Machine Learning Engineer Nanodegree

Capstone Report

Guillermo Aure
April 9, 2017

I. Definition

Project Overview

Doctors in the health industry consider obesity a risk factor for multiple types of illness. Conditions like diabetes type 2, hypertension and cardiovascular diseases and various types of tumors, in overweight patients [1] are associated with higher rates of mortality. It seems clear that obesity has a negative impact on humans.

In the medical field, there is a hypothesis know as The obesity paradox [2]. The obesity paradox hypothesis state that overweight elderly patients have a better chance to survive illness like the ones describes in the previous paragraph than underweight patients. Multiples studies try to proof or invalidate such hypothesis [3, 4, 8]. Some of the studies have found evidence that there is indeed some time of protection granted by a high BMI (body mass index) and others argue that confounding variables and lack of follow up invalidate such proofs [5]. One thing is sure, that despite which study is the correct all of them prove one way or another that there is a correlation between the: BMI, the age, the medical condition and the smoking habits; and that this association influence negative or positive the resilient capacity of a human being.

This project was based on the background knowledge of the correlation of the features described in the previous paragraph; it should be possible to train a learner to predict how much time a patient will stay hospitalized, based in the afford mentioned features.
To prove the above paragraph hypothesis, we will create a machine learning model to predict how long a patient will be hospitalized based on the mentioned features. To train the model, we will use the available public MIMI-III database. MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with 40,000 critical care patients.
Patient electronic record systems tend to store patient data based on its visit to the hospital; of all the possible visits information for the purpose of this project, we will use only the first.

Problem Statement

Preoperative evaluation purpose is not to explain patients for elective surgery, but rather to put in place controls that minimize the risk of potential surgery complications [6]. As was describe in the Project Overview on this document obesity as a general rule is consider

a surgery risk factor, and as consequence hospital tries to prepare patients for surgery, so they dont come to the procedure overweight. The problem with this approach is that considers overweight as a factor risk and not as a potential benefit. The obesity paradox hypothesis opens the door to a more custom preoperative evaluation and a possible reduction of the time a patient that gets complicated during a high-risk surgery will spend on the ICU unit. The key indicator here is the time spent hospitalized for surgery, the elapse time indicator, represents the resilient condition on a patient body based on the features describe before, in other words, the label of our predictor.
The elapsed time is a multivariable function; it depends on several features including the BMI. "The problem statement is to build a model to predict such time using the patient's electronic medical record (EHR.)" The problem is we humans have problems visualizing beyond three variables. An EHR has more than three variables; this represents a challenge for a person and an opportunity for the utilization of computers and a Machine Learning algorithms. The EHR is an excellent source of information, but the different types of data (continue and categorical) will require to encode and transform some of its features. Because the elapsed time is in nature continues, the model needs to predict also a no discrete output, because of this our problem requires a regression learner.
Details of the type of input and model that we will use are explained in the following sections of this report.
The expected solution is a machine learning model that can predict to a certain degree of accuracy how many days a patient diagnostic with a condition during its admission time will stay in the hospital.

Datasets and Inputs

In order to train and test our predictor we will use some of the feature of the MIMIC-III (MIT Lab for Computational Physiology) dataset. The MIMIC-III dataset contains more than 40,000 critical patients records. Information and details related to it can be found in the web site `https://mimic.physionet.org`, this is not an open source data a access must be requested. I will select the sample of the dataset by answering the questions documented in the solution statement of this documents. The resulting sample size after filtering the dataset is of 6862 patients. I selected the sample of the dataset by answering the questions documented in the solution statement of this documents. The resulting sample size after filtering the dataset is of 6862 patients. Along with the project files, there is an ECLIPSE BIRT workbench report design file; this file can be imported back to an ECLIPSE BIRT installation to recreate the dataset used in this project. The only modification to the BIRT report design file is the "data source" object that has to be pointed to an instance of the MIMICIII database. The file can be opened in any text editor to see the SQL queries used to create the dataset. It is importance to note that the final data set was created using BIRT joins, so I extracted three separated data sets using PostgreSQL queries and joined them using BIRT; again the dataset can be recreated using the companion BIRT report design file.
The sampled dataset will be reduce (some features will be calculate from some of the dataset values) to the following features:

FEATURES

- Body Mass Index: CHARTEVENTS table
- Sub Set of Medical Conditions: DIAGNOSESICD table
- Age: PATIENT table
- Sex: PATIENT table

LABEL

- Admission Time Discharge Time (MIMIC admissions table): ADMISSION table

NOTE: The "Body Mass Index"needs to be calculated from the heigh and weight values of the patients data. The formula we will use is:

$$BMI = WeightKg/(Heightmtrs)^2 \tag{1}$$

The features were selected based on the medical studies related to the Obesity Paradox.' The selected label is continue,' as most of the features values; some of them like the conditions (ICD9) and sex is categorical data and will be encoding or converted into dummy variables.

Solution Statement

As was described at the beginning of this document, obesity is considered a health factor risk in some cases and a protection feature for some others. Previous medical studies are proof one versus the other with no clear conclusion. My study proposes not agreed with any of the two positions but to use the existing information to measure how resilient the body is to surgery using the features of those previous studies. The metric is the time a patient spent hospitalized when it subjected to a medical operation. A potential more complete metric will be the time to recovery, but unfortunately, we do not pose such data set, so we propose to use only the hospitalize time. We are assuming that no doctor will discharge a patient which condition is inadequate for release. After assembling the dataset and scale the features we will identify the unit of the label time (days, hours, minutes).

I will propose the use of a linear regression machine learning algorithm to predict the elapse time. The main reasons for choosing regression are the constraints in time and computing power (less time to train), and second, I expect a linear correlation (positive or negative) based on the previous studies about the BMI and its implications. I will verify the two previous assumptions during the data exploration phase. A proposed benchmark model is the nearest neighbors model, this type of models also requires less time to train. I will consider a regression decision tree model but to find the optimal size of our training data. The MIMICIII data set contains 40,000 patientes records which I filtered based on the answer to the below questions:

- Did the patient survive, yes or no?

- Does the patient has pre-existing conditions, other than the one selected in the obesity paradox studies?
- Does the patient has both metrics recorded, height and weight.
- The MIMICIII database contains information on each visit of the patient to the hospital, for the purpose of this project we only selected the first visit information.

After applying the below filter the 40,000 records, I ended up with a sample of 6862 patients. I will use to an unsupervised clustering algorithm (k-mean) to classify it. From the resulting classes, I will select 1000 records randomly, representing each class. The resulting dataset will be used to plot the training accuracy versus the test accuracy and determine the optimal size of our training data set.

Evaluation Metrics

To measure our model's accuracy, we will use the R2 or the coefficient of determination. This metric tells us how much our prediction is explained by the vector features. The highest good value for R2 is 1.0, and the worst can be negative. A model that predicts correctly not matter what features are input will always give an R2 of 0; this is a clear indication of an overfitted model. The Mathematically operation to calculate R2 is as follow: the square difference between the real label value and the predicted value,' divided by the square variance of the original label's distribution minus 1. I am dividing the actual error of my prediction between the spread of the values of the fitted data distribution, in and an ideal scenario where the prediction is close to the real value the different between one and the division is close to 1. R2 metric help to tune the model parameters so this error gets lower, and the R2 gets higher. The mathematical representation of the R2 is:

$$R^2(y, \hat{y}) = 1 - \Sigma(yi, \hat{yi})^2 / \Sigma(yi, \bar{y})^2 \tag{2}$$

II. Analysis

Data Exploration

I mentioned during the definition section of the project that I would have used the MIMIC-III public data set to train and test our predictor. I selected some of the features of the MIMIC-III (MIT Lab for Computational Physiology) data set, those mentioned in the "obese paradox" studies. The MIMIC-III dataset contains more than 40,000 critical patients records. Information and details related to it can be found on the website https://mimic.physionet.org; this is not an open source data set so I can't share the data set I used to train my model, although any person in the academic area can get access to it by requesting it.

The MIMICIII database contains information on each visit of the patient to the hospital, for the purpose of this project we only selected the first visit information.

Because the selected sample of the data set contained categorical information, I converted such data into dummy variables using the one hot encoder method. The converted features were the patient gender and the medical condition. I switched the medical condition from ICD9 code to its text description.

After applying the above sub-sample methodology, I ended up with data set similar to the one showed in "Figure 1."

Figure 1: Data set sample.

	Patient_Age	BIM	F	M	cardiovascular_disorders	diabetes	hypertension	lipid_disorders	neoplasia
0	71	29.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
1	71	29.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
2	68	30.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
3	68	30.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
4	68	30.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0

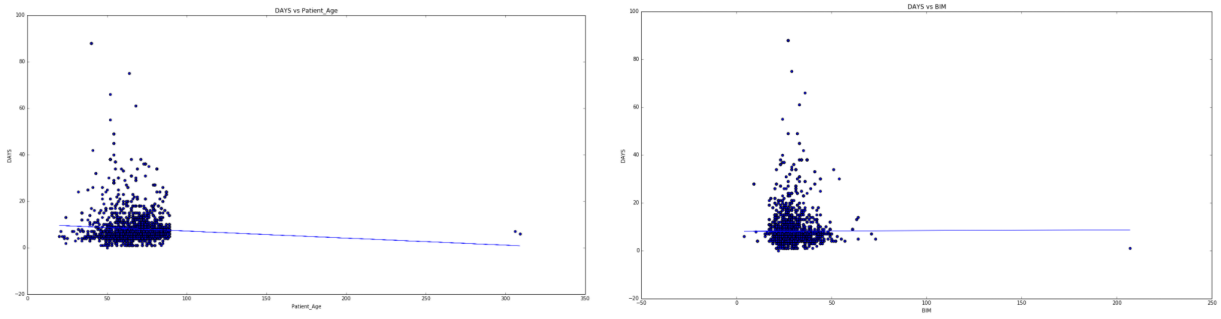
I executed several other pre-processing activities like scaling and outliers removal. Further details of such activities as the methods and code I used are detailed in the companion HTML file of this project.
The statistical summary of the data set is presented below; I took these metrics before getting rid of the outliers, so the min and max values changed.

One of the first observations I noted, was the fact that the data variance wasn't that big. I will show this graphical in the next section.

- Minimum time in days: 0.00
- Maximum time in days: 88.00
- Mean time: 8.16
- Median time 6.00
- Standard deviation of time: 6.91

Exploratory Visualization

The solution to the problem presented at the beginning of this project consisted in the creation of a model capable of predicting a continues target value. Because of the type of label that I needed to predict the logical choice of model was a model based on regression. A regression model works better when the correlation between the features and the target label is strong enough and linear enough to fit a simple linear function. It was necessary for me to visualize and understand the relations of the features (those that where numerical) and the target label, that is the reason why I plotted such relation. The result of the plot can be seen in "Figure 2 (a) and (b)" below.



(a) Figure 2 (b) Figure 2

Figure 2: Visualization of data correlation

I extracted the following conclusions from the above graphic.

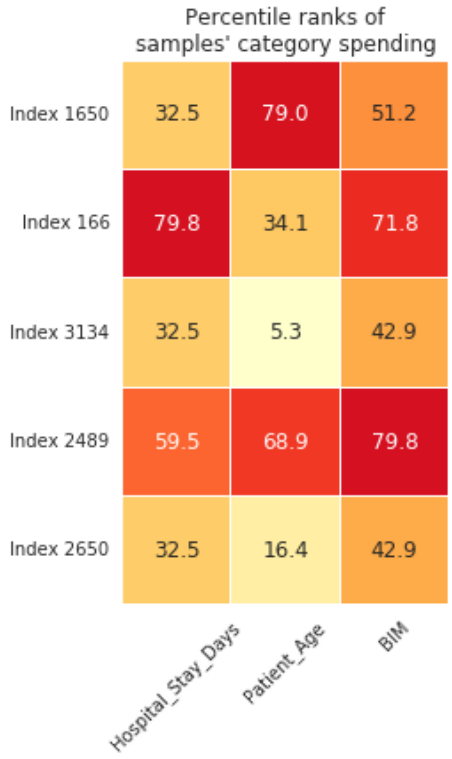
- The variance of the data was small.
- The correlation between the age and the days was slightly negative, which it makes sense because older patients tend to recover slower than young patients.
- Most of the variance is accounting by the hospitalization days and not by the age or BMI.

- Correlation between the features and the target can't be described by a linear function.
- There are outliers that need to be removed from the data set prior the training phase.

I took a sample of five patients and ranking them with the rest of the data set. The resulting raking visualization is shown below.

	Hospital_Stay_Days	Patient_Age	BIM
1650	32.5	79.0	51.2
166	79.8	34.1	71.8
3134	32.5	5.3	42.9
2489	59.5	68.9	79.8
2650	32.5	16.4	42.9

(a) Figure 3



(b) Figure 3

Figure 3: Visualization of patients sample

I used the above visualization to describe groups of patients to help me balance the training data set. Further details on the balance exercise can be obtained by reading the companion HTML file.

Algorithmsl and Techniques

The type of algorithm I selected for this project are of three type, (1) instance-based learning algorithms, (2) regression algorithms, decision tree algorithms and (3) clustering algorithms. To predict a continue's label, it is necessary to select a continues prediction

algorithm. What follow is an explanation of each one of the above selected algorithms. The explanation does not include the clustering algorithm because this was used only with the purposes of give insight to balance the data set.

- Linear Regression (chosen for its simplicity)
- Decision Tree Regression (chosen after explorer the data and understand its correlation)
- Nearest Neighborly (chosen to benchmark the LR and DT)

A Linear Regressor based model is the simplest regression algorithm available; its sole parameter is the Normalization parameter. The normalization parameter is not useful as we will proof in the learning curve because we normalized the data way before passing it to the regressor.

Decision Tree Regressor, we noticed early in the project that the features will be a mix of continue and categorical data, linear regressors performance when some of the features are discrete is not acceptable, a perfect replacement is the DTR. DTR's models are capable of managing numerical and categorical data at the same time. The sklearn Decision Tree Regressor uses the CART algorithm, which is based on the C4.5 algorithm that in turns is an improvement of the ID3 algorithm. The differences between the ID3 and the C4.5 algorithm is that the C4.5 is capable of use numerical features, it does that by creating a discrete range of numerical features, example patients between 30 and 50 is a range, and a value between them will test true for that range. Finally, the differences between the C4.5 and the CART is that CART does the ranges also for the target, which in our case is the days in the hospital. Is because of the previous explanation that this algorithm was selected. For the DT model, the parameter selected for tuning was the tree "maximum depth". Another possible setting could be minimum samples per leaf, but because of the numerical nature of the target and the fact that it will become a discrete value this parameter when set beyond to 1 the model performs poorly.

Nearest Neighborly, this is an instance base learner because it works by using the existing data a not by calculating its relation like the previous two algorithms. It is also suitable for a mix of numerical and categorical features and target. It works by querying the existing training data using the features of the target that need to predict; then it measure the distance between the closest features, and finally, it averages the target of those features to calculate the prediction. We selected this algorithm to use it as a benchmark of the DTR and LN. The metric selected was the number of neighbors. Because of the small variance of the data we expect the algorithm to perform better than the other selected algorithms.

Benchmark Model

We will use a supervise learning method to predict the time a patient spend in the hospital. Because our vector label continues we will use a regression predictor, and we will compare its prediction with an information based learner (K-nearest neighbor), this will allows to benchmark one model versus the other. We will use cross-validation to make sure with dont overfit the model and the R2 Coefficient of Determination to measure the accuracy of our model or what is the same how well it explains the label based on the features. Our learner contains some bias introduced during the features selection; this is because we are selecting the medical conditions used in previous obesity paradox studies, selecting only patients that survive a surgery and finally only taking in consideration the first visit to a hospital. The added bias is based on the observable features related to the paradox, and the sampling process. Introducing new features will be only a guess that can create model complexity due to the additional dimensions.

III. Methodology

Data Processing

As stated in the previous section of this report, the dataset had outliers; these were removed in such cases where a single vector contained multiple outliers. The method I used, was the elimination of data points with multiple values in the first and last quartile.

Another pre-processing method I used was scaling. Using an algorithmic transformation, I scaled all continues features, so all of them were on the same scale.

Finally, I balanced the data set by selecting an equal number of a medical condition, gender, and the combination of age, BMI and hospitalization days. To achieve the later, I used a clustering algorithm that helps me to group the numeric data in two separate clusters.

Full details of the above methods are explained in the companion HTML.

Implementation

During this project, I implemented three supervised learning algorithms and one unsupervised learning algorithm. The supervised learning algorithm used were of different classes, Regression, Instance-Based and Information Gain. I used the python "Science Kit Learning Framework" to implement the three supervised algorithms. I also used an unsupervised learning algorithm to identify groups in the sample data to balance the dataset. Other data pre-processing methodologies were used, like cross-validation, OneHot encoder; the reasons for the utilization of such methods is described along this document and the project's companion HTML file. To improve the accuracy of the built models, I used an ensemble model based on Boosting. I implemented the ensemble using the supervised model with the best R2 score.

During the exploration of the data, I noted that the features variance and its correlation with the label weren't suitable for a linear regression. Also, the strategy used to select the sample from the population introduced a high level of BIAS that caused the model to perform less optimally that how was expected.

Refinement

The initially proposed model was of type "linear regression." After graphical fitting, a linear function over imposed the correlation scatters plot of the features and the data I noted that this kind of model wouldn't perform well; this realization made me put more

emphasis in tuning the benchmark models instead of keeping working with the original proposed design.

I used the python sk-learning framework "Grid Search" function to tune the "Decision Tree" and the "Nearest Neighbor" models. The parameters selected for each model respectively were the "Max Depth" and "Number of Neighbors". The fact I was using a dataset with a mix of numerical and categorical data types limited the number of possible parameters I could select for tuning. For example, I used the "Decision Tree" CART model; it transforms the target labels into ranges which reduce the number of leaves to a small number because of lack of variance in the data; so manipulating the CART number of leaves did not produce any improvement.

Another improvement technic I used, was the ensemble of multiple "Nearest Neighbor" models. Each model added some weights to the labels it got it wrong, so the next model put more emphasis on this for training purposes, this strategy improved the model score at least ten decimal points.

One conclusion I got is that data with a high degree of variance plus additional features to reduce the project BIAS are required to improve even further the model score.

IV.Results

Model Evaluation and Validation

Even after using a boosted ensemble which helps with the BIAS still the best model "Nearest Neighbor" produced only a 0.57 score of accuracy. I believe the model is robust enough, but still can benefit from adding some extract features to help boost the precision. In a later exercise at the end of the project, I tested the model with three fake patients and plotted the result over imposing to the Patients Day standard distribution curve. The predicted values were within the norm deviations what proof the model is predicting a value that belongs to the data distribution. The final model is reasonable but not aligned with what I was expecting. When I started to balance the data, I used a clustering algorithm to determine if BMI variance explained the hospitalization days, but the result was negative. A expected result of solving the proposed problem of my project was to proof the hypothesis of the BMI be responsible for the patient recovery speed, and the conclusions is that is not, or at least not on its own.

Justification

A test was conducted using the three fake patients, shown in Figure 4 below. The results where within one standard deviation and where within the original data distribution so with can at least guarantee that the model learned to a some degree from the original dataset. The Figure 5 shows the statement of the previous paragraphs represented by plotting the dataset normal distribution and over imposing the results predicted by the model.

Figure 4: Test Patients.

Features	Patient 1	Patient 2	Patiente 3
Patient_Age	25	50	80
BIM	20	45	35
Gender	M	M	F
Medical Condition	diabities	cardiovascular disorders	high collesterol

Follow the results predicted by the model.

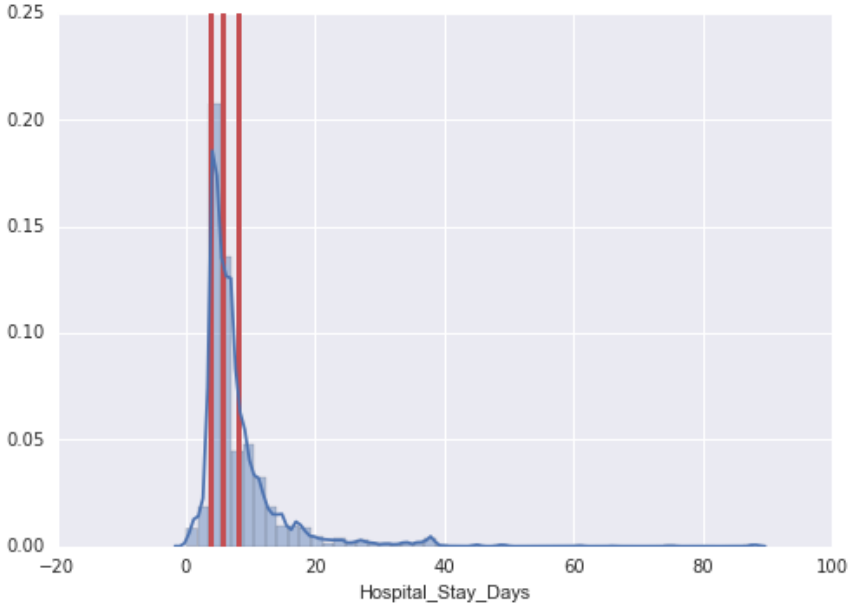
- Predicted days in Hospital patient 1: 5.75 days
- Predicted days in Hospital patient 2: 3.75 days

- Predicted days in Hospital patient 3: 8.0 days

What follows is the graphical representation of the predicted values within the normal distribution of the dataset's labels.

Note that the Y axis represents the days scaled values. I had to scale the days' values because otherwise the normal distribution representation will be skew to the left.

Figure 5: Predicted Values Plotting.



IV.Conclusion

Free-Form Visualization

During the data balancing activity of the project I utilized a cluster algorithm to identify groups within the dataset I could use to balance the training and test data points. As a result of that we were able to conclude that the data point were better explained by the "hospitalization days" variance than the actual BMI variance. The Figure 6 below shows a plot with the resulting cluster and on it we can see the location of the sample patients shown in the Figure 3 (a) of this report.

Figure 6: Clustering Result.

```
Sample point 0 predicted to be in Cluster 0
Sample point 1 predicted to be in Cluster 0
Sample point 2 predicted to be in Cluster 0
Sample point 3 predicted to be in Cluster 0
Sample point 4 predicted to be in Cluster 0
Number of element cluster 0: 4007 Number of Elements Cluster 1: 1318
```

The data point further to the left closer to the cluster one is the sampled patient with 10 days of hospitalization (see Figure 3 a). Another important information is the fact that 4007 data point are in one cluster and 1318 are in the other, this is a clear indication of an unbalanced dataset.

Reflection & Improvements

The model score was a 0.57 precision. It is evident to me that the dataset requires another feature that helps to explain better the hospitalization time. One thing I would do different is worked in tandem with a physician to help me understand base on his/her experience what that missing feature could have been. Another thing I will be different is to select the days in "intensive care" instead of the hospitalization days of the patient first visits. It is strange that when I plotted the variance of the features versus the hospitalization time the patients within 0 and 15 days were all grouped by age or BMI. After 15 days in the hospital, it seems like the age and BMI start to spread as more time the patient spent in the hospital; this could be because I selected the specific medical conditions instead of picking all type of medical conditions. Another hypothesis is that because the records came from the same hospital, such hospital has standard time for the first visit based on the patient condition. A way to overcome these problems is to collect data from different hospitals and get all medical conditions. Also, most of the patients in my sample were overweight or obese; very few were within the normal weight range, which means that more variance in the samples is also required.

One of the biggest challenges I faced was to balance the dataset; this was because the features were a mix of categorical data and numerical data. This type of values requires separate strategies for balancing so I had to balance them separately which requires a good deal of manipulation (split and join).

References

[1] Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA 2013;309:7182.

[2] Obesity paradox. https://en.wikipedia.org/wiki/Obesity_paradox

[3] Braun, N., Hoess, C., Kutz, A., Christ-Crain, M., Thomann, R.,Henzen, C., ... & Schuetz, P. (2017). Obesity paradox in patients with community-acquired pneumonia: Is inflammation the missing link?. Nutrition, 33, 304-310

[4] Carbone, S., Lavie, C. J., & Arena, R. (2017, January). Obesity and Heart Failure: Focus on the Obesity Paradox. In Mayo Clinic Proceedings. Elsevier.).

[5] BMI and all cause mortality: systematic review and non-linear dose-response meta-analysis of 230 cohort studies with 3.74 million deaths among 30.3 million participants

[6] Dagfinn Aune, PhD student and research associate1 2, Abhijit Sen, postdoctoral fellow1, Manya Prasad, resident3, Teresa Norat, principal research fellow2, Imre Janszky, professor1, Serena Tonstad, head physician3, Pl Romundstad, professor1, Lars J Vatten, professor1

[7] Preoperative Evaluation MITCHELL S. KING, M.D, Northwestern University Medical School, Chicago, Illinois Am Fam Physician. 2000 Jul 15;62(2):387-396. http://www.aafp.org/afp/2000/0715/p387.html

[8] TIENE IMPACTO LA HIPERGLUCEMIA Y LA OBESIDAD EN LA PROLONGACION DE LOS DIAS DE HOSPITALIZACION EN LOS PACIENTES EN EL CENTRO MEDICO DOCENTE LA TRINIDAD Author: Madeleyn Santaella, Tutor: Gestne Aure Centro Medico Docente La Trinidad Direccion De Educacion e Investigaciones