Adherence to Medication to Cardiovascular Diseases

Karan Vishwakarma
Chicago, USA
kvishw5@uic.edu

Md Anisur Rahman
Chicago, USA
mrahma46@uic.edu

INTRODUCTION

Cardiovascular diseases (CVDs) is one of the leading causes of death worldwide[12][11] with an estimated 17.9 million people in 2019 suffering from it, according to the World Health Organization (WHO) data. With the increasing influx of data on the causes of these diseases, it has become one of the possibility for providing an early prediction for these diseases and taking the required preventive necessary measures at early stages. Some major aspects contributing to these diseases have become prevalent in the recent years in various reports. One of these factors is Adherence.

Adherence is defined as the ability of any individual to follow a recommended treatment by any doctor or healthcare provider within a prescribed period. It is one of the most imperative factors which alter the levels of medication given to the patient completely. For producing a machine-processable network for all these factors, ontology provide simpler and efficient way for mapping different entities and objects into one semantic network for better scalability. These entities cover the different factors which contribute to these effects.

This report helps in initiating a preliminary analysis with the ultimate goal of providing a semantic framework for encompassing all the factors which contribute to adherence to medication[8][9][1] in cardiovascular diseases. Additionally, the paper sheds some light on the supervised learning algorithms utilizing healthcare data set to predict[10] the vulnerability of patients to heart diseases. These data set provides a numerical evaluation of the different factors included in the ontology for training a model for optimal approximation.

The initial sections discuss the common factors contributing to the adherence for cardiovascular diseases and defines the different sub-entities. The discussion extends to the ontology which provides the required relations between these entities and the different individuals with data properties. Finally, to predict the vulnerability of these diseases based on the data gathered, the various steps for training the supervised learning algorithms are evaluated and an enhanced visualization for the accuracy of the model is provided.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04. . . $15.00
DOI: https://doi.org/10.1145/3313831.XXXXXXX

ONTOLOGICAL DEVELOPMENT

Adherence to medication influences the medication process to greater level. The semantic ontology for this effect is distributed around factors which are highlighted below. Note that the ontology covers factors that have been reported to have the most significant effect on medication adherence. This section provides a theoretical overview of the entities involved in adherence in medication for CVDs and how these entities are related through object properties.

- **Cognitive**: The cognitive factor covers the urge and motivation to recover from the diseases. This is one of the initial stages which provokes the patient to think in a positive direction. Cognitive factors can be subdivided into areas like Belief factors, Overcoming age barriers, exercise motivation, etc.
- **Demographic**: An individual's geographical location and cultural connections are included in the demographic factors. These factors constitute Ethnicity, gender, race, socio-economic, etc, which are inherited by an individual. Some diseases are hereditary which become a contributing factor for cardiovascular diseases in various individuals. These become predominant and need to be addressed during evaluations.
- **Mental Health Factor**: Adherence to medication is affected by the mental health issues which the patient is suffering from additional to the diseases to which the medication is pertaining to. Several stages like avoidance and negligence initiate these factors which might become major issues.
- **Environmental**: Ambiance influence on medication adherence has been one of the most contributing aspects. The ontology includes sub-entities like resident location, transport factors, etc, to address some issues about it. A proper binary representation is utilized to address these factors for the supervised learning model.
- **Healthcare Service**: Accessibility and availability of proper medical providers in the area for patients affect their follow-up treatment. These might delay the dosage quantity and timestamp for every medication event. The knowledge graph addresses entities like doctor visit factor, location accessibility, etc, to cover the health care services domain.
- **Healthcare Literacy**: Awareness level influences the decision-making process and changes our domain of consideration of options for medication. Individuals lacking proper knowledge of diseases and proper treatments might deteriorate their health.

- **Physical Health**: Human body vitals like glucose levels, cholesterol, current blood pressure, are a part of the medical reports and final output of the current condition of the patient. The levels of adherence are analyzed based on these metrics which include dosage quantity, frequency of visits, medication window, etc, for a patient.

- **Psychological**: Mental health awareness has been rising with the increased concern for its influence on daily life activities. The ontology provides entities like depression, anxiety, lack of interest, etc, to cover a small block of these aspects contributing as a hindrance to medication adherence. Several other disorders and PTSD contribute to these factors which need to be addressed.

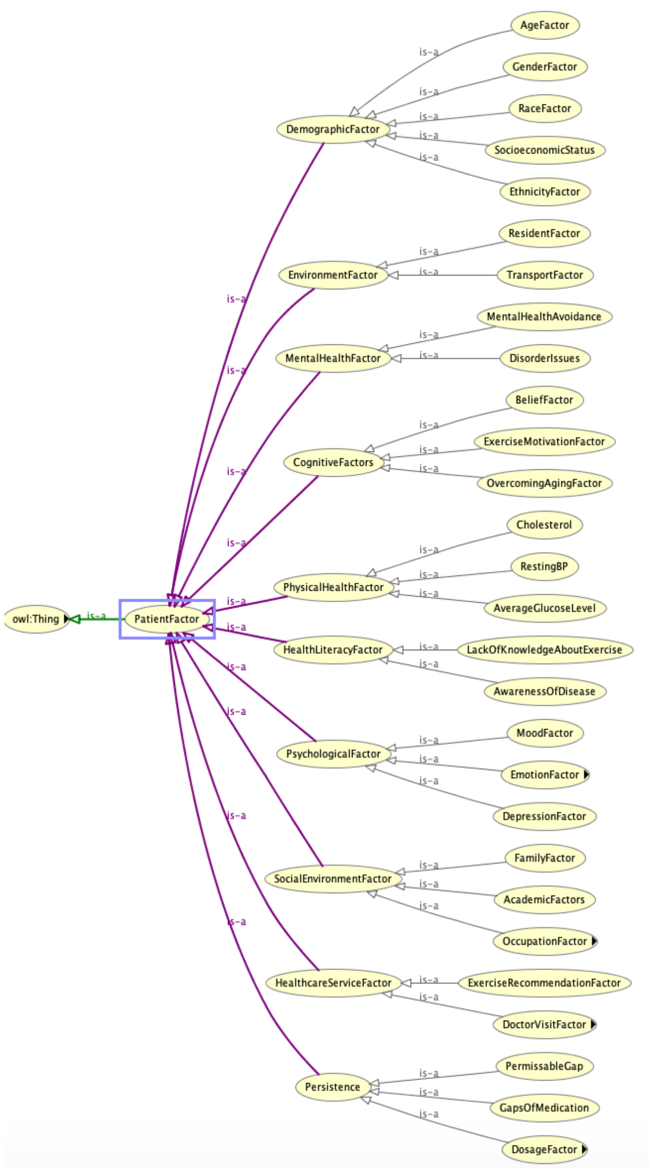

Figure 1. Entities covering the adherence factors in the semantic graph

- **Social Environment**: Social factors include the engagement of people from work, academics, social gatherings, etc. Change of conditions in these places greatly influences an individuals well-being at a mental and physical level. The ontology below shows the major entities covering the above factors discussed.

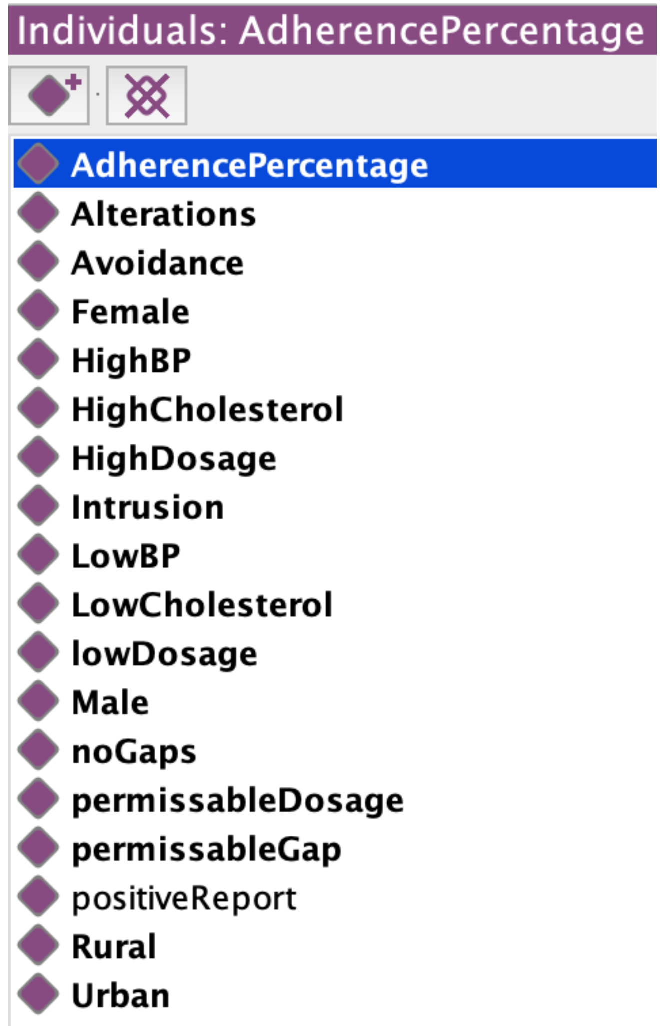

Figure 2. Individuals for mapping important edge vitals to the ontology

- **Persistence**: Persistence is defined as the total medication taken by an individual within a period of time. This includes the dosage and duration of the medication taken by an individual during persistence period. Persistence is influenced by the window for the medication period duration, minimum gap between two medication periods, the dosage amount, dosage quality etc.

  These factors affect the persistence level of the patient which in turn affects the overall adherence to the medication for the prescribed period. Figure 3 shows the treatment record of patients with the time gaps between different medications. The red and blue lines show the two medicines which were consumed by the patients and the yellow window is the period for evaluation chosen to analyze the persistence.

  For having a comprehensive overview of how persistence is affected by patients factors, few factors which greatly influence the persistence level have been narrowed down. These factors include

1. Patient Identifier: denotes the patient being considered for the medication period.
2. Medication Type: type of medicines described.
3. Event data: The period in which the medication is prescribed. It includes the start date of medication and the end date for the medication.
4. Daily quantity of dosage: The quantity of dosage taken by the patient.
5. Duration: The duration for which the medication has been prescribed for the patient.

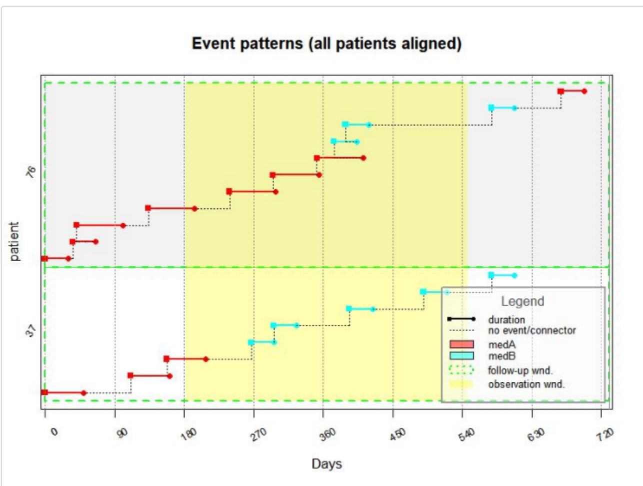

Figure 3. Patterns of dosage for patient and gaps between those medication

Persistence includes entities like permissible Gap, Dosage Factor, Dosage Quantity, etc. Permissible Gap specifies the time duration which is permissible to take during a medication. Dosage Factor consists of the duration of the dosage and the quantity of the dosage required to be taken by the patient during the medication event. Dosage Quantity might be fluctuating based on the availability of the medication available at that time.

In figure 3 the medication events of two patients are compared by plotting the data from Electronic Healthcare data (EHD) and processed through the AdhereR [2] library. AdhereR is defined to handle data from EHD and provide a systematic visualization of the data using the CMA (Continuous Multiple interval Measures of Medication Availability) objects. The data plotted is compared by changing the parameters like event date, amount of dosages, the permissible gap for medication, etc. To have a better visual representation, CMA objects are utilized with the required parameters set for the particular scenario.

Individuals and Data properties

The ontology provides a variety of object properties for addressing the relationship between the entities for the ontology. At a foundational level, the patient Factor entity influences the Adherence Level which as a result adds to the vulnerability of the patient to a disease.

Figure 2 shows some of the individuals for the entities added in the ontology which highlights the maximum and minimum metric for a factor for some of the attributes included in the data set. Note: Some of the individuals included in the ontology have a non-numeric format and can only be specified using binary format.

DATA MODELING

For providing an accurate prediction of the vulnerability of an individual towards heart diseases, various supervised learning algorithms have been trained for different research studies. In this report, for mapping the semantic network entities of the developed ontology, data for some features have been extracted to match the attributes. Furthermore, this data is processed in different stages for training an appropriate algorithm.

For the simplicity of this use case, we are considering few of these features which contribute to adherence for cardiovascular diseases. These features include age, hypertension, effects of employment, smoking status, residence type, glucose level, BMI, blood pressure, cholesterol, and marital status. The predicted value is labeled in the range of 0 and 1 for better classification boundaries between the classes. Various steps involved in training the model have been exemplified below for the cardiovascular disease data set. Note that for processing data with additional attributes, the required parameters need to be included to better fit the classification algorithm. For the current report, we are considering binary Logistic regression classification for simplifying the factors denoted by non-numerical data. The steps required for modeling the data are as follows

- **Cleaning the data**: The data collected from the sources is converted into a form compatible for training the supervised learning algorithms. For example factors like hypertension and effects of employment are converted into binary indexes for evaluation and training of the algorithms.

  For calculating the best and worst-performing features in the algorithm, we performed Recursive Feature Elimination on the data and tried to repeatedly construct the model. This process helped in highlighting the best features which were contributing as an influencing factor with the ultimate goal of using fewer features as possible.

- **Supervised learning algorithm**: For the data set, logistic regression[6][7] is utilized as the classification algorithm for predicting the stroke rate, from the skit-learn library. The classifier helps to incorporate some parameters for efficient classification of the model. The regression model for this report utilizes two attributes for training.

  1. Solver: The labeled attribute or the target attribute in the data set is the vulnerability of an individual to cardiovascular diseases which can be represented by binary values. Because of this, the Binary Logistic Regression algorithm is integrated for the data. Additionally, for optimizing the regression algorithm, one

of the solver algorithms is used named lbfgs (Limited Memory Broyden - Fletcher GoldFarb - Shanno). It saves a few computed updates and hence enhances the performance. For preventing the over-fitting or under-fitting of data, the original data is split into a 1:4 ratio for the testing sample.

2. max_iter: the max_iter attribute for the logistic regression classifier is taken 1000. This value should be optimally chosen for the data set to prevent under-fitting the over-fitting of the training model. For minimizing the cost function value using gradient descent, this value helps in getting the minima for the given data set. The sign for the derivative of the cost function helps in descending the value and making it as optimal as possible.

- **Visualization curves**: For a clear presentation of the performance of the model, a ROC (Receiver Operating Characteristic[4] Curve) is plotted with the sensitivity and specificity aligned at the x and y-axis respectively. It is another way for visualizing binary classifiers. The area under the ROC curve is calculated by integrating the curve over the range, which provides the accuracy with which the model is predicting the trained data set. More the area under the curve, better the accuracy of the algorithm. Accuracy of 0.8333 in the figure 4 denotes the number of the true positives and true negatives predicted by the algorithm for the given training data set.

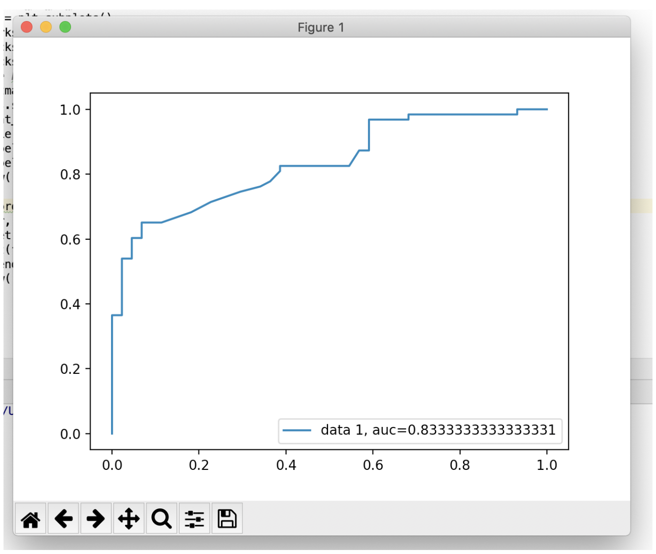

Figure 4. Receiver Operating Characteristic Curve

- **Accuracy of the algorithm**: Table 1 shows the actual reading of the Logistic regression algorithm on the data set. Precision denotes the performance of the classifier of not being able to predict the false positives on the training data set. On the other hand, recall specifies the overall positives found by the model. The f1-score (calculated from the precision and recall[3][5] of the predicted sample) is around 0.77. After multiple testing, the f1-score fluctuates around the range 0.75 - 0.85. It reaches best at 1 and worse at 0.

	Precision	Recall	f1-score
Not Affected	0.77	0.91	0.84
Affected	0.83	0.62	0.71
Accuracy			0.79
Macro Average	0.80	0.77	0.78
Weighted Average	0.80	0.79	0.79

Table 1. Classification report for logistic regression for Cardiovascular disease data.

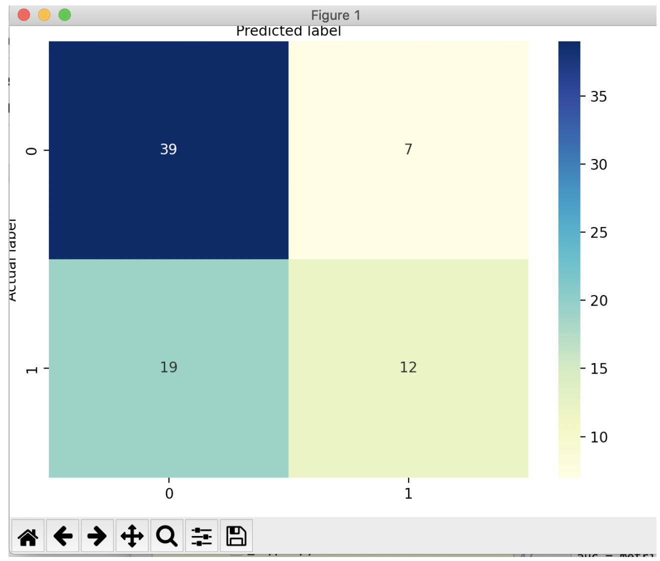

Figure 5. Heat map for true positives and false positives

Additionally, figure 5 shows a pictorial representation of the accuracy of the model. A heatmap is generated for the confusion matrix. This matrix includes the true positive, false positive, true negatives, and false negative generated along with a 2-D matrix. The matrix can be extended to other dimensions depending on the attributes considered. In Figure 5, the top left square and bottom-right square show the true positive and true negatives which were predicted. Similarly, the other two squares show the false negatives and false positives.

FUTURE WORKS

The semantic frameworks cover a section of the entities which contribute to the adherence to medication at a foundational level. The ontology provides a section of the complex network that contributes to the adherence to medication factors. For gaining more accuracy for medical procedures, a complex machine processible network of factors contributing to heart diseases is required which can be utilized for training data sets having more variable parameters. Due to variability with the inclusion of human-induced factors, a optimal approximation to the vulnerability of diseases is still a challenging task from numerical computation.

CONCLUSION

The report helps to provide a preliminary analysis with the ultimate goal of providing a comprehensive semantic network for adherence to medication for cardiovascular diseases. The semantic framework highlights some of the data set utilized for training the supervised learning algorithms for predicting the vulnerability of diseases.

For the current data set, the logistic regression model achieves an optimal accuracy limit over the data set highlighting the different factors interfering with adherence to medication for cardiovascular diseases. Additionally, a pictorial representation from AdhereR helps to account of factors for dosage and duration through CMA objects and get a better overview of how these factors can influence the medication treatment progress and hence influencing the adherence.

REFERENCES

[1] https://blog.cureatr.com/what-is-medication-adherence.

[2] https://cran.r-project.org/web/packages/AdhereR/vignettes/AdhereR-overview.html#adherence-continuous-multiple-interval-measures-of-medication-availability gaps-cma.

[3] https://en.wikipedia.org/wiki/Precision_and_recall.

[4] https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

[5] https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[7] https://towardsdatascience.com/dont-sweat-the-solver-stuff-aea7cddc3451.

[8] https://www.bmj.com/content/374/bmj.n1493.

[9] https://www.frontiersin.org/articles/10.3389/fphar.2020.510339/full.

[10] https://www.kaggle.com/emrearslan123/machine-learning-on-stroke-prediction-dataset.

[11] https://www.who.int/health-topics/cardiovascular-diseases.

[12] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).