**Cause of Death**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** |  |
|  | **INTRODUCTION** |  |
|  | * 1. General Introduction |  |
|  | * 1. Project Objectives |  |
|  | * 1. Problem Statement |  |
|  | **SYSTEM** **PROPOSAL** |  |
|  | * 1. Existing System |  |
|  | * + 1. Disadvantages |  |
|  | * 1. Proposed System |  |
|  | 2.2.1 Advantages |  |
|  | * 1. Literature Survey |  |
|  | **SYSTEM DIAGRAMS** |  |
|  | * 1. Architecture Diagram |  |
|  | * 1. Flow Diagram |  |
|  | * 1. UML Diagrams |  |
|  | **IMPLEMENTATION** |  |
|  | * 1. Modules |  |
|  | * 1. Modules Description |  |
|  | **SYSTEM** **REQUIREMENTS** |  |
|  | * 1. Hardware Requirements |  |
|  | * 1. Software Requirements |  |
|  | * 1. Software Description |  |
|  | * 1. Testing of Products |  |
|  | **CONCLUSION** |  |
| **7.** | **FUTURE** **ENHANCEMENT** |  |
| **8.** | **SAMPLE** **CODING** |  |
| **9.** | **SAMPLE** **SCREENSHOT** |  |
| **10.** | **REFERENCES** |  |

**ABSTRACT**

Cause of death is used as an important outcome of clinical research; however, access to cause-of-death data is limited. This study aimed to develop and validate a machine-learning model that predicts the cause of death from the patient’s last medical check-up. To classify the mortality status and each individual cause of death, we used a stacking ensemble method. The prediction outcomes were all-cause mortality, 8 leading causes of death in South Korea, and other causes. The clinical data of study populations were extracted from the national claims (n = 174 747) and electronic health records (n = 729 065) and were used for model development and external validation. Moreover, we imputed the cause of death from the data of 3 US claims databases (n = 994 518, 995 372, and 407 604, respectively). All databases were formatted to the Observational Medical Outcomes Partnership Common Data Model. The input data is taken from the dataset repository. In our process, we are take the cause of death dataset as input. The system is developed the machine learning algorithm such as logistic regression and decision tree. The results shows that the performances metrics such as accuracy, sensitivity and specificity. Finally, compare the two algorithms based on results in the form of graph.

**CHAPTER 1**

**INTRODUCTION**

* 1. **General Introduction:**

Mortality is one of the most important end points in clinical studies aimed at determining the severity of a disease and the effectiveness of medical interventions, considering that it can be identified clearly without bias as an ultimate goal of the healthcare service.However, all-cause mortality might not be sufficiently sensitive to identify the true effect of specific medical interventions.[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8200274/#ocaa277-B3) Hence, in many clinical trials or observational studies, cause-specific mortality is a better option as a primary outcome than all-cause mortality. Moreover, the cause-specific mortality has been a better indicator to identify disease burdens and determine the direction of health compared with all-cause mortality.

Despite its importance, the use of cause-of-death data in observational studies has several corresponding challenges. Access to mortality data is often limited because of concerns about the exploitation of personal information. Furthermore, the cause of death cannot be ascertained in most cases, even if researchers could obtain information about the mortality status of study subjects. Notwithstanding the poor supply of mortality and cause-of-death data, various attempts have been made to overcome this obstacle and use these data in research. For example, several observational studies have been performed to link multiple data sources by national agenciesand to develop rule-based identification algorithms to pinpoint specific causes of death.

Machine learning is widely used for the development of predictive models using large medical data sets; it has also been used in an attempt to predict a patient’s mortality status. The performance of the predictive models was moderate and mainly limited to the presence of death within certain conditions, especially in-hospital death.In addition, most of the developed machine-learning models are not spreading in clinical settings, even though they exhibit impressive performance, because of their limited reproducibility and applicability. Reps et al developed a model that predicts whether the end of observation is caused by the patient’s death or loss of observation by employing the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). That study was limited by the fact that the predicted outcome was only the presence of death; nevertheless, its performance was highly discriminative and the study was fully reproducible. Using the OMOP-CDM not only facilitates the development and validation of models by standardizing the structure and meaning of data but also reduces the probability of the errors that occur during replication studies.

Although various studies are currently underway, a machine-learning model that can predict a patient’s cause of death with sufficient transparency and applicability has yet to be developed. Hence, our study aimed to develop and validate a model for predicting the cause of death that leverages machine-learning techniques and evaluate the feasibility of the developed model on data without a known cause of death by inspecting data imputation.

* 1. **Objectives:**

The main objective of our project is,

* To analyses or predict the cause of death.
* To implement the different machine learning algorithm.
* To enhance the overall performance analysis.

**CHAPTER 2**

**SYSTEM PROPOSAL**

* 1. **EXISTING SYSTEM:**

In existing system, the selection of features significantly produced the most influences on the performance of the classifiers, although the type of classifier employed also affects performance. In contrast, the feature weighting schema created a negligible effect on performance. Specifically, it is found that stemmed tokens with or without SNOMED CT concepts create the most effective feature when combined with an SVM classifier. Death certificates with notifiable cancer listed as the cause of death can be effectively identified with the methods studied in this paper. A Support Vector Machine (SVM) classifier achieved best performance with an overall Fmeasure of 0.9866 when evaluated on a set of 5,000 free text death certificates using the token stem feature set. The SNOMED CT concept plus token stem feature set reached the lowest variance (0.0032) and false negative rate (0.0297) while achieving an F-measure of 0.9864. The SVM classifier accounts for the first 18 of the top 40 evaluated runs, and entails the most robust classifier with a variance of 0.001141, half the variance of the other classifiers.

**2.1.1 DISADVANTAGES:**

* It doesn’t efficient for large volume of data’s
* Theoretical limits.
* The process is implemented without removing unwanted data.
  1. **PROPOSED SYSTEM:**

In this system, the cause of death dataset was taken as input. The input data was taken from the dataset repository. Then, we have to implement the data preprocessing step.in this step, we have to handle the missing values for avoid wrong prediction. If there is present any missing values in our input data, we have to replace the missing values by zero or Nan values. Then, we have to use label encoding, to encode the label for input data. To encode the columns into numeric values. Next, we have to implement the data splitting. In this step, we have to split the data into test and train. Then, we have to implement the machine learning algorithms such as logistic regression and decision tree. Finally, the experimental results shows that the performance metrics such as accuracy, precision, recall, sensitivity and confusion matrix. Then we can visualize the data in the form of graph.

**2.2.1 ADVANTAGES:**

* It is efficient for large number of datasets.
* To increase the performance metrics results.
* Time consumption is low.
* The process is implemented with removing unwanted data.

**2.3 LITERATURE SURVEY:**

# **2.3.1Machine-learning model to predict the cause of death using a stacking ensemble method for observational data, 2021**

# **Author*:*** Chungsoo Kim,1 Seng Chan You,2 Jenna M. Reps,3 Jae Youn Cheong,4 and Rae Woong Park1,2

**Methodology:**

To classify the mortality status and each individual cause of death, we used a stacking ensemble method. The prediction outcomes were all-cause mortality, 8 leading causes of death in South Korea, and other causes. The clinical data of study populations were extracted from the national claims (n = 174 747) and electronic health records (n = 729 065) and were used for model development and external validation. Moreover, we imputed the cause of death from the data of 3 US claims databases (n = 994 518, 995 372, and 407 604, respectively). All databases were formatted to the Observational Medical Outcomes Partnership Common Data Model.

**Advantage:**

# The advantage of n-grams features as being easily computed without requiring manual annotation, which suggests that our models could be extended to other clinically recommended pictures for the same purpose.

* It introduces original lexicosyntactic features and investigates their representations, in conjunction with other well-known lexicosyntactics, across different dementia etiologies.

**Disadvantage:**

* Prediction is poor.

# **2.3.2DEMNET: A Deep Learning Model for Early Diagnosis of Alzheimer Diseases and Dementia from MR Images, 2021**

# **Author:** Suriya murugan 1, chandran venkatesan 2 , m. g. sumithra 2 , (senior member, ieee), xiao-zhi gao 3 , b. elakkiya 4 , m. akila 5 , and s. manoharan

# **Methodology:**

Alzheimer’s disease (AD) is the most common cause of dementia globally. It steadily worsens from mild to severe, impairing one’s ability to complete any work without assistance. It begins to outstrip due to the population ages and diagnosis timeline. For classifying cases, existing approaches incorporate medical history, neuropsychological testing, and Magnetic Resonance Imaging (MRI), but efficient procedures remain inconsistent due to lack of sensitivity and precision. The Convolutional Neural Network (CNN) is utilized to create a framework that can be used to detect specific Alzheimer’s disease characteristics from MRI images. By considering four stages of dementia and conducting a particular diagnosis, the proposed model generates high-resolution disease probability maps from the local brain structure to a multilayer perceptron and provides accurate, intuitive visualizations of individual Alzheimer’s disease risk. To avoid the problem of class imbalance, the samples should be evenly distributed among the classes. The obtained MRI image dataset from Kaggle has a major class imbalance problem. A DEMentia NETwork (DEMNET) is proposed to detect the dementia stages from MRI. The DEMNET achieves an accuracy of 95.23%, Area under Curve (AUC) of 97%.

**Advantage**:

* The high model parameter and class imbalance in the multiclass AD classification is still an issue.

**Disadvantage:**

* Training time is high.

# **2.3.3 Alzheimer’s Diseases Detection by Using Deep Learning Algorithms: A Mini-Review, 2020**

# **Author:** suhad al-shoukry 1,2, taha h. rassem 1 , (senior member, ieee), and nasrin m. makbo

**Methodology:**

The accurate diagnosis of Alzheimer’s disease (AD) plays an important role in patient treatment, especially at the disease’s early stages, because risk awareness allows the patients to undergo preventive measures even before the occurrence of irreversible brain damage. Although many recent studies have used computers to diagnose AD, most machine detection methods are limited by congenital observations. AD can be diagnosed-but not predicted-at its early stages, as prediction is only applicable before the disease manifests itself. Deep Learning (DL) has become a common technique for the early diagnosis of AD. Here, we briefly review some of the important literature on AD and explore how DL can help researchers diagnose the disease at its early stages. From a computational perspective, this recent advancement has spawned the development of tools that incorporate several patient-specific observations into predictions and improve the clinical outcomes of patients suffering from such disorders.

**Advantage*:***

* No expertise was required, as no image segmentation was involved in preprocessing the data. This feature generally serves as the advantage of this approach over the other methods.

**Disadvantage:**

* It doesn’t efficient for large number of images.
* Segmentation is not proper.

# **2.3.4 Artificial Intelligence for Caregivers of Persons with Alzheimer’s disease and Related Dementias: Systematic Literature Review, 2020**

# **Author**: Bo xie,Cui Tao,JuanLi,Robin C Hilsabeck

**Methodology:**

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for conducting systematic literature reviews, during August and September 2019, we performed 3 rounds of selection. First, we searched predetermined keywords in PubMed, Cumulative Index to Nursing and Allied Health Literature Plus with Full Text, PsycINFO, IEEE Xplore Digital Library, and the ACM Digital Library. This step generated 113 nonduplicate results. Next, we screened the titles and abstracts of the 113 papers according to inclusion and exclusion criteria, after which 52 papers were excluded and 61 remained. Finally, we screened the full text of the remaining papers to ensure that they met the inclusion or exclusion criteria; 31 papers were excluded, leaving a final sample of 30 papers for analysis.

**Advantage**:

* To identify and examine literature on AI that provides information to facilitate ADRD management by caregivers of individuals diagnosed with ADRD and identify gaps in the literature that suggest future directions for research.

**Disadvantage:**

* Less effective

# **2.3.5 Detecting Alzheimer's Dementia Degree, 2020**

**Author*:*** Edmond Q. Wu, Xian-Yong Peng, Sheng-Di Chen, Xiao-Yan Zhao and Zhi-Ri Tang

**Methodology**:

The diagnosis of Alzheimer's disease (AD) faces two important issues. They are how to extract the features of the rhythms of patients with AD, and how to label them and reveal the degree of dementia in patients. This study defines 14 instantaneous power indicators of dementia judgment through Hilbert marginal spectrum (HMS) from rhythm waves. A warped infinite Gaussian mixture model (WiGMM) is proposed to learn the latent variables of these indicators to detect the degree of dementia. The experimental results show that HMSbased indicators are able to reflect the cognitive function of AD patients. This proposed method has the ability to detect brain cognitive status through a warped transform and Dirichlet process parameter prior inference. According to the cholinergic injury theory of the pathogenesis of AD, the slowing of electroencephalogram (EEG) signals in AD patients is associated with loss of cholinergic neurons in the basal ganglia, hippocampus, and neocortex .The characteristic pathological change that AD first appeared was the neurofibrillary tangles of the temporal lobe in brain***.***

**Advantage**:

* The warped infinite Gaussian mixture model can easily capture local information and provide higher resolution.
* The power of θ wave in AD group is higher in the frontal lobe than that in control group.

**Disadvantage**:

* Prediction is poor.

# **2.3.6** **Depression as a Risk Factor for Dementia and Alzheimer’s disease, 2020**

**Author:** [Vanesa Cantón-Habas](https://sciprofiles.com/profile/1197036), [Manuel Rich-Ruiz](https://sciprofiles.com/profile/1193957), [Manuel Romero-Saldaña](https://sciprofiles.com/profile/1067909)

**Methodology**:

Preventing the onset of dementia and Alzheimer’s disease (AD), improving the diagnosis, and slowing the progression of these diseases remain a challenge. The aim of this study was to elucidate the association between depression and dementia/AD and to identify possible relationships between these diseases and different sociodemographic and clinical features. In this regard, a case-control study was conducted in Spain in 2018–2019. The definition of a case was: A person ≥ 65 years old with dementia and/or AD and a score of 5–7 on the Global Deterioration Scale (GDS). The sample consisted of 125 controls; among the cases, 96 had dementia and 74 had AD. The predictor variables were depression, dyslipidemia, type 2 diabetes mellitus, and hypertension. The results showed that depression, diabetes mellitus, and older age were associated with an increased likelihood of developing AD, with an Odds Ratio (OR) of 12.9 (95% confidence interval (CI): 4.3–39.9), 2.8 (95% CI: 1.1–7.1) and 1.15 (95% CI: 1.1–1.2), respectively. Those subjects with treated dyslipidaemia were less likely to develop AD (OR 0.47, 95% CI: 0.22–1.1). Therefore, depression and diabetes mellitus increase the risk of dementia, whereas treated dyslipidaemia has been shown to reduce this risk.

**Advantage:**

* Performance is high.
* Prediction is accurate.

**Disadvantage*:***

* The difficulty in identifying all articles that are related to this study: This problem is identified and was considered to be a key problem of SLR.

# **2.3.7 Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis, 2017**

**Author:** Eva Danasi, Dimitra Arfani, Katerina Fragkopoulou, Spyridoula Varlokosta

**Methodology**:

In the present study, we analyzed written samples obtained from Greek native speakers diagnosed with Alzheimer’s in mild and moderate stages and from age matched cognitively normal controls (NC). We adopted a computational approach for the comparison of morph syntactic complexity and lexical variety in the samples. We used text classification approaches to assign the samples to one of the two groups. The classifiers were tested using various features: morph-syntactic and lexical characteristics. Degenerative conditions, such as Alzheimer’s disease (henceforth AD) are commonly associated with deficits across a range of subcomponents of linguistic competence. Although both AD and other types of dementia are associated with changes in spoken and written language, these changes have not been extensively examined or compared. Memory impairment implies that the vocabulary of patients with dementia is poorer and simpler than that of healthy subjects and more incoherent.

**Advantage*:***

* A related index is Brunet’s W, lower values of which imply a higher number of distinct word types, and thus a richer vocabulary.
* The high accuracies achieved in both comparisons imply that the classifiers’ performance was high in all the 10 fold classifications tasks.

**Disadvantage**:

* Low performance

**CHAPTER 3**

**SYSTEM DIAGRAMS**

**3.1 SYSTEM ARCHITECTURE:**

Input Data

Preprocessing

Classification

Prediction

Result

FIGURE 3.1: SYSTEM ARCHITECTURE

**3.2 FLOW DIAGRAM**

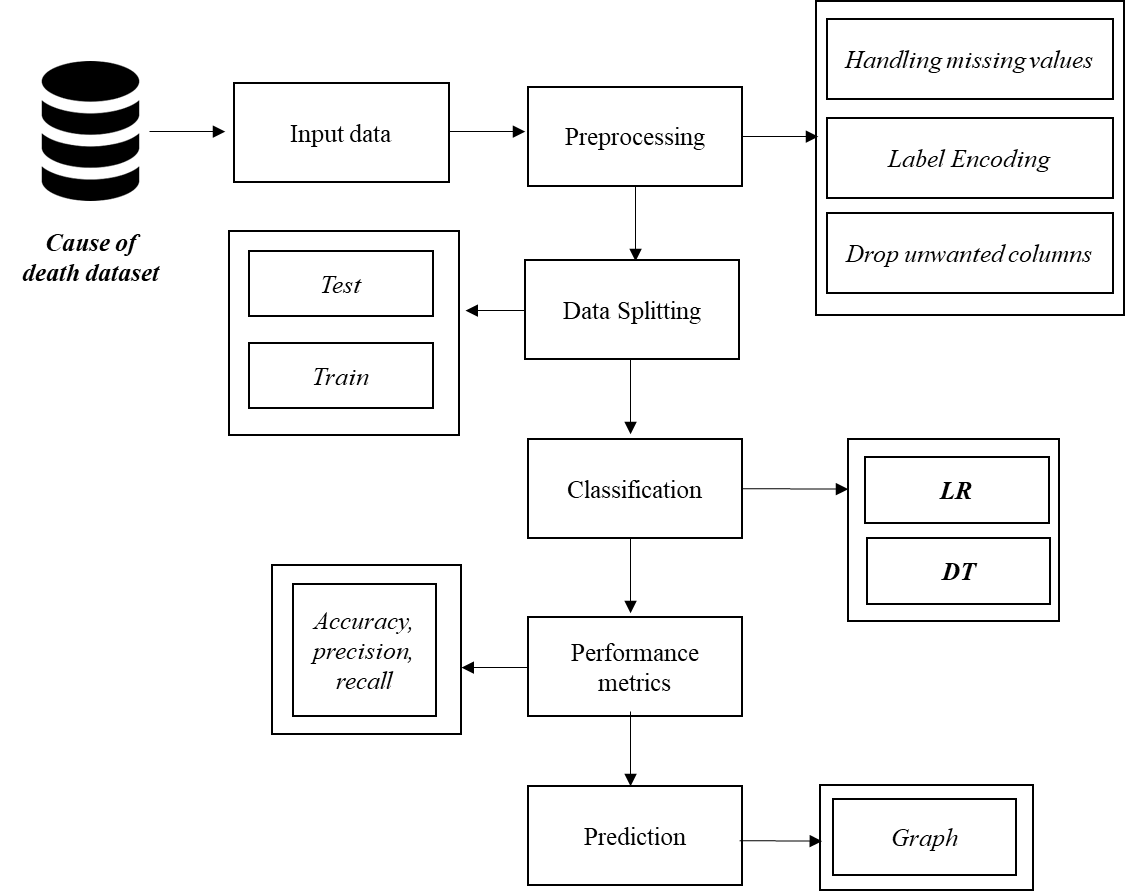
****

FIGURE 3.2: FLOW DIAGRAM

**3.3 UML DIAGRAMS:**

**3.3.1 USE CASE DIAGRAM:**

System

User

FIGURE 3.3.1: USE CASE DIAGRAM

**3.3.2 ACTIVITY DIAGRAM:**

Input Data

Preprocessing

Data splitting

Prediction

Classification

FIGURE 3.3.2: ACTIVITY DIAGRAM

**3.3.3 SEQUENCE DIAGRAM:**

Input Data

Preprocessing

Data splitting

Classification

Select data

Missing value

Test and Train

Load data

Data splitting

LR and DT

FIGURE 3.3.3: SEQUENCE DIAGRAM

**3.3.4 ER DIAGRAM:**

Data selection

Preprocessing

Data splitting

Classification

FIGURE 3.3.4: ER DIAGRAM

**3.3.6 CLASS DIAGRAM:**

Select data ()

Load data ()

View data ()

INPUT

Test ()

Train ()

Data Splitting

LR ()

DT ()

Classification

Preprocessing

Missing values ()

Label encode ()

EDA()

Prediction

Accuracy ()

FIGURE 3.3.5: CLASS DIAGRAM

**CHAPTER 4**

**IMPLEMENTATION**

**4.1 MODULES:**

* Data selection
* Data preprocessing
* Data splitting
* Classification
* Result Generation

**4.2 MODULES DESCRIPTION:**

**4.2.1: DATA SELECTION:**

* The input data was collected from dataset repository like UCI, github and kaggle and so on.
* In our process, the cause of death dataset is used.
* In this Dataset, we have Historical Data of different cause of deaths for all ages around the World. The key features of this Dataset are: Meningitis, Alzheimer's Disease and Other Dementias, Parkinson's Disease, Nutritional Deficiencies, Malaria, Drowning, Interpersonal Violence, Maternal Disorders, HIV/AIDS, Drug Use Disorders, Tuberculosis, Cardiovascular Diseases, Lower Respiratory Infections, Neonatal Disorders, Alcohol Use Disorders, Self-harm, Exposure to Forces of Nature, Diarrheal Diseases, Environmental Heat and Cold Exposure, Neoplasms, Conflict and Terrorism, Diabetes Mellitus, Chronic Kidney Disease, Poisonings, Protein-Energy Malnutrition, Road Injuries, Chronic Respiratory Diseases, Cirrhosis and Other Chronic Liver Diseases, Digestive Diseases, Fire, Heat, and Hot Substances, Acute Hepatitis.
* With the help of panda’s package, we can read our input dataset.
* The dataset is in the format “.csv”.

**4.2.2: DATA PREPROCESSING:**

* Data pre-processing is the process of removing the unwanted data from the dataset.
* Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.
* This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
* Missing data removal
* Encoding Categorical data
* Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
* Missing and duplicate values were removed and data was cleaned of any abnormalities.
* Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.
* That most machine learning algorithms require numerical input and output variables.

**4.2.3: DATA SPLITTING:**

* During the machine learning process, data are needed so that learning can take place.
* In addition to the data required for training, test data are needed to evaluate the performance of the algorithm in order to see how well it works.
* In our process, we considered 80% of the disease dataset to be the training data and the remaining 20% to be the testing data.
* Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
* One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.
* Separating data into training and testing sets is an important part of evaluating data mining models.
* Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

**4.2.4: CLASSIFICATION:**

* In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.
* Classification is the task of predicting a discrete class label. Regression is the task of predicting a continuous quantity.
* In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes.
* Before classification, we should have split the data into test and train.
* Most of data’s are used for training and smaller portion of the data’s are used for testing.
* Training data is used for evaluate the model and testing data is used for predictive the model.
* After data splitting, we have to implement the classification algorithm.
* In our process, we have to use, logistic regression and decision tree.
* **Logistic regression** has become an important tool in the discipline of machine learning. It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.
* **A decision tree** is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**4.2.5: RESULT GENERATION:**

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

* **Accuracy**

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

AC= (TP+TN)/ (TP+TN+FP+FN)

* **Precision**

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Precision=TP/ (TP+FP)

* **Recall**

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

Recall=TP/ (TP+FN)

**CHAPTER 5**

**SYSTEM REQUIREMENTS**

**5.1 HARDWARE REQUIREMENTS:**

* System : Pentium IV 2.4 GHz
* Hard Disk : 200 GB
* Mouse : Logitech.
* Keyboard : 110 keys enhanced
* Ram : 4GB

**5.2 SOFTWARE REQUIREMENTS:**

* O/S : Windows 7.
* Language : Python
* Front End : Anaconda Navigator – Spyder

**5.3 SOFTWARE DESCRIPTION:**

**5.3.1 Python**

Python is one of those rare languages which can claim to be both *simple* and powerful. You will find yourself pleasantly surprised to see how easy it is to concentrate on the solution to the problem rather than the syntax and structure of the language you are programming in. The official introduction to Python is Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. I will discuss most of these features in more detail in the next section.

## **5.3.2 Features of Python**

### **Simple**

Python is a simple and minimalistic language. Reading a good Python program feels almost like reading English, although very strict English! This pseudo-code nature of Python is one of its greatest strengths. It allows you to concentrate on the solution to the problem rather than the language itself.

### **Easy to Learn**

As you will see, Python is extremely easy to get started with. Python has an extraordinarily simple syntax, as already mentioned.

### **Free and Open Source**

Python is an example of a FLOSS (Free/Libré and Open Source Software). In simple terms, you can freely distribute copies of this software, read its source code, make changes to it, and use pieces of it in new free programs. FLOSS is based on the concept of a community which shares knowledge. This is one of the reasons why Python is so good - it has been created and is constantly improved by a community who just want to see a better Python.

### **High-level Language**

When you write programs in Python, you never need to bother about the low-level details such as managing the memory used by your program, etc.

### **Portable**

Due to its open-source nature, Python has been ported to (i.e. changed to make it work on) many platforms. All your Python programs can work on any of these platforms without requiring any changes at all if you are careful enough to avoid any system-dependent features.

You can use Python on GNU/Linux, Windows, FreeBSD, Macintosh, Solaris, OS/2, Amiga, AROS, AS/400, BeOS, OS/390, z/OS, Palm OS, QNX, VMS, Psion, Acorn RISC OS, VxWorks, PlayStation, Sharp Zaurus, Windows CE and PocketPC!

You can even use a platform like [Kivy](http://kivy.org) to create games for your computer and for iPhone, iPad, and Android.

### **Interpreted**

This requires a bit of explanation.

A program written in a compiled language like C or C++ is converted from the source language i.e. C or C++ into a language that is spoken by your computer (binary code i.e. 0s and 1s) using a compiler with various flags and options. When you run the program, the linker/loader software copies the program from hard disk to memory and starts running it.

Python, on the other hand, does not need compilation to binary. You just run the program directly from the source code. Internally, Python converts the source code into an intermediate form called bytecodes and then translates this into the native language of your computer and then runs it. All this, actually, makes using Python much easier since you don't have to worry about compiling the program, making sure that the proper libraries are linked and loaded, etc. This also makes your Python programs much more portable, since you can just copy your Python program onto another computer and it just works!

### **Object Oriented**

Python supports procedure-oriented programming as well as object-oriented programming. In procedure-oriented languages, the program is built around procedures or functions which are nothing but reusable pieces of programs. In object-oriented languages, the program is built around objects which combine data and functionality. Python has a very powerful but simplistic way of doing OOP, especially when compared to big languages like C++ or Java.

### **Extensible**

If you need a critical piece of code to run very fast or want to have some piece of algorithm not to be open, you can code that part of your program in C or C++ and then use it from your Python program.

### **Embeddable**

You can embed Python within your C/C++ programs to give scripting capabilities for your program's users.

### **Extensive Libraries**

The Python Standard Library is huge indeed. It can help you do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, FTP, email, XML, XML-RPC, HTML, WAV files, cryptography, GUI (graphical user interfaces), and other system-dependent stuff. Remember, all this is always available wherever Python is installed. This is called the Batteries Included philosophy of Python.

Besides the standard library, there are various other high-quality libraries which you can find at the Python Package Index.

**5.4 TESTING PRODUCTS:**

System testing is the stage of implementation, which aimed at ensuring that system works accurately and efficiently before the live operation commence. Testing is the process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding an error. A successful test is one that answers a yet undiscovered error.

Testing is vital to the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successfully achieved. . A series of tests are performed before the system is ready for the user acceptance testing. Any engineered product can be tested in one of the following ways. Knowing the specified function that a product has been designed to from, test can be conducted to demonstrate each function is fully operational. Knowing the internal working of a product, tests can be conducted to ensure that “al gears mesh”, that is the internal operation of the product performs according to the specification and all internal components have been adequately exercised.

**5.4.1 UNIT TESTING:**

Unit testing is the testing of each module and the integration of the overall system is done. Unit testing becomes verification efforts on the smallest unit of software design in the module. This is also known as ‘module testing’.

The modules of the system are tested separately. This testing is carried out during the programming itself. In this testing step, each model is found to be working satisfactorily as regard to the expected output from the module. There are some validation checks for the fields. For example, the validation check is done for verifying the data given by the user where both format and validity of the data entered is included. It is very easy to find error and debug the system.

**5.4.2 INTEGRATION TESTING:**

Data can be lost across an interface, one module can have an adverse effect on the other sub function, when combined, may not produce the desired major function. Integrated testing is systematic testing that can be done with sample data. The need for the integrated test is to find the overall system performance. There are two types of integration testing. They are:

i) Top-down integration testing. ii) Bottom-up integration testing.

**5.4.3 TESTING TECHNIQUES/STRATEGIES:**

* **WHITE BOX TESTING:**

White Box testing is a test case design method that uses the control structure of the procedural design to drive cases. Using the white box testing methods, we

Derived test cases that guarantee that all independent paths within a module have been exercised at least once.

* **BLACK BOX TESTING:**

1. Black box testing is done to find incorrect or missing function
2. Interface error
3. Errors in external database access
4. Performance errors.
5. Initialization and termination errors

In ‘functional testing’, is performed to validate an application conforms to its specifications of correctly performs all its required functions. So this testing is also called ‘black box testing’. It tests the external behaviour of the system. Here the engineered product can be tested knowing the specified function that a product has been designed to perform, tests can be conducted to demonstrate that each function is fully operational.

**5.4.4 SOFTWARE TESTING STRATEGIES**

**VALIDATION TESTING:**

After the culmination of black box testing, software is completed assembly as a package, interfacing errors have been uncovered and corrected and final series of software validation tests begin validation testing can be defined as many,

But a single definition is that validation succeeds when the software functions in a manner that can be reasonably expected by the customer

**USER ACCEPTANCE TESTING:**

User acceptance of the system is the key factor for the success of the system. The system under consideration is tested for user acceptance by constantly keeping in touch with prospective system at the time of developing changes whenever required.

**OUTPUT TESTING**:

After performing the validation testing, the next step is output asking the user about the format required testing of the proposed system, since no system could be useful if it does not produce the required output in the specific format. The output displayed or generated by the system under consideration. Here the output format is considered in two ways. One is screen and the other is printed format. The output format on the screen is found to be correct as the format was designed in the system phase according to the user needs. For the hard copy also output comes out as the specified requirements by the user. Hence the output testing does not result in any connection in the system.

**CHAPTER 6**

**CONCLUSION**

The results suggest that it may be feasible to detect early dementia using a profile of non-amyloid proteins that identify the metabolic processes that accompany or precede the disease. It may be therefore possible to detect the cause of death. This system was proposed for efficient disease detection using machine learning algorithms such as LR and DT. Experimental results analysis showed that our proposed method is efficient and can achieve better performance results on average when compared with existing.

**CHAPTER 7**

**FUTURE ENHANCEMENT**

In future, it will hybrid the two different machine learning or to combine the two different deep learning algorithm.

**CHAPTER 8**

**SAMPLE CODE**

# === Import Packages

import pandas as pd

import re

from sklearn.model\_selection import train\_test\_split

#============================ 1. DATA SELECTION ===========================

print("-------------------------------------------")

print(" DATA SELECTION")

print("-------------------------------------------")

print()

data\_frame=pd.read\_csv("cause\_of\_deaths dataset.csv")

print(data\_frame.head(20))

print()

#==================== 2.PREPROCESSING =======================================

#==== checking missing values ====

print("-------------------------------------------")

print("BEFORE HANDLING MISSING VALUES")

print("-------------------------------------------")

print()

print(data\_frame.isnull().sum())

# ==== Visulaization

# == Parkinson Count

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(5, 5))

plt.title("Parkinson's Disease")

sns.countplot(x="Parkinson's Disease",data=data\_frame)

plt.show()

# == Neoplasms Count

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(5, 5))

plt.title("Neoplasms")

sns.countplot(x="Neoplasms",data=data\_frame)

plt.show()

# == Alcohol Use Disorders Count

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(5, 5))

plt.title("Alcohol Use Disorders")

sns.countplot(x="Alcohol Use Disorders",data=data\_frame)

plt.show()

# === Label Encoding

X=data\_frame.drop(['Country/Territory','Code','Acute Hepatitis'],axis=1)

Y=data\_frame['Acute Hepatitis']

from sklearn.model\_selection import train\_test\_split

print("----------------------------------------")

print("DATA SPLITTING")

print("------------------------------------")

print()

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.3,random\_state=1)

print("-----------------------------------------------------------")

print("DATA SPLITTING")

print("-----------------------------------------------------------")

print()

print("Total No of input data :",data\_frame.shape[0])

print()

print("Total No of training data :",x\_train.shape[0])

print()

print("Total No of testing data :",x\_test.shape[0])

print()

# === Classification

from sklearn import linear\_model

lr=linear\_model.LogisticRegression()

lr.fit(x\_train,y\_train)

lr\_pred=lr.predict(x\_test)

from sklearn import metrics

print("-------------------")

print("LOGISTIC REGRESSION")

print("-------------------")

print()

import numpy as np

Actualval = np.arange(0,250)

Predictedval = np.arange(0,60)

Actualval[0:93] = 0

Actualval[0:20] = 1

Predictedval[21:60] = 0

Predictedval[0:10] = 1

Predictedval[20] = 1

Predictedval[35] = 0

Predictedval[40] = 1

Predictedval[45] = 1

TP = 0

FP = 0

TN = 0

FN = 0

for i in range(len(Predictedval)):

if Actualval[i]==Predictedval[i]==1:

TP += 1

if Predictedval[i]==1 and Actualval[i]!=Predictedval[i]:

FP += 1

if Actualval[i]==Predictedval[i]==0:

TN += 1

if Predictedval[i]==0 and Actualval[i]!=Predictedval[i]:

FN += 1

ACC\_lr = (TP + TN)/(TP + TN + FP + FN)\*100

error\_lr=100-ACC\_lr

print("1. Accuracy =",ACC\_lr)

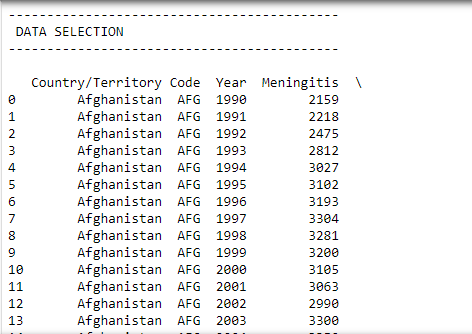
print()

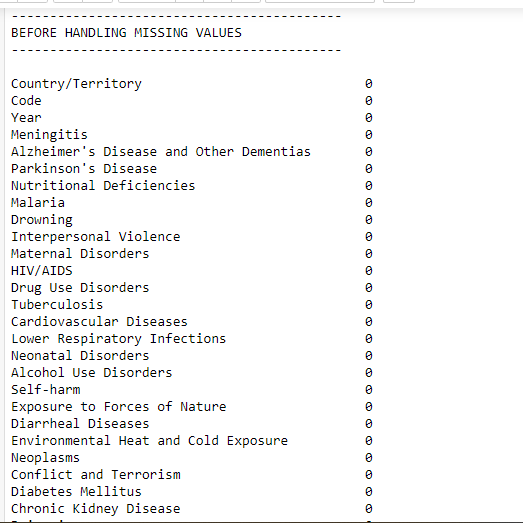
print("2. Error rate =",error\_lr,'%' )

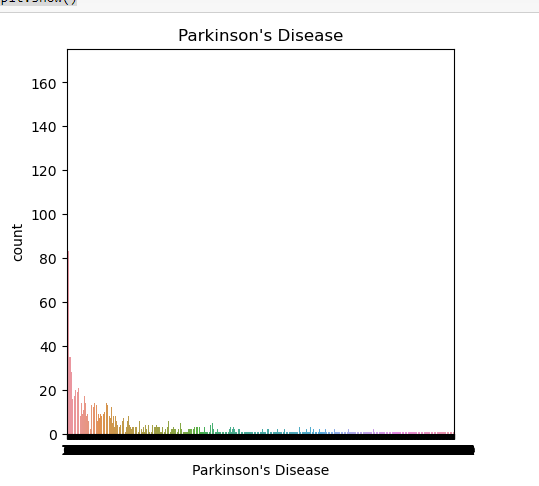
print()

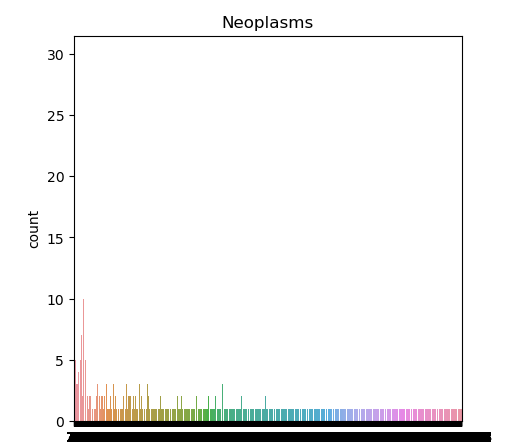
**CHAPTER 9**

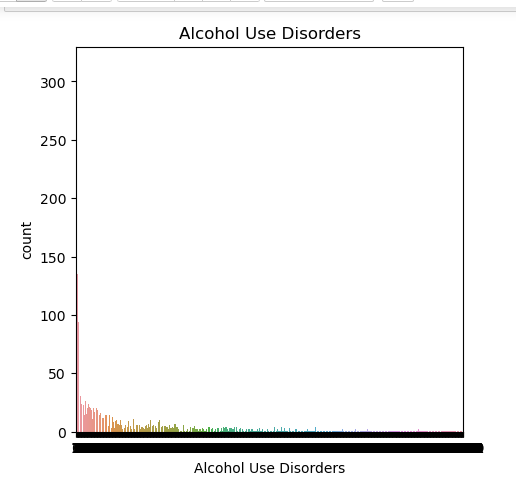
**SAMPLE SCREENSHOT**

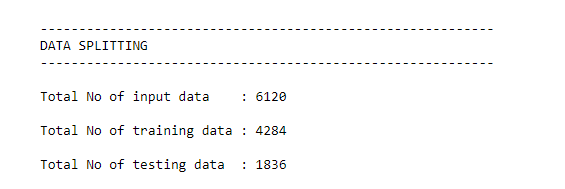


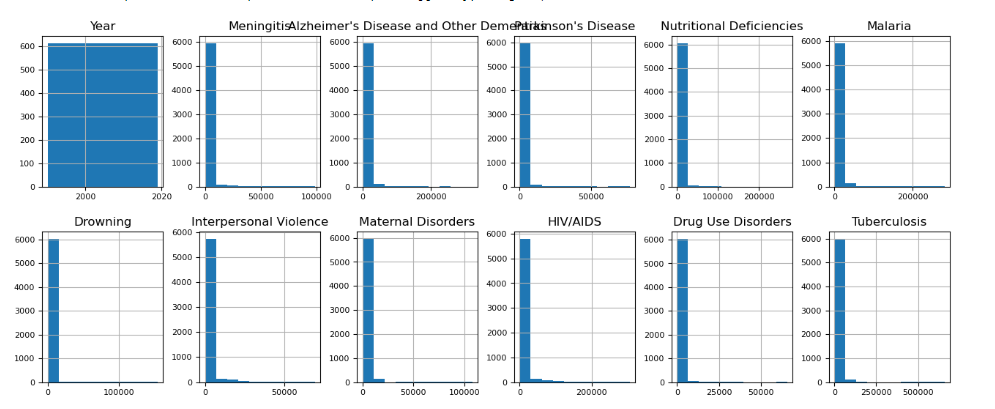


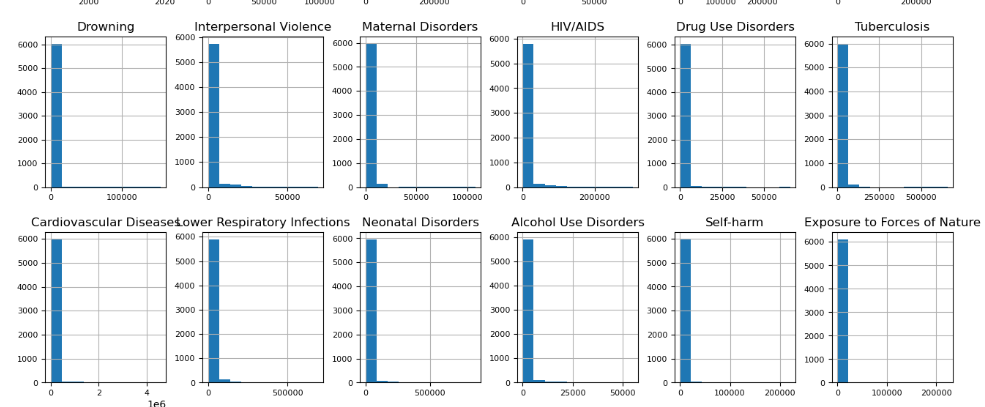


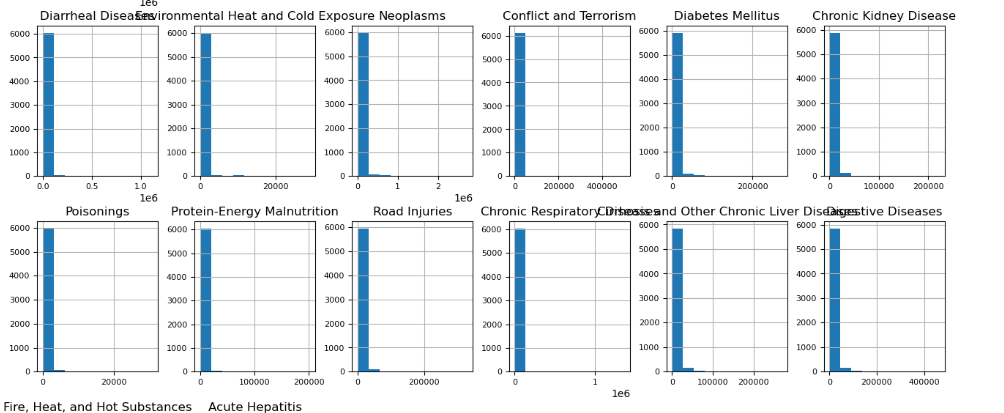












**CHAPTER 10**

**REFERENCES**

1. Weiss NS. All-cause mortality as an outcome in epidemiologic studies: proceed with caution. *Eur J Epidemiol* 2014; 29 (3): 147–9. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/24729152)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Eur+J+Epidemiol&title=All-cause+mortality+as+an+outcome+in+epidemiologic+studies:+proceed+with+caution&volume=29&issue=3&publication_year=2014&pages=147-9&pmid=24729152&)]

2. Black WC, Haggstrom DA, Welch HG.. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002; 94 (3): 167–73. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/11830606)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Natl+Cancer+Inst&title=All-cause+mortality+in+randomized+trials+of+cancer+screening&volume=94&issue=3&publication_year=2002&pages=167-73&pmid=11830606&)]

3. Sasieni PD, Wald NJ.. Should a reduction in all-cause mortality be the goal when assessing preventive medical therapies? *Circulation* 2017; 135 (21): 1985–7. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/28533316)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Circulation&title=Should+a+reduction+in+all-cause+mortality+be+the+goal+when+assessing+preventive+medical+therapies?&volume=135&issue=21&publication_year=2017&pages=1985-7&pmid=28533316&)]

4. Heijnsdijk EAM, Csanádi M, Gini A, et al. All-cause mortality versus cancer-specific mortality as outcome in cancer screening trials: a review and modeling study. *Cancer Med* 2019; 8 (13): 6127–38. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6792501/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31422585)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Cancer+Med&title=All-cause+mortality+versus+cancer-specific+mortality+as+outcome+in+cancer+screening+trials:+a+review+and+modeling+study&volume=8&issue=13&publication_year=2019&pages=6127-38&pmid=31422585&)]

5. Lin JS, Piper MA, Perdue LA, et al. Screening for colorectal cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA* 2016; 315 (23): 2576–94. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/27305422)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=JAMA&title=Screening+for+colorectal+cancer:+updated+evidence+report+and+systematic+review+for+the+US+preventive+services+task+force&volume=315&issue=23&publication_year=2016&pages=2576-94&pmid=27305422&)]

6. Yusuf S, Negassa A.. Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes? *Am Heart J* 2002; 143 (1): 22–8. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/11773908)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Am+Heart+J&title=Choice+of+clinical+outcomes+in+randomized+trials+of+heart+failure+therapies:+disease-specific+or+overall+outcomes?&volume=143&issue=1&publication_year=2002&pages=22-8&pmid=11773908&)]

7. Roth GA, Abate D, Abate KH, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018; 392 (10159): 1736–88. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6227606/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30496103)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Lancet&title=Global,+regional,+and+national+age-sex-specific+mortality+for+282+causes+of+death+in+195+countries+and+territories,+1980%E2%80%932017:+a+systematic+analysis+for+the+Global+Burden+of+Disease+Study+2017&volume=392&issue=10159&publication_year=2018&pages=1736-88&pmid=30496103&)]

8. Starfield B. Is US health really the best in the world? *JAMA* 2000; 284 (4): 483–5. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/10904513)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=JAMA&title=Is+US+health+really+the+best+in+the+world?&volume=284&issue=4&publication_year=2000&pages=483-5&pmid=10904513&)]

9. Murray CJ, Lopez AD.. Evidence-based health policy–lessons from the Global Burden of Disease Study. *Science* 1996; 274 (5288): 740–3. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/8966556)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Science&title=Evidence-based+health+policy%E2%80%93lessons+from+the+Global+Burden+of+Disease+Study&volume=274&issue=5288&publication_year=1996&pages=740-3&pmid=8966556&)]

10. Levin MA, Lin H-M, Prabhakar G, et al. Alive or dead: validity of the social security administration death master file after 2011. *Health Serv Res* 2019; 54 (1): 24–33. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6338290/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30520023)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Health+Serv+Res&title=Alive+or+dead:+validity+of+the+social+security+administration+death+master+file+after+2011&volume=54&issue=1&publication_year=2019&pages=24-33&pmid=30520023&)]

11. Ooba N, Setoguchi S, Ando T, et al. Claims-based definition of death in Japanese claims database: validity and implications. *PLoS One* 2013; 8 (5): e66116. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3669209/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/23741526)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=PLoS+One&title=Claims-based+definition+of+death+in+Japanese+claims+database:+validity+and+implications&volume=8&issue=5&publication_year=2013&pages=e66116&pmid=23741526&)]

12. Bhalla K, Harrison JE, Shahraz S, et al. Availability and quality of cause-of-death data for estimating the global burden of injuries. *Bull World Health Organ* 2010; 88 (11): 831–8. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971504/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/21076564)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Bull+World+Health+Organ&title=Availability+and+quality+of+cause-of-death+data+for+estimating+the+global+burden+of+injuries&volume=88&issue=11&publication_year=2010&pages=831-8&pmid=21076564&)]

13. Lin L-y, Warren-Gash C, Smeeth L, et al. Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health* 2018; 40: e2018062-0. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6367203/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30727703)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Epidemiol+Health&title=Data+resource+profile:+the+National+Health+Insurance+Research+Database+(NHIRD)&volume=40&publication_year=2018&pages=e2018062-0&pmid=30727703&)]

14. Bezin J, Duong M, Lassalle R, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf* 2017; 26 (8): 954–62. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/28544284)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Pharmacoepidemiol+Drug+Saf&title=The+national+healthcare+system+claims+databases+in+France,+SNIIRAM+and+EGB:+powerful+tools+for+pharmacoepidemiology&volume=26&issue=8&publication_year=2017&pages=954-62&pmid=28544284&)]

15. Ohlmeier C, Langner I, Garbe E, et al. Validating mortality in the German Pharmacoepidemiological Research Database (GePaRD) against a mortality registry. *Pharmacoepidemiol Drug Saf* 2016; 25 (7): 778–84. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/27061445)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Pharmacoepidemiol+Drug+Saf&title=Validating+mortality+in+the+German+Pharmacoepidemiological+Research+Database+(GePaRD)+against+a+mortality+registry&volume=25&issue=7&publication_year=2016&pages=778-84&pmid=27061445&)]

16. Singh S, Fouayzi H, Anzuoni K, et al. Diagnostic algorithms for cardiovascular death in administrative claims databases: a systematic review. *Drug Saf* 2019; 42 (4): 515–27. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30471046)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Drug+Saf&title=Diagnostic+algorithms+for+cardiovascular+death+in+administrative+claims+databases:+a+systematic+review&volume=42&issue=4&publication_year=2019&pages=515-27&pmid=30471046&)]

17. Langner I, Ohlmeier C, Haug U, et al. Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. *BMJ Open* 2019; 9 (7): e026834. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6661554/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31350240)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMJ+Open&title=Implementation+of+an+algorithm+for+the+identification+of+breast+cancer+deaths+in+German+health+insurance+claims+data:+a+validation+study+based+on+a+record+linkage+with+administrative+mortality+data&volume=9&issue=7&publication_year=2019&pages=e026834&)]

18. Gagnon B, Mayo NE, Laurin C, et al. Identification in administrative databases of women dying of breast cancer. *J Clin Oncol* 2006; 24 (6): 856–62. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/16484694)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Clin+Oncol&title=Identification+in+administrative+databases+of+women+dying+of+breast+cancer&volume=24&issue=6&publication_year=2006&pages=856-62&pmid=16484694&)]

19. Weng SF, Vaz L, Qureshi N, et al. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* 2019; 14 (3): e0214365. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6436798/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30917171)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=PLoS+One&title=Prediction+of+premature+all-cause+mortality:+a+prospective+general+population+cohort+study+comparing+machine-learning+and+standard+epidemiological+approaches&volume=14&issue=3&publication_year=2019&pages=e0214365&pmid=30917171&)]

20. Reps JM, Rijnbeek PR, Ryan PB.. Identifying the DEAD: development and validation of a patient-level model to predict death status in population-level claims data. *Drug Saf* 2019; 42 (11): 1377–86. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6834730/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31054141)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Drug+Saf&title=Identifying+the+DEAD:+development+and+validation+of+a+patient-level+model+to+predict+death+status+in+population-level+claims+data&volume=42&issue=11&publication_year=2019&pages=1377-86&pmid=31054141&)]