2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Preliminary Analysis of the Risk Factor Identification Embedding Model for Cardiovascular Disease

Jihye Moon, *Student Member, IEEE*, Hugo F. Posada-Quintero, *Member, IEEE,* Insoo Kim, *Member, IEEE,* and Ki H. Chon, *Senior Member, IEEE*

Abstract—**Cardiovascular Disease (CVD) is responsible for a large part of healthcare costs every year, but susceptibility to it is affected by complex biological and physiological variables including patients' genetics and lifestyles. There has not been much work to develop a framework that incorporates these important and clinically relevant risk factors into a comprehensive model for CVD research. Moreover, the data labeling required to do so, such as annotating gene functions, is an extremely challenging, tedious, and time-consuming process. In this work, our goal was to develop and validate a risk factor embedding model, which incorporates genotype, phenotype without pre-labeled information to identify various risk factors of CVD. We hypothesize that (1) the knowledge background that does not require data labeling could be gathered from published abstract data, (2) the phenotype, genotype risk factors could be represented in an embedding vector space. We collected 1,363,682 published abstracts from PubMed using the keyword "heart" and 19,264 human gene names, then trained our model using the collected abstracts. We evaluated our CVD risk factor identification model using both intrinsic and extrinsic evaluations: for the intrinsic evaluation, we examined whether or not the captured top-10 words and genes have references related to the input query "myocardial infarction", as one of CVDs, and our model correctly identified them. For the extrinsic evaluation, we used our model to the dimensionality reduction task for classifications, and our method outperformed other popular methods. These results show the feasibility of our approach for disease-associated risk factors of CVD which incorporates genotype, phenotype.**

Clinical Relevance—**Our model provides a comprehensive tool to incorporate various risk factors without any a priori data labeling knowledge for CVD. Our approach shows a potential to provide discovered knowledge that contributes to better understanding and treatment of CVD.**

I. Introduction

Cardiovascular disease (CVD) is leading to 31% of all global deaths in 2016 [1]. CVD results from complex deleterious feedback between multiple organs such as kidneys and lungs, and the disease is also strongly affected by genetic, social, and environmental factors [2]. In the past several years, genome-wide association studies (GWAS) and electronic health record (EHR) [3, 4] based methods have been introduced for risk factor identification in CVD study. Since most CVDs are caused by heritable gene components [5], several approaches based on GWAS to identify genotype-phenotype correlation have been developed. However, due to the complexity of CVDs, for example, a single genotype associated to CVDs may be involved in a wide range of biomarkers, many of which could be found in several diseases, as noted by [5] such that some of the genotype-phenotype approaches are ineffective for most CVDs. Moreover, data labeling tasks such as the functional annotations of genes are still challenging to identify the risk factors associated with diseases in bioinformatics research [6, 7].

EHR data provide real-world information about patients' disease data captured in unstructured clinical narratives. The EHR also contains diverse conditions of multiple co-morbid factors, which are typically not identified in the existing classification and phenotyping methods [3, 4]. However, the EHR based CVD research works have some limitations such as (1) unstructured data management, (2) missing data, (3), selection bias (4) limited source ability. The limited source access especially interrupts the improvement of CVD risk identification research. Recently, Gopalakrishnan et al. [8] reported that text mining of the published biomedical literature has a potential to provide undiscovered knowledge and hidden correlated information for a specific disease. Text mining enables extraction of information from entities such as gene/proteins and diseases which are often readily available in the literature [9]. However, there has not been much work to identify CVD risk factors from the literature databases. In addition, no there are many studies that have incorporated multiple factors (e.g. phenotype, genotype) that are known to affect CVD into a comprehensive model [10].

Embedding model based on a neural network enables integration of multi-modalities such as images and audio by capturing correlations among these data, due to network's ability to map representations of different data types into a numerical vector space [11, 12]. We hypothesize that (1) data labeling tasks (e.g., annotations of gene function/clinical records) for risk factors could be obtained from published abstracts, and (2) the multi-modalities of the phenotype, genotype can be represented in an embedding vector space. To the best of our knowledge, this is the first work on building an embedding model by identifying genotype and phenotype associated risk factors from the published abstracts without any labeling tasks.

In this paper, we propose a framework and validation for our CVD risk factor embedding model. We collected 1,363,682 published abstracts using the word "heart" and

Jihye Moon is with the Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269 USA (corresponding author, e-mail: jihye.moon@uconn.edu).

Hugo. F. Posada is with the Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: hugo.posada-quintero@uconn.edu).

Insoo Kim is with the UConn Health Center, 263 Farmington Ave., Farmington, CT 06030 USA (e-mail: ikim@uchc.edu).

Ki H. Chon is with the Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: ki.chon@uconn.edu).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

19,264 human gene names from PubMed. We validated the embedding model using intrinsic and extrinsic evaluations. For the intrinsic evaluations, we captured CVD-associated words and genes using our model and checked whether or not the captured words and genes have references related to the input queries. The published papers were used as the benchmark of the results. For the extrinsic evaluation, we validated the model by applying the embedding model for the Multi-Ethnic Study of Atherosclerosis (MESA) [13] dataset to discriminate between CVD and healthy subjects.

II. Proposed Method

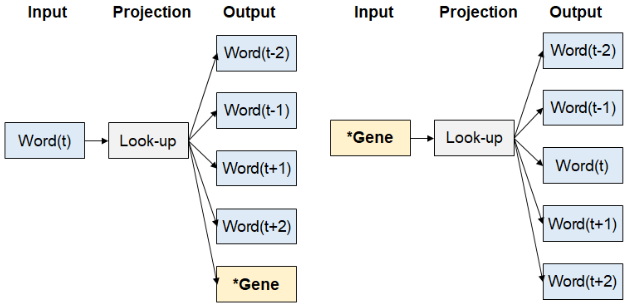
Our method consists of three steps: (a) data collection, (b) proposed CVD risk factor embedding model design, and (c) model validation. Each of these steps is detailed below.

A. *Data collection*

To train the risk factor embedding model, we collected 1,363,682 published abstracts (years published were between 1960s-2020). 1,088,288 of the abstracts were collected using the keyword "heart" and 275,394 abstracts were obtained using human gene names from 19,264 queries of the PubMed database. The gene name lists were obtained using SNP lists provided by the dbGaP accession phs000883.v1.p1. The gene name extraction from SNP lists and abstract collection were implemented using BioPython library with Python.3.7.1. in June 2020.

B. *Proposed model structure*

Our risk factor embedding model is an extension of Word2Vec [14], as shown in Fig. 1, to train genomic information and risk-associated words. Word2vec is one of the well-known models in the Natural Language Processing (NLP) tasks, which consists of a neural network with an input layer, embedding lookup layer, and output layer. The model learns the representation of words in the same sentence or documents by considering co-appearances of the words. The result of the embedding lookup layer is the distributed representation of the input and output words. Thus, the model provides embedding vectors which preserve semantic and syntactic relationships among vocabularies.



(a) The step 1 portion of the proposed model

(b) The step 2 portion of the proposed model

Figure 1. The structure of the proposed embedding model

The CVD risk factor embedding model consists of three steps: the Word2Vec structure, and the models shown in Fig. 1(a), and Fig. 1(b). Word2Vec is designed to train vector sets for each unique word, which captures the correlation between words. Our proposed structure shown in Fig. 1 trains gene vectors by adding gene names into the Word2Vec structure. Note that we collected two types of documents depending on the abstract search query: word-based (e.g. "heart") and the human gene names-based (19,264 gene names). If we use word-based abstracts, our embedding model trains word-only results by using Word2Vec with Skip-gram structure. When published abstract data from a gene-name-based query are provided as training data, the model trains genes and words together from the same documents, as depicted in Figs. 1(a)-(b). The model structures and training process enable vast distributed representations of words and genes using published abstracts. Therefore, the model does not require any knowledge associated with the specific disease. The details of Word2Vec are further described in [14].

C. *Model Validation*

We evaluated the proposed model using both (1) intrinsic and (2) extrinsic evaluations [15]. The intrinsic methods measure semantic relationships between data points represented as embedding vectors to assess the quality of the model. This study used the aggregate scores among the sets of query terms c_i and target words/gene names c_j captured in the embedding space as intrinsic evaluation metrics. The similarity is formulated as follows.

$$Cosine\ Similarity(c_i,\ c_j) = \frac{c_i^T c_j}{\|c_i\|\|c_j\|} \quad (1)$$

The similarity metric ranges from -1 to 1, with 0 as the least similar and 1 as the most similar among data points. We defined words/gene names captured by the proposed model with high values (e.g., close to 1) for the risk factors for each input query (e.g. heart). The published papers were used as the benchmark of the results.

The extrinsic evaluation used vector embedding as input features for supervised learning models. We used the proposed model as a dimensionality reduction approach for classification tasks to illustrate the quality of the data representation. For the classification task, the Multi-Ethnic Study of Atherosclerosis (MESA) [13] from BioLINCC was selected. The MESA is a prospective cohort study which consists of data from 6,814 men and women subjects with ages ranging from 45 to 84 years old, members of four racial/ethnic groups. We selected 775 phenotype features from the baseline examination (2000–2002) and ancillary studies for cardiovascular disease (CVD) patient classifications. The detail of the attributes is described in [13].

The proposed dimensionality reduction is used to validate the quality of the embedding. The inner-product between two independent vectors is a valid measure of similarity [16]. Our embedding model contains the distributed representations for each word/gene name. The MESA datasets provide names of these phenotype features, referred to as variables (e.g. hpylori1: Helicobacter pylori antibodies). By using the embedding vectors from variables' names, the inner-product can reduce the data dimension by the following procedure:

$$Inner\ Product(u,\ v) = |u||\ v|cos\theta = uv^T \quad (2)$$

In Eq. (2), u is a variables matrix (variables for each subject) in the MESA data and v is the variable embedding matrix using the variable's name. The variable embedding vectors are generated by averaging the embedding vectors of the variables' names if the number of the words in the name is two or more. The variable embedding vectors are normalized before the inner-product. In order to compare with the proposed work, we examined Uniform Manifold Approximation and Projection (UMAP) and Principal Component Analysis (PCA). Accuracy, recall, precision, and F1 scores were used as the performance metrics. The performance metrics were calculated as an average of the five results obtained via 5-fold cross-validation.

III. Experiment

This section consists of (a) text data preprocessing, (b) MESA data processing, and (c) model training. The details are described below.

A. Text data preprocessing

Natural language pre-processing techniques were applied for conversion to all lower case (e.g. "Heart" -> "heart"); replacing number next '-' as # (e.g. covid-19 -> covid-#); word tokenization; and removal of unwanted words such as stop-words (e.g. "are," "where"), prepositions /subordinating conjunctions, determiners, personal /possessive pronouns, wh-adverbs (e.g., how, when, where), modals, comparative adverbs, superlative adverbs, coordinating conjunctions, and the existential there (e.g., there is/are). Gene names used as queries for abstract search were labeled with capital letters with the "#" symbol appended. An example of data preprocessing is shown in Table 1.

TABLE I. AN EXAMPLE OF THE PRE-PROCESSING

Type	An example for gene name and sentence processing	
Original	LINC01128	LINC01128 resisted acute myeloid leukemia through regulating miR-4260/NR3C2
Pre-processed	#LINC01128	linc# resisted acute myeloid leukemia regulating mir-# nr#c#

B. MESA data processing

For the binary classification task of CVD vs. non-CVD, we combined 19 CVD-related labels as a CVD label. The 19 events are: Atrial Fibrillation Diagnosis (via ICD9 Code), Atrial Fibrillation (Self-Report), Atrial Fibrillation Diagnosis (ICD-9 or Self-Report), Myocardial Infarction, Resuscitated Cardiac Arrest, Angina Pectoris, Percutaneous Transluminal Coronary Angioplasty, Coronary stent or Coronary atherectomy, Coronary Bypass Graft, Other Revascularization, Congestive Heart Failure, Peripheral Vascular Disease, Stroke, Transient Ischemic Attack, Death, Coronary Heart Disease-Hard, Coronary Heart Disease-All, Cardiovascular Disease-Hard, Cardio-vascular Disease-All, Coronary Revascularization.

We used only 543 attributes among 775 from the MESA data, as some of them had missing values that were <5%. Thus, 543 attributes from 6,814 subjects were used to classify between CVD (2,147 subjects) versus non-CVD (4,667 subjects).

C. Model Training

We trained different models for two tasks: a risk factor identification embedding model using abstract documents, classification models with the proposed embedding model using the MESA dataset for extrinsic evaluation.

The risk factor embedding model was trained with 1,363,682 published abstracts, using the structure shown in Fig. 2. We used negative sampling of 12, a minimum word count of 6, a window size of 2, an epoch of 20, a learning rate of 1.0 with a gradient descent optimizer, and a dimension of 128. The model was trained using Python 3.7 with Tensorflow ver. 1.18.3. Our model trained 280,138 unique words and 19,264 unique gene names.

For the extrinsic evaluation, two popular machine learning algorithms—Random Forest (RF) and Logistic Regression (LR)—and a basic deep neural network model (DNN) were used as classifiers. We trained each method (RF, LR, and DNN) five times with 80% of the data, tested with 10% of the data, and validated the model with the remaining 10% of the data. All features were standardized before feature selection or dimensionality reduction tasks to handle biases towards larger feature values. The classification tasks were evaluated using stratified 5-fold cross validation. The best RF and LR classifiers were selected using a grid search. For the DNN model, the best classifier was chosen based on the highest accuracy in the validation dataset post-training. For determining hyper-parameters with the grid search, an estimator of {256, 512}, a max feature of 128, and a max depth of {3,5} were used for RF. For LR, *C* parameter was varied {0.01, 0.1, and 1.0}. The DNN classifier with 10 hidden layers, a learning rate of 0.001, a gradient descent optimizer, a batch size of 128, and an epoch of 1,000 were used. The machine learning models—RF and LR with a grid search—were trained using scikit-learn v. 0.23.2 and the DNN model was trained using Tensorflow v. 1.18.3 with Python 3.7.

IV. Results and discussion

A. Results for the Intrinsic Evaluation

To assess the risk factor embedding model, we used "myocardial infarction" as the input query to obtain risk factors (associated words and genes). Embedding vectors of two words such as "myocardial + infarction" were averaged into an embedding vector. The captured words and genes were sorted by the highest score based on the cosine similarity for the input queries, as shown in Table II.

TABLE II. THE LIST OF TOP 10 DISEASE ASSOIBED WORDS/GENES IDENTIFIED BY OUR MODEL FOR "MYOCARDIAL INFARCTION"

Word	Gene
mi	**#PPP6R3**[17]
ami	**#GNB2**[18]
infarct	**#HRK**[19]
infarctions	**#NUP43**[20]
st-elevation	**#CAPZA3**[21]
infraction	**#NUFIP1**[22]
non-q-wave	**#PABPC4**[22]
post-myocardial	**#BRD2**[22]
stemi	**#ACADS**[22]
q-wave	**#WFDC9**[23]

To validate the performance of the embedding model for the intrinsic evaluation, we checked whether or not the captured genes have references related to the input queries. We found that all 10 genes for myocardial infarction have associated reference, as shown in Table II. The details of the captured genes are described in their references. Our model using the query of myocardial infarction provided relevant associated terms such as: mi (myocardial infarction), ami (acute myocardial infarction), stemi (ST-Elevation Myocardial Infarction), and mis (myocardial infarctions) as well as ECG-related indicators such as st-elevation, non-q-wave, and q-wave which suggest current or prior myocardial infarction. These results suggest that our embedding model provides a good representation of the phenotype and genotype.

B. *Results for the Extrinsic Evaluation*

To validate the quality of the embedding model, we conducted CVD classifications using the MESA dataset.

TABLE III. THE PERFORMANCE OF DIMENSIONALITY REDUCTION(DR) FOR CVD CLASSIFICATION (MESA)

DR	ML	Acc.	Pre.	Re.	F1	Input size	Ave. time
Non-DR	DNN	0.75	0.70	0.67	0.68	543	N/A
	RF	0.76	0.73	0.68	0.70		
	LR	0.75	0.71	0.67	0.68		
Proposed Work	DNN	0.73	0.69	0.65	0.66	128	**0.00 sec**
	RF	0.73	0.69	0.64	0.65		
	LR	**0.74**	**0.70**	**0.65**	**0.66**		
UMAP	DNN	0.54	0.37	0.50	0.36	128	13.68 sec
	RF	0.32	0.33	0.50	0.25		
	LR	0.46	0.23	0.50	0.30		
PCA	DNN	0.66	0.59	0.57	0.57	128	0.16 sec
	RF	0.60	0.53	0.57	0.50		
	LR	0.55	0.27	0.50	0.34		

As shown in Table III, our approach outperformed two well-known dimensionality reduction methods, UMAP and PCA, with much faster execution time (~1,368 times faster). Note that the dimensionality reduction task is to validate the quality of the embedding models for the phenotype variable representations, which preserve the correlations between data points. The outstanding performance suggests our embedding model provides a superior vector representation for phenotype variables. These results shown in Tables II, III provide good support for our approach to risk factor identifications of CVD using embedding vectors.

V. CONCLUSION

We propose a CVD risk factor identification embedding approach incorporating phenotype, genotype using published abstracts. Our approach was validated using both intrinsic and extrinsic evaluations. The model showed accurate risk factor identification results and outperformed other dimensionality reduction methods for classification tasks. This proposed framework is not limited to specific fields. Our approach could be applied to a variety of disease data sets. In addition, it is easy to build the model in any environment. Many information retrieval tasks require expensive computing power such as multiple GPUs to address the large number of parameters required by other models. However, our model was computed on a CPU with 8 cores (Xeon E5-2690 v3, @2.60 (GHz)) and the training time took only 26 hours. While further improvements are necessary with our approach, the proposed risk factor embedding model has the potential to provide better understanding and treatment of CVD by identifying accurate risk factors of CVD.

REFERENCES

[1] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. J Am Coll Cardiol 2017 Jul 04;70(1):1-25

[2] Bhatnagar, Aruni. (2017). Environmental Determinants of Cardiovascular Disease. Circulation Research. 121. 162-180.

[3] Jin, B., Che, C., Liu, Z, et al. (2018). Predicting the Risk of Heart Failure with EHR Sequential Data Modeling. IEEE Access. PP. 1-1.

[4] Nagamine, T., Gillette, B., Pakhomov, A. et al. (2020). Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci Rep 10, 21340.

[5] Leopold, J. A., Maron, B. A., Loscalzo, J. (2020). The application of big data to cardiovascular disease: paths to precision medicine. The Journal of clinical investigation, 130(1), 29–38.

[6] Du, J., Jia, P., Dai, Y. et al. (2019). Gene2vec: distributed representation of genes based on co-expression. BMC Genomics 20, 82.

[7] Alshahrani M, Hoehndorf R. (2018). Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. Bioinformatics.34(17) 901-i907.

[8] Vishrawas Gopalakrishnan, Kishlay Jha, Wei Jin, Aidong Zhang, A survey on literature based discovery approaches in biomedical domain, Journal of Biomedical Informatics, Volume 93, 2019, 103141, ISSN 1532-0464,

[9] Lavrač, N., Martinc, M., Pollak, S. et al. Bisociative Literature-Based Discovery: Lessons Learned and New Word Embedding Approach. New Gener. Comput. 38, 773–800 (2020).

[10] https://www.nhlbi.nih.gov/sites/default/files/media/docs/NHLBI-Big-Data-Analysis-Challenge-Prize-Announcement-02.27.20.pdf

[11] Baltrusaitis, T., Ahuja, C., Morency, L. (2017). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[12] Hancock, J.T., Khoshgoftaar, T.M. (2020). Survey on categorical data for neural networks. J Big Data 7, 28.

[13] Bild DE., Bluemke DA., Burke GL., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol; 156:871–81.

[14] Mikolov, T., Corrado, G.., Chen, K. et al. (2013). Efficient Estimation of Word Representations in Vector Space. 1-12.

[15] Wang, B., Wang, A., Chen, F., et al. (2019). Evaluating word embedding models: methods and experimental results. APSIPA Transactions on Signal and Information Processing. 8.

[16] Vilnis, L., McCallum, A. (2014). Word Representations via Gaussian Embedding.

[17] Tan, W., Lim, B., Anene-Nzelu, C. (2017). A landscape of circular RNA expression in the human heart, Cardiovascular Research, Volume 113, Issue 3, Pages 298–309.

[18] Stallmeyer, B., Kuß, J., Kotthoff, S., et al. (2017). A Mutation in the G-Protein Gene GNB2 Causes Familial Sinus Node and Atrioventricular Conduction Dysfunction. Circulation Research. 120.

[19] Dlamini, Z., Tshidino, S. C., Hull, R. (2015). Abnormalities in Alternative Splicing of Apoptotic Genes and Cardiovascular Diseases. International journal of molecular sciences, 16(11), 27171–27190.

[20] Kontou, P., Pavlopoulou, A., Braliou, G. et al. (2018). Identification of gene expression profiles in myocardial infarction: a systematic review and meta-analysis. BMC Med Genomics 11, 109.

[21] Urashima, T., Zhao, M., et al. (2008). Molecular and physiological characterization of RV remodeling in a murine model of pulmonary stenosis. American journal of physiology. Heart and circulatory physiology, 295(3), H1351–H1368.

[22] Wang, Y., Huang, Y., et al. (2018). Bioinformatic Analysis of the Possible Regulative Network of miR-30a/e in Cardiomyocytes 2 Days Post Myocardial Infarction. Acta Cardiologica Sinica, 34(2), 175–188.

[23] Mejía O.A.A., Pulido A.J.P. (2014). Bioinformatic Analysis of Two Proteins with Suspected Linkage to Pulmonary Atresia with Intact Ventricular Septum. In: Castillo L., Cristancho M., Isaza G., Pinzón A., Rodríguez J. (eds) Advances in Computational Biology. Advances in Intelligent Systems and Computing, vol 232. Springer, Cham.