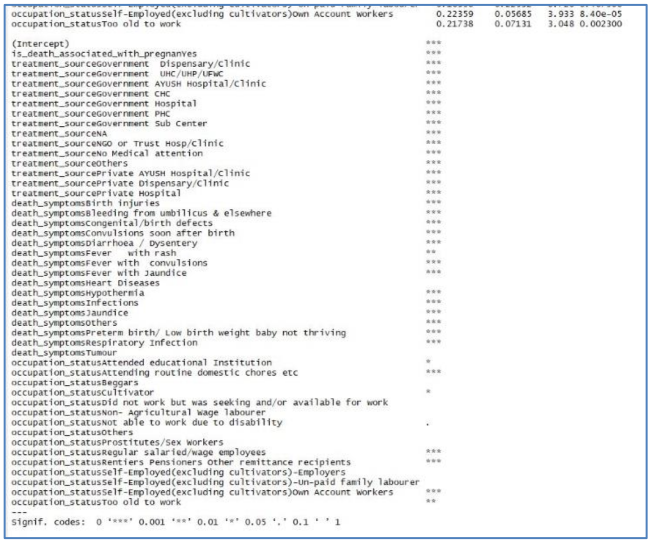


```
Call:
glm(formula = is_death_reg ~ is_death_associated_with_pregnan + 
    treatment_source + death_symptoms + occupation_status, family = "binomial", 
    data = Train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0045  -0.2884  -0.1102   0.0000   3.7132  

Coefficients:
                                                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                               -4.02737    0.08090 -49.781  < 2e-16
is_death_associated_with_pregnanYes                                       -0.20968    0.03303  -6.348 2.18e-10
treatment_sourceGovernment  Dispensary/Clinic                              1.16777    0.20214   5.777 7.60e-09
treatment_sourceGovernment  UHC/UHP/UFWC                                   1.22622    0.18543   6.613 3.77e-11
treatment_sourceGovernment AYUSH Hospital/Clinic                           1.08223    0.31401   3.447 0.000568
treatment_sourceGovernment CHC                                             1.82638    0.14176  12.884  < 2e-16
treatment_sourceGovernment Hospital                                        1.70662    0.07384  23.113  < 2e-16
treatment_sourceGovernment PHC                                             1.46743    0.08621  17.022  < 2e-16
treatment_sourceGovernment Sub Center                                      0.94398    0.12455   7.579 3.49e-14
treatment_sourceNA                                                        -1.02035    0.11932  -8.551  < 2e-16
treatment_sourceNGO or Trust Hosp/Clinic                                   2.33609    0.21845  10.694  < 2e-16
treatment_sourceNo Medical attention                                      -0.31350    0.05795  -5.409 6.32e-08
treatment_sourceOthers                                                    -0.81034    0.06437 -12.589  < 2e-16
treatment_sourcePrivate AYUSH Hospital/Clinic                              1.95346    0.13618  14.345  < 2e-16
treatment_sourcePrivate Dispensary/Clinic                                  0.86839    0.05736  15.139  < 2e-16
treatment_sourcePrivate Hospital                                           2.37739    0.06309  37.683  < 2e-16
death_symptomsBirth injuries                                               1.86349    0.17192  10.839  < 2e-16
death_symptomsBleeding from umbilicus & elsewhere                          2.06666    0.27075   7.633 2.29e-14
death_symptomsCongenital/birth defects                                     1.66683    0.13157  12.669  < 2e-16
death_symptomsConvulsions soon after birth                                -1.13814    0.09677 -11.761  < 2e-16
death_symptomsDiarrhoea / Dysentery                                       -0.66376    0.14233  -4.664 3.11e-06
death_symptomsFever   with rash                                            0.65428    0.21103   3.100 0.001933
death_symptomsFever with  convulsions                                      1.10893    0.13672   8.111 5.03e-16
death_symptomsFever with Jaundice                                         -2.05513    0.16182 -12.700  < 2e-16
death_symptomsHeart Diseases                                             -17.26304  133.28896  -0.130 0.896949
death_symptomsHypothermia                                                  3.62233    0.06339  57.148  < 2e-16
death_symptomsInfections                                                   5.70426    0.06805  83.823  < 2e-16
death_symptomsJaundice                                                     3.66338    0.09387  39.028  < 2e-16
death_symptomsOthers                                                       1.80378    0.06471  27.875  < 2e-16
death_symptomsPreterm birth/ Low birth weight baby not thriving            1.64833    0.11802  13.966  < 2e-16
death_symptomsRespiratory Infection                                        3.33571    0.06557  50.875  < 2e-16
death_symptomsTumour                                                     -17.60576  122.60692  -0.144 0.885820
occupation_statusAttended educational Institution                          0.47278    0.21912   2.158 0.030955
occupation_statusAttending routine domestic chores etc                     0.17775    0.04626   3.843 0.000122
occupation_statusBeggars                                                   0.62346    0.52776   1.181 0.237471
occupation_statusCultivator                                                0.11277    0.04663   2.418 0.015597
occupation_statusDid not work but was seeking and/or available for work    0.31653    0.20272   1.561 0.118425
occupation_statusNon- Agricultural Wage labourer                          -0.02445    0.04054  -0.603 0.546357
occupation_statusNot able to work due to disability                       -0.30313    0.17791  -1.704 0.088401
occupation_statusOthers                                                    0.03454    0.13996   0.247 0.805092
occupation_statusProstitutes/Sex Workers                                  -0.14949    0.78450  -0.191 0.848872
occupation_statusRegular salaried/wage employees                           0.37257    0.07733   4.818 1.45e-06
```

Fig. 16. Summary of Improved model part 1



```
occupation_statusSelf-Employed(excluding cultivators)Own Account Workers   0.22359    0.05685   3.933 8.40e-05
occupation_statusToo old to work                                           0.21738    0.07131   3.048 0.002300

(Intercept)                                                               ***
is_death_associated_with_pregnanYes                                       ***
treatment_sourceGovernment  Dispensary/Clinic                             ***
treatment_sourceGovernment  UHC/UHP/UFWC                                  ***
treatment_sourceGovernment AYUSH Hospital/Clinic                          ***
treatment_sourceGovernment CHC                                            ***
treatment_sourceGovernment Hospital                                       ***
treatment_sourceGovernment PHC                                            ***
treatment_sourceGovernment Sub Center                                     ***
treatment_sourceNA                                                        ***
treatment_sourceNGO or Trust Hosp/Clinic                                  ***
treatment_sourceNo Medical attention                                      ***
treatment_sourceOthers                                                    ***
treatment_sourcePrivate AYUSH Hospital/Clinic                             ***
treatment_sourcePrivate Dispensary/Clinic                                 ***
treatment_sourcePrivate Hospital                                          ***
death_symptomsBirth injuries                                              ***
death_symptomsBleeding from umbilicus & elsewhere                         ***
death_symptomsCongenital/birth defects                                    ***
death_symptomsConvulsions soon after birth                                ***
death_symptomsDiarrhoea / Dysentery                                       ***
death_symptomsFever   with rash                                           **
death_symptomsFever with  convulsions                                     ***
death_symptomsFever with Jaundice                                         ***
death_symptomsHeart Diseases
death_symptomsHypothermia                                                 ***
death_symptomsInfections                                                  ***
death_symptomsJaundice                                                    ***
death_symptomsOthers                                                      ***
death_symptomsPreterm birth/ Low birth weight baby not thriving           ***
death_symptomsRespiratory Infection                                       ***
death_symptomsTumour
occupation_statusAttended educational Institution                         *
occupation_statusAttending routine domestic chores etc                    ***
occupation_statusBeggars
occupation_statusCultivator                                               *
occupation_statusDid not work but was seeking and/or available for work
occupation_statusNon- Agricultural Wage labourer
occupation_statusNot able to work due to disability                       .
occupation_statusOthers
occupation_statusProstitutes/Sex Workers
occupation_statusRegular salaried/wage employees                          ***
occupation_statusRentiers Pensioners Other remittance recipients          ***
occupation_statusSelf-Employed(excluding cultivators)-Employers
occupation_statusSelf-Employed(excluding cultivators)-Un-paid family labourer
occupation_statusSelf-Employed(excluding cultivators)Own Account Workers  ***
occupation_statusToo old to work                                          **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 17. Summary of Improved model part 2

In this model, we can see that almost all the categories have very less p values indicating high confidence intervals. Most of the categories have three stars indicating that they are all important in prediction. Please find below the plot generated for this logistic regression model.

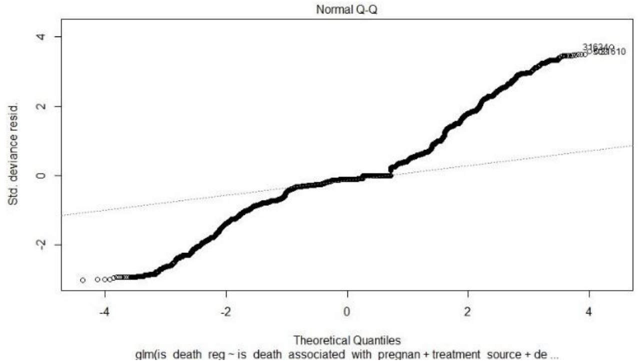
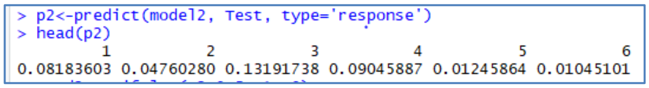


Fig. 18. Plot of logistic regression model

Let us run this prediction model on test dataset. You can see below the prediction for the first 6 records in the test dataset.

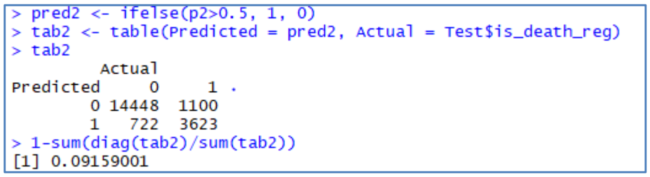


```
> p2<-predict(model2, Test, type='response')
> head(p2)
         1          2          3          4          5          6 
0.08183603 0.04760280 0.13191738 0.09045887 0.01245864 0.01045101 
```

Fig. 19. Prediction probabilities for first 6 records of test set

There is 0.08 percent that the first candidate death was registered. In fact all the first 6 candidates deaths are highly unlikely to be registered.To correctly validate the predictive power of our model, we can calculate a confusion matrix as For this we have defined that if the probability of death being registered is > 0.5, then consider it as 1 (death is registered) else 0 (death not registered).

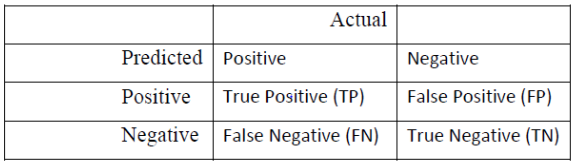
14448 deaths have been correctly predicted as being not registered. 3623 deaths have been correctly predicted as deaths being registered. Overall out of the 19893 entries, 18071 of them have been correctly predicted and the misclassification rate is very low (0.091Sensitivity and Specificity analysis is used to assess the performance of a test Those metrics can be calculated from the confusion matrix.



```
> pred2 <- ifelse(p2>0.5, 1, 0)
> tab2 <- table(Predicted = pred2, Actual = Test$is_death_reg)
> tab2
         Actual
Predicted     0     1
        0 14448  1100
        1   722  3623
> 1-sum(diag(tab2)/sum(tab2))
[1] 0.09159001
```

Fig. 20. Generating confusion matrix

True positive (TP): Number of cases that the test declares positive and that are truly positive. False positive (FP): Number of cases that the test declares positive and that in reality are negative. True negative (VN): Number of cases that the test declares negative and that are truly negative. False negative (FN): Number of cases that the test declares negative and that in reality are positive.

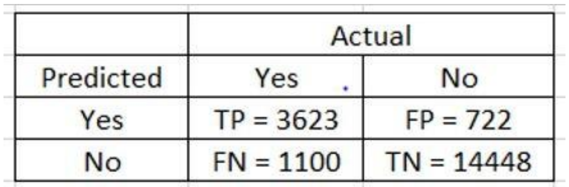


	Actual	
Predicted	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Fig. 21. Confusion matrix format

Lets consider the confusion matrix obtained for the logical regression prediction model (Figure 23Sensitivity (equivalent to the True Positive Rate): Propor- tion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals (in our case the registered deaths) Sensitivity = TP/(TP + FN) = 3623/(3623+1100) =0.76

Specificity (also called True Negative Rate): Proportion of negative cases that are well detected by the test. In other words, specificity measures how the test is effective when used on negative individuals (in our case the negative deaths) Specificity = TN/(TN + FP) = 14448/(14448+722) = 0.95

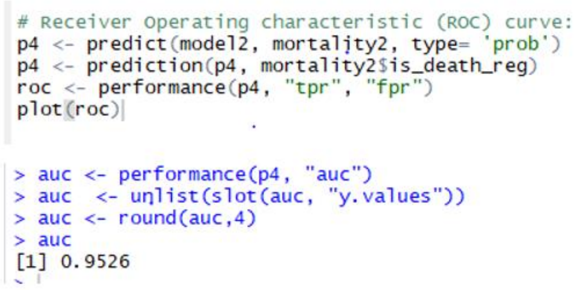


	Actual	
Predicted	Yes	No
Yes	TP = 3623	FP = 722
No	FN = 1100	TN = 14448

Fig. 22. Confusion matrix for our logistic prediction model

C. ROC curve and AUC

An ROC curve (receiver operating characteristic curve) is a graph displaying the performance of a classification model at all the different thresholds. This curve plots two parameters: True Positive Rate (probability of detection) = TP/(TP + FN) = 0.76 for our model False Positive Rate (probability of false alarm) = FP/(FP + FN) = 0.39 for our model TPR 0.76 and FPR 0.39 are the classification thresholds for our logistic prediction model. An ROC curve plots the curve for all the thresholds of TPR and FPR from 0 to 1. To compute the points in an ROC curve, we could assess a logistic regression model numerous times with various classification thresholds, however this would be very time consuming and inefficient. Fortunately, there's a more productive, sorting-based algorithm that can provide this information for us, called AUC (Area under Curve). AUC measures the whole two-dimensional area under the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of defining AUC is as the probability that the model ranks a random positive sample more highly than a random negative sample. AUC is alluring to use for the following two reasons: AUC is scale-invariant. It measures how good the predictions are ranked, rather than their outright values. It measures the quality of the model's forecasts independent of what classification threshold is chosen. The value of AUC can be anything between 0 to 1. The higher the value of AUC, the better is the model performance. A prediction model whose predictions are 100 percent wrong has an AUC of 0.0; one whose predictions are 100 percent correct has an AUC of 1.0.



```
# Receiver Operating characteristic (ROC) curve:
p4 <- predict(model2, mortality2, type= 'prob')
p4 <- prediction(p4, mortality2$is_death_reg)
roc <- performance(p4, "tpr", "fpr")
plot(roc)
```

```
> auc <- performance(p4, "auc")
> auc  <- unlist(slot(auc, "y.values"))
> auc <- round(auc,4)
> auc
[1] 0.9526
```

Fig. 23. Generating ROC and AUC

As you can see, the ROC of our prediction model has an AUC of 0.9526 which means, the prediction model is pretty good.

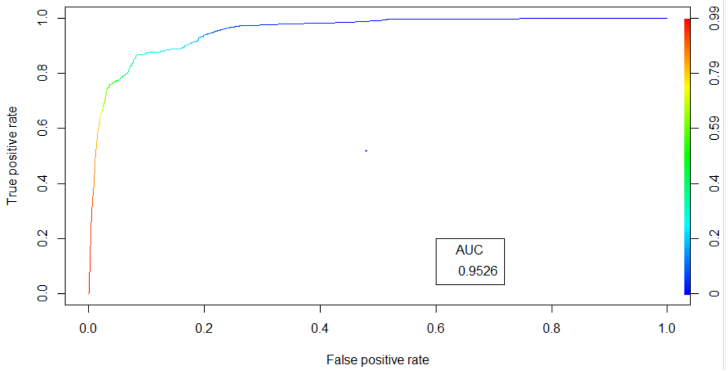


Fig. 24. ROC curve and AUC

V. Conclusion

In this paper, the individual level survey level mortality dataset of Bihar (a state in India) containing information regarding the various deaths of individuals from the year 2007 to 2011 has been analyzed.The various important statistics have been represented in the form of visual models. Some of the key points from the statistics are in all those 5 years, majority of the deaths are in the age group of 30 to 50. Khagaria is the district with highest number of deaths (5528). Deaths in rural area are significantly higher than deaths in rural area. Male deaths are very much higher than female deaths. Majority of deaths are unregistered although the relative percentage registration of deaths is slowly increasing year by year.

The registration of death is very important as it is a legal requirement. It gives imperative data around: the decedent, the cause of passing, and final mien. This data is utilized within the application for protections benefits, settlement of benefits claims, and exchange of title of genuine and individual property. The certificate is prima facie prove of the reality of passing and, so, can be presented in court as proof when a a question about the death arises The death certificate is the source for state and national mortality measurements. It is required for a assortment of restorative and health-related inquire about endeavors. It is utilized to decide which therapeutic conditions get inquire about and improvement subsidizing, to set open wellbeing objectives and arrangements, and to degree wellbeing status at neighborhood, state, national, and worldwide levels. This information is important as a investigation apparatus and by affecting investigate subsidizing. Statistical information determined from death certificates can be no more precise than the data on the certificate. In this manner, it is critical that everybody included with the registration of deaths endeavor for complete, exact, and prompt reporting of these events.

In this paper a logistic regression prediction model has been developed which analyses the deaths in Bihar state from 2007 to 2011 and predicts whether the death has been registered or not. This prediction model has been tested on the test set and 18071 deaths registrarion status has been correctly out of the total 19893 deaths acheiving an accuracy of 90.8 percent.

REFERENCES

[1]. Knott, C. S., Coombs, N., Stamatakis, E., & Biddulph, J. P. (2015). All cause mortality and the case for age specific alcohol consumption guidelines: Pooled analyses of up to 10 population-based cohorts. Bmj, 350(Feb10 2). doi:10.1136/bmj.h384

[2]. Kruk, M. E., Gage, A. D., Joseph, N. T., Danaei, G., García-Saisó, S., & Salomon, J. A. (2018). Mortality due to low-quality health systems in the universal health coverage era: A systematic analysis of amenable deaths in 137 countries. The Lancet, 392(10160), 2203-2212. doi:10.1016/s0140-6736(18)31668-4

[3]. Cao, H., Wang, J., Li, Y., Li, D., Guo, J., Hu, Y., . . . Zhang, L. (2017). Trend analysis of mortality rates and causes of death in children under 5 years old in Beijing, China from 1992 to 2015 and forecast of mortality into the future: An entire population-based epidemiological study. BMJ Open, 7(9). doi:10.1136/bmjopen-2017-015941

[4]. Silcocks, P. B. (2001). Life expectancy as a summary of mortality in a population: Statistical considerations and suitability for use by health authorities. Journal of Epidemiology & Community Health, 55(1), 38-43. doi:10.1136/jech.55.1.38

[5]. Carr, M. J., Ashcroft, D. M., Kontopantelis, E., While, D., Awenat, Y., Cooper, J., . . . Webb, R. T. (2017). Premature Death Among Primary Care Patients with a History of Self-Harm. The Annals of Family Medicine, 15(3), 246-254. doi:10.1370/afm.2054

[6]. Laatikainen, T. (2003). Increased mortality related to heavy alcohol intake pattern. Journal of Epidemiology & Community Health, 57(5), 379-384. doi:10.1136/jech.57.5.379

[7]. W. Krahwinkel E. Schuler M. Liebetrau A. Meier-Hellmann J. Zacher R. Kuhlen (2016). International Journal for Quality in Health Care, Volume 28, Issue 5, 10 October 2016, doi.org/10.1093/intqhc/mzw072

[8]. Accortt, N. A. (2002). Chronic Disease Mortality in a Cohort of Smokeless Tobacco Users. American Journal of Epidemiology, 156(8), 730-737. doi:10.1093/aje/kwf106

[9]. Rehm, J., Baliunas, D., Borges, G. L., Graham, K., Irving, H., Kehoe, T., . . . Taylor, B. (2010). The relation between different dimensions of alcohol consumption and burden of disease: An overview. Addiction, 105(5), 817-843. doi:10.1111/j.1360-0443.2010. 02899.x

[10]. Thun, M. J., Carter, B. D., Feskanich, D., Freedman, N. D., Prentice, R., Lopez, A. D., . . . Gapstur, S. M. (2013). 50-Year Trends in Smoking-Related Mortality in the United States. New England Journal of Medicine, 368(4), 351-364. doi:10.1056/nejmsa1211127

[11]. Henley, S. J., Connell, C. J., Richter, P., Husten, C., Pechacek, T., Calle, E. E., & Thun, M. J. (2007). Tobacco-related disease mortality among men who switched from cigarettes to spit tobacco. Tobacco Control, 16(1), 22-28. doi:10.1136/tc.2006.018069

[12]. Nilsen, C., Andel, R., Fritzell, J., & Kåreholt, I. (2016). Work-related stress in midlife and all-cause mortality: Can sense of coherence modify this association? The European Journal of Public Health, 26(6), 1055-1061. doi:10.1093/eurpub/ckw086

[13]. Bonsdorff, M. B., Seitsamo, J., Bonsdorff, M. E., Ilmarinen, J., Nygård, C., & Rantanen, T. (2012). Job strain among blue-collar and white-collar employees as a determinant of total mortality: A 28-year population-based follow-up. BMJ Open, 2(2). doi:10.1136/bmjopen-2012-000860

[14]. Brennan, T. A., & Leape, L. L. (2009). Adverse events, negligence in hospitalized patients: Results from the Harvard Medical Practice Study. Perspectives in Healthcare Risk Management, 11(2), 2-8. doi:10.1002/jhrm.5600110202

[15]. Mohan, P., Lando, H. A., & Panneer, S. (2018). Assessment of Tobacco Consumption and Control in India. Indian Journal of Clinical Medicine, 9, 117991611875928. doi:10.1177/1179916118759289

[16]. Katikireddi, S. V., Leyland, A. H., Mckee, M., Ralston, K., & Stuckler, D. (2017). Patterns of mortality by occupation in the UK, 1991–2011: A comparative analysis of linked census and mortality records. The Lancet Public Health,2(11). doi:10.1016/s2468-2667(17)30193-7

[17]. Stirton, F. D., & Heslop, P. (2018). Medical Certificates of Cause of Death for people with intellectual disabilities: A systematic literature review. Journal of Applied Research in Intellectual Disabilities, 31(5), 659-668. doi:10.1111/jar.12448

[18]. Forman-Hoffman, V. L., Ault, K. L., Anderson, W. L., Weiner, J. M., Stevens, A., Campbell, V. A., & Armour, B. S. (2015). Disability Status, Mortality, and Leading Causes of Death in the United States Community Population. Medical Care, 1. doi:10.1097/mlr.0000000000000321

[19]. Christopher J. L. Murray. (2018). The State of US Health, Burden of Diseases, Injuries, and Risk Factors Among US States. doi:10.1001/jama.2018.0158

[20] Hamid Reza Hassanzadeh A Review on Diagnosis of Diabetes in Data Mining. Volume 4 Issue 6, June 2015. https://arxiv.org/ftp/arxiv/papers/1705/1705.03508.pdf

[21] Anthony Lipphardt Using cuase of death literal text from death certificate for classification.
https://github.com/alipphardt/dmi-deaths-classification/blob/master/dmi-classif

[22] PBS Sicocks, D A Jenner, R Reza Life expectancy as a summary of mor- tality in a population: statistical considerations and suitability for use by health authorities https://jech.bmj.com/content/55/1/38

[24] Han Sao, Jing Wang Trend analysis of mortality rates and causes of death in children under 5 years old in Beijing, China from 1992 to 2015 and forecast of mortality into the future: an entire population-based epidemi- ological study. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5623503/

[25] Margaret E Kruk, Anna D Gage Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries.
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31668-4/full

[26] Craig S Knott, Jane P Biddulph All cause mortality and the case for age specific alcohol consumption guidelines: pooled analyses of up to 10 population based cohorts. https://www.bmj.com/content/350/bmj.h384

[27]. (2017). Smoking Prevalence and Attributable Disease Burden in 195 Countries and Territories, 1990-2015: A Systematic Analysis from the Global Burden of Disease Study 2015. doi: 10.1016/S0140-6736(17)30819