ScienceDirect

Computers in Biology and Medicine

Volume 174, May 2024, 108413

Symptom-based drug prediction of lifestyle-related chronic diseases using unsupervised machine learning techniques

Sudipto Bhattacharjee [a], Banani Saha [a], Sudipto Saha [b]

Show more

https://doi.org/10.1016/j.compbiomed.2024.108413

Get rights and content

Highlights

- Computed novel associations between drugs and symptoms of lifestyle-related diseases.
- Developed unsupervised machine learning models to predict drugs from symptoms as features.
- Good chemical and biological similarity was found among clustered drugs.
- SDLDpred, a user-friendly web application was developed for prediction with the optimal model.
- Can aid clinicians in decision-making for early treatment of life-style related diseases.

Abstract

Background and objectives

Lifestyle-related diseases (LSDs) impose a substantial economic burden on patients and health care services. LSDs are chronic in nature and can directly affect the heart and lungs. Therapeutic interventions only based on symptoms can be crucial for prompt treatment initiation in LSDs, as symptoms are the first information available to clinicians. So, this work aims to apply unsupervised machine learning (ML) techniques for developing models to predict drugs from symptoms for LSDs, with a specific focus on pulmonary and heart diseases.

Methods

The drug-disease and disease-symptom associations of 143 LSDs, 1271 drugs, and 305 symptoms were used to compute direct associations between drugs and symptoms. ML models with four different algorithms – K-Means, Bisecting K-Means, Mean Shift, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) – were developed to cluster the drugs using symptoms as features. The optimal model was saved in a server for the development of a web application. A web application was developed to perform the prediction based on the optimal model.

Results

The Bisecting K-means model showed the best performance with a silhouette coefficient of 0.647 and generated 138 drug clusters. The drugs within the optimal clusters showed good similarity based on i) gene ontology annotations of the gene targets, ii) chemical ontology annotations, and iii) maximum common substructure of the drugs. In the web application, the model also provides a confidence score for each predicted drug while predicting from a new set of input symptoms.

Conclusion

In summary, direct associations between drugs and symptoms were computed, and those were used to develop a symptom-based drug prediction tool for LSDs with unsupervised ML models. The ML-based prediction can provide a second opinion to clinicians to aid their decision-making for early treatment of LSD patients. The web application (URL - http://bicresources.jcbose.ac.in/ssaha4/sdldpred↗) can provide a simple interface for all end-users to perform the ML-based prediction.

Introduction

Lifestyle-related diseases (LSD) are defined as the diseases that are associated with a person's lifestyle, occupation and relationship with the environment [1]. These are mainly non-communicable diseases and are chronic in nature [2]. The health condition deteriorates throughout a patient's lifetime and, thus, incurs a huge financial burden in terms of healthcare costs [3,4]. The leading causes of LSDs include smoking, unhealthy diet, alcohol consumption, and lack of physical activity [5,6]. The growth in industrialization and urbanization has led to a considerable increase in pollution, which has also been responsible for the LSDs at the societal level [7]. The burden of LSDs is often measured by quality-adjusted life years (QALY) and disability-adjusted life years (DALY) [8,9]. The LSDs diminish the QALY and DALY for patients in all age groups, especially for older persons [10,11]. Pulmonary diseases constitute a majority of the LSDs, such as chronic obstructive pulmonary disease (COPD), asthma, and interstitial lung disease (ILD) [[12], [13], [14]]. Lifestyle factors are also known to be associated with the severity and mortality of infectious lung diseases such as tuberculosis and COVID-19 and with the management of post-recovery patients [[15], [16], [17]]. Recent studies have concluded that even lung cancer is also a LSD, with smoking and air pollution being its risk factors [18]. Pulmonary diseases are often diagnosed with other comorbid LSDs. These comorbidities include nutritional or metabolic diseases like diabetes and obesity, cardiovascular diseases like hypertension and coronary artery disease, and musculoskeletal diseases like arthritis and osteoporosis [[19], [20], [21]].

The symptoms are the first information the doctors receive during a patient's visit. These are the most direct characteristics observed in the patient. Doctors analyze the symptoms data to diagnose the disease and prescribe drugs or any other therapy. The Human Symptom Disease Network (HSDN) is a publicly available database that provides symptom-disease associations curated using text-mining methods [22]. Furthermore, the therapeutic associations between the drugs and diseases are well explored. The Comparative Toxicogenomics Database (CTD) is a public database containing drug-disease associations and other scientific data [23]. CTD integrates data from several databases, such as DrugBank [24]. The Therapeutic Target Database (TTD) also provides a comprehensive set of drug-disease associations [25]. Although the symptom-disease and drug-disease associations are crucial for clinical diagnosis and therapeutics, the construction of direct associations between symptoms and drugs is a major aspect that remains overlooked.

In recent years, machine learning (ML) models have utilized these association data to perform drug predictions for diseases with the aim of introducing robust automation in healthcare. Kim *et al.* [26] used a combination of graph mining, matrix factorization, and deep learning techniques to predict novel drug-disease associations using heterogeneous association networks (disease-disease, drug-drug, gene-gene, disease-gene, drug-gene, drug-disease) as input [26]. Models were also developed for drug prediction using an

integration of drug-disease, disease-gene, and drug-gene associations as input to train ensembles of generalized tensor decomposition and neural network [27]. Jiang *et al.* [28] combined the associations between proteins, miRNAs, lncRNAs, diseases, and drugs to develop ML models for drug prediction [28]. Liu *et al.* [29] designed a fusion of random-walk-with-restart and ML algorithms for novel drug predictions of diseases [29]. Wang *et al.* [30] developed a novel ensemble strategy to predict drug-disease associations [30]. The ML models were also developed for symptom-based disease prediction. Shah and Dhawan (2023) compared seven different ML algorithms for disease prediction from symptoms [31]. Divya *et al.* [32] developed a ML-based disease diagnosis model using symptom features [32]. Islam *et al.* [33] trained deep learning models for disease prediction from symptoms [33]. Hema *et al.* [34] and Kosarkar *et al.* [35] developed tools with graphic user interface (GUI) for ML-based prediction of diseases from symptoms [34,35]. These recently published models can be viewed as piecemeal models that predict either diseases from symptoms or drugs from diseases. Recently, a ML-based framework for drug recommendation from symptoms, named 4SDrug, was developed that used set-aggregation and set-augmentation methods followed by a multiclass classification approach [36]. Still, ML-based tools for predicting drugs directly from symptom data are not explored extensively, even though such tools can be beneficial for ensuring early and improved treatment to patients.

In this work, to fill the void of drug-symptom association, we leveraged the existing drug-disease and symptom-disease associations of LSDs – with a prime focus on pulmonary and cardiovascular diseases – to compute direct associations between drugs and symptoms. If symptoms are associated with a disease, and if there are specific drugs for that disease, then there can be a hidden relationship between the symptoms and drugs. Next, we used the drug-symptom associations to develop ML models for the novel task of drug prediction with the symptoms as features in an unsupervised manner. The models were trained using four different clustering algorithms. The symptom-based drug clusters generated by the optimal model were validated by evaluating the similarities between the drugs within the clusters. The drug similarities were evaluated in terms of gene ontology (GO) annotations, chemical ontology (CO) annotations, and maximum common substructure (MCS). Finally, we developed a web application that provides a user-interface to perform the drug prediction from symptoms using the optimal ML model.

Section snippets

Datasets

A list of 143 LSDs was obtained by searching the International Classification of Diseases 11th Revision (ICD-11) nomenclature. ICD-11 provides a hierarchical classification of diseases [37]. The obtained LSDs include pulmonary, cardiovascular, diabetes, obesity, and musculoskeletal diseases. The complete ICD-11 hierarchy of these LSDs is given in Table ST1 of Supplementary data 1. A set of 3159 associations between 1271 drugs and the LSDs were obtained from Comparative Toxicogenomics Database…

Clustering performance

The clusterability of the input data was assessed by computing the Hopkins statistic. The datasets, $A_{drug\text{-}symptom<cosine>}$, $A_{drug\text{-}symptom<pearson>}$, and $A_{drug\text{-}symptom<jaccard>}$, achieved Hopkins statistic values of 0.034, 0.035 and 0.041 respectively. Since all the three values tend to zero, the null hypotheses that these three datasets were uniformly distributed can be rejected. It suggests that a clustering analysis on these datasets are feasible. Next, the ML models for symptom-based clustering …

Using SDLDpred

The SDLDpred homepage snapshot is given in Fig. SF19 of Supplementary data 1. A web form is provided on the homepage for users to input a set of symptoms along with their intensities on a scale of 1–10. Users can select the symptom from a drop-down list for each symptom input and insert the intensity value in the corresponding text box. Users can click on the “Add symptom” button to add more symptoms. The “Delete symptom” button can be used to delete one or more symptoms. Users can clear the…

Discussion

The drug-disease and symptom-disease associations are well established and several databases are available where such associations are curated and compiled [[22], [23], [24]]. Recently published ML-based studies predict diseases from symptoms [31,[33], [34], [35]] or drugs from diseases [[26], [27], [28], [29], [30]]. These studies mainly employ supervised techniques by using the known associations as ground truth labels. Also, in studies that work with real patient data, the actual drugs given …

Conclusion

In this work, unsupervised ML models were created for symptom-based drug prediction for lifestyle-related diseases. It can aid clinicians to provide a rapid treatment initiation for patients. Furthermore, direct associations between drugs and symptoms were computed and used as features to train the ML models. A web application was also

developed that enables users to access the ML models from any web browser and perform drug prediction.…

Data availability

SDLDpred, is available at http://bicresources.jcbose.ac.in/ssaha4/sdldpred↗. The source code is available at https://github.com/ttsudipto/sdldpred↗. The datasets are given at http://bicresources.jcbose.ac.in/ssaha4/sdldpred/about.html↗.…

CRediT authorship contribution statement

Sudipto Bhattacharjee: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Banani Saha:** Investigation, Supervision, Writing – review & editing. **Sudipto Saha:** Conceptualization, Investigation, Project administration, Resources, Supervision, Writing – review & editing.…

Declaration of competing interest

The authors declare that they have no competing interests.…

Acknowledgments

The authors thank Bose Institute, Kolkata, India, for the web-hosting facilities. SB and BS acknowledge the Department of Computer Science and Engineering, University of Calcutta. SS acknowledges the BIC COE project funded by the Dept. of Biotechnology, Govt. of India (sanction no. BT/PR40174/BTIS/137/45/2022).…

Recommended articles

References (93)

L. Engelen *et al.*
Who is at risk of chronic disease? Associations between risk profiles of physical activity, sitting and cardio-metabolic disease in Australian adults
Aust. N. Z. J. Publ. Health (2017)

S.M. Nyenhuis *et al.*

Impact of lifestyle interventions targeting healthy diet, physical activity, and weight loss on asthma in adults: what is the evidence?
J. Allergy Clin. Immunol. Pract. (2018)

H. Liang *et al.*
Association of outdoor air pollution, lifestyle, genetic factors with the risk of lung cancer: a prospective cohort study
Environ. Res. (2023)

T.S. Prior *et al.*
Clusters of comorbidities in idiopathic pulmonary fibrosis
Respir. Med. (2021)

T. Wu *et al.*
clusterProfiler 4.0: a universal enrichment tool for interpreting omics data
Innovation (2021)

F.M. Couto *et al.*
Semantic similarity definition

F. Victoria-Muñoz *et al.*
Cheminformatics analysis of molecular datasets of transcription factors associated with quorum sensing in *Pseudomonas aeruginosa*
RSC Adv. (2022)

S.S. Braman
Chronic cough due to acute bronchitis
Chest (2006)

P.T. King *et al.*
Characterisation of the onset and presenting clinical features of adult bronchiectasis
Respir. Med. (2006)

W. Ulmer
Fleroxacin versus amoxicillin in the treatment of acute exacerbation of chronic bronchitis
Am. J. Med. (1993)

View more references

Cited by (0)

View full text