International Journal of Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal.
Vol-14 Issue-01 Mar 2025

Disease Drug Prediction using ML

[1]SHAIK ABDUL QUADEER, [2]B. GNANAPRATAP

[1]PG Scholar, Dept. of CSE, Sri Annamacharya Institute of Technology & Science, Rajampet, AP, India.

[2]Assistant Professor, Dept. of CSE, Sri Annamacharya Institute of Technology & Science, Rajampet, AP, India.

Abstract: The advancement of computing technology has significantly contributed to the medical field, yet accurate disease diagnosis remains a complex challenge due to numerous influencing factors. This paper presents a Medical Decision Support System (MDSS) that leverages machine learning techniques to enhance disease prediction and drug recommendations. The proposed system utilizes the K-Means clustering algorithm to group diseases based on patient symptoms, reducing the complexity of diagnosis by identifying the most probable conditions. Additionally, a Service-Oriented Architecture (SOA) ensures accessibility, while the LAMSTAR network is employed to optimize weight calculations, improving diagnostic accuracy and processing speed. This approach enhances medical decision-making by reducing the dependency on human expertise, minimizing errors, and improving overall healthcare efficiency.

INDEX TERMS: Medical Decision Support System, Disease Prediction, K-Means Clustering, Machine Learning, Service-Oriented Architecture, LAMSTAR Network, Diagnosis Automation, Drug Recommendation..

1. INTRODUCTION

With the advancement of computing technology, the medical field has integrated various technological tools such as surgical representation processes and X-ray photography. However, medical decision-making still heavily relies on doctors' expertise and experience due to the influence of multiple factors, including medical records, environmental conditions, blood pressure, and other physiological variables. The complexity of analyzing these vast numbers of factors makes it difficult to develop an accurate and reliable diagnostic model. To address this challenge, Medical Decision Support Systems (MDSS) are essential, assisting doctors in making informed decisions by leveraging computational models for disease diagnosis.

Medical decision-making involves identifying possible diseases based on symptoms and patient data. The diagnostic process considers both medical and non-medical factors such as ethical concerns, financial incentives, and communication protocols. Automated decision support systems use rule-based approaches to handle repetitive management problems, reducing human errors and improving efficiency. However, medical diagnosis remains challenging, especially in cases involving rare diseases, stress-induced conditions, or incomplete patient data.

A standard diagnostic approach involves differential diagnosis, where doctors systematically analyze symptoms to identify potential causes. This process includes gathering patient information, listing possible causes, prioritizing the most critical conditions, and eliminating unlikely options through tests and observations. While effective, this method requires extensive medical knowledge and experience, making it prone to errors when dealing with complex cases.

To improve disease diagnosis, this study proposes the use of the K-Means clustering algorithm to categorize diseases based on patient symptoms. K-Means is an unsupervised learning technique suitable for grouping large datasets, efficiently identifying disease patterns by determining centroids for different clusters. By incorporating Service-Oriented Architecture (SOA), the system enables online accessibility, allowing users to interact with the model from any location. Additionally, the LAMSTAR network is used to enhance weight calculations, increasing the accuracy and speed of diagnosis. This approach aims to reduce diagnostic errors, streamline the decision-making process, and provide better healthcare outcomes through machine learning-driven disease prediction.

2. LITERATURE SURVEY

2.1 A standard database for drug repositioning:

https://pubmed.ncbi.nlm.nih.gov/28291243/

ABSTRACT: Academics and business are becoming more interested in drug repositioning, which is the process of finding, evaluating, and promoting already authorised medications for new uses. This is because repositioned medications take less time and money to produce. Because they ostensibly identify the most promising candidate medications for a certain indication, computational approaches for repositioning are attractive. However, inconsistent method validation in the field makes it difficult to compare the vast diversity of computational repositioning techniques. Furthermore, it is cognitively unsatisfactory and impairs repeatability to make the frequent simple assumption that all innovative predictions are incorrect. By offering a gold standard database, repoDB, that includes both true positives (authorised medications) and true negatives (failed medications), we address this assumption. In addition to creating a web application that lets people explore sections of the data, we have made the whole database and all of the code needed to create it publicly available (http://apps.chiragjpgroup.org/repoDB/).

2.2 Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data:

https://link.springer.com/chapter/10.1007/978-3-642-38288-8_14

ABSTRACT: The greatest network of linked data for the life sciences is now offered by Bio2RDF. Here, we outline a major improvement to improve the overall quality of RDFized datasets produced from open scripts that are driven by an API to produce downloadable RDF and database files, registry-validated IRIs, dataset provenance and metrics, and SPARQL endpoints. We show how to use the Semanticscience Integrated Ontology (SIO) to integrate semantics in federated SPARQL searches both inside and outside of the Bio2RDF network. A solid basis for expanded coverage and ongoing data

integration in the biological sciences is established by this effort.

2.3 Drug target identification using side-effect similarity

https://pubmed.ncbi.nlm.nih.gov/18621671/

ABSTRACT: Drug targets have so far been identified using molecular or cellular characteristics, such as by taking advantage of similarities in chemical structure or activity between cell lines. To determine if two medications share a target, we looked for commonalities in phenotypic side effects. A network of 1018 drug-drug interactions based on side effects was identified when 746 marketed medications were examined; 261 of them were created by chemically dissimilar medications with distinct therapeutic purposes. Twenty of these surprising drug-drug interactions were explored experimentally, and thirteen suggested drug-target relations were confirmed by in vitro binding experiments; eleven of them showed inhibition constants below 10 micromolar. Nine of these were examined and validated in cell experiments, indicating potential novel applications for commercially available medications and demonstrating the viability of utilising phenotypic data to deduce molecular relationships.

2.4 Systematic evaluation of drug-disease relationships to identify leads for novel drug uses

https://pubmed.ncbi.nlm.nih.gov/19571805/

ABSTRACT: The term "drug repositioning" describes the process of finding new applications for medications that differ from the ones for which they were created. Selecting the indication for which a medicine of interest may be prospectively investigated presents one difficulty in this endeavour. To tackle this difficulty, we conducted a thorough evaluation of a medication treatment-based perspective on disorders. Using a "guilt by association" method, ideas for new medication applications were produced. The proposed novel medication applications produced by this method were substantially enriched in relation to past and ongoing clinical studies when compared to a control group of drug uses.

2.5 PREDICT: a method for inferring novel drug indications with application to personalized medicine

https://pubmed.ncbi.nlm.nih.gov/21654673/

ABSTRACT: A crucial stage in the development of new or authorised medications is the inference of possible pharmacological indications. medication repositioning or matching medication and illness gene expression patterns have been the major topics of previous computational approaches in this field. Here, we provide a brand-new approach for large-scale drug indication prediction (PREDICT) that works with both new and established medications. Our approach, which uses many drug-drug and disease-disease similarity metrics for the prediction job, is founded on the finding that comparable medications are recommended for similar disorders. It outperforms current techniques in cross-validation, achieving excellent specificity and sensitivity (AUC=0.9) in medication indication prediction. Our predictions are confirmed by their consistency with tissue-specific expression data on the therapeutic targets and by their overlap with medication indications that are presently undergoing clinical trials. We further demonstrate that pharmacological

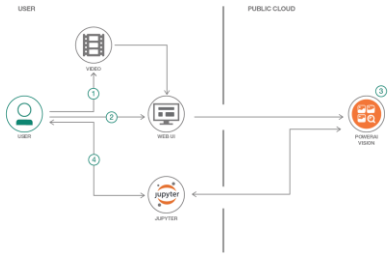
indications for novel disorders may be reliably predicted using disease-specific genetic markers (AUC=0.92). In the future, gene expression profiles from unique individuals will replace disease-specific signatures in personalised medication therapies, laying the computational groundwork for this.

3. METHODOLOGY

a) Proposed Work:

To use the K-Means technique to narrow down the enormous number of factors and identify the most likely illnesses. The more illnesses there are, the better this method is at clustering them. One of the unsupervised learning methods used to address the clustering issue is K-Mean. Finding the k centroids—one for each cluster—is the major goal. The many tests that are administered to the patients will be used as clustering characteristics. This method produces reliable results for every diagnosis by reducing the amount of iterations and ensuring that cluster boundaries are clearly defined and do not overlap. Anyone with an internet connection may use this system thanks to its Service Orientated Architecture (SOA), and the LAMSTAR Network can be utilised to compute weight, improve algorithm correctness, test speed, and provide better results.

b) System Architecture:



A system architecture, sometimes known as a systems architecture, is a conceptual model that describes the behaviour, structure, and other aspects of a system. A formal description and representation of a system that is structured to facilitate reasoning about its behaviours and structures is called an architectural description. System components, their outwardly evident characteristics, and their interrelationships can all be included in the system architecture.

c) Dataset:

The dataset consists of various symptoms associated with different medical conditions. It includes general symptoms such as cough, fever (high and mild), nausea, loss of appetite, headache, and sweating, which are commonly observed in viral infections and flu. Additionally, there are symptoms indicating potential liver or digestive issues, such as yellowish skin, dark urine, yellowing of eyes, abdominal pain, diarrhoea, and indigestion. Certain neurological or pain-related symptoms like back pain, pain behind the eyes, and constipation are also present. The dataset further includes critical symptoms such as acute liver failure, fluid overload, swelling of the stomach, swelled lymph nodes, blurred vision, sinus pressure, chest pain, and congestion, which could indicate severe infections or underlying chronic diseases. Respiratory symptoms like breathlessness, throat irritation, phlegm, runny nose, and redness of eyes suggest respiratory infections or allergic reactions. This dataset can be useful in machine learning models for disease prediction and medical decision support, helping to classify diseases based on symptom patterns.

Fig 2 Dataset

d) Data preprocessing:

Data preprocessing is an essential step in disease prediction to ensure the dataset is clean, structured, and suitable for analysis. It begins with data cleaning, where missing values are handled using techniques like mean/mode imputation, and duplicate entries are removed for consistency. Next, data transformation standardizes symptom representation by encoding categorical values such as "high fever" and "mild fever" into numerical forms, using methods like label encoding or one-hot encoding. Normalization techniques like Min-Max Scaling or Z-score normalization help maintain uniform weightage across features. Dimensionality reduction methods, such as Principal Component Analysis (PCA) and correlation analysis, eliminate irrelevant features to improve model efficiency. To prevent bias in disease classification, data balancing techniques like SMOTE (Synthetic Minority Over-sampling Technique) are applied to ensure equal representation of all classes. Once preprocessing is complete, the dataset is split into training and testing sets, optimizing it for accurate disease prediction and enhancing the model's reliability in medical decision-making.

e) Feature Extraction

Feature extraction plays a crucial role in disease prediction using machine learning by identifying the most relevant symptoms from the dataset. The given dataset consists of a wide range of symptoms, including general, gastrointestinal, respiratory, neurological, and severe conditions. Extracting meaningful features from these symptoms helps in improving model accuracy and reducing complexity.

The first step in feature extraction is data preprocessing, which includes handling missing values, standardizing symptom representation, and converting categorical symptoms into numerical form using techniques like one-hot encoding or label encoding. Next, dimensionality reduction techniques such as Principal Component Analysis (PCA) or Feature Selection Methods are applied to eliminate redundant or less significant symptoms. These techniques help in reducing computation time while maintaining prediction accuracy.

Additionally, symptom clustering using K-Means helps group related symptoms, allowing for better disease classification. For example, symptoms like high fever, cough, and breathlessness may be clustered under respiratory diseases, while yellowing of eyes, dark urine, and abdominal pain may indicate liver-related diseases. Another approach is using weight-based feature selection with a LAMSTAR neural network, which assigns importance scores to

symptoms based on their impact on disease prediction.

e) Modules:

Chemical structure Drug structure at the molecular level describes its binding

activity. Chemical _ngerprints are the most commonly used structural pro_lingmarker for drugs [13]. Fingerprints are bit vectors that indicate the presence (1) or absence (0) of certain chemical features (e.g. a C=N group, a six member ring,).We used the OpenBabel 2.3 library to take an input chemical formula (SMILESID) and generate Molecular Access System (MACCS) binary structural featurelists with lengths of 166.

Drug targets The set of targets for a drug can shed light on a_ected biologicalprocesses. We represent the set of drug targets obtained from DrugBank andKEGG as a bit vector in which 1 represents a target of the drug, and a 0represents not a target for the drug. This results in a sparse matrix, since thedrugs have a median of one putative target each.

f) Algorithms:

K Means: K-Means is an unsupervised machine learning algorithm used for clustering diseases based on patient symptoms, helping in efficient medical diagnosis. It works by selecting a predefined number of clusters (K), assigning each data point (patient symptoms) to the nearest cluster centroid, and updating centroids iteratively until they stabilize. This algorithm effectively groups similar diseases, making it useful for predicting illnesses based on symptom patterns. By categorizing patient data into distinct clusters, it enhances disease identification, reduces computational complexity, and provides clearer insights for medical professionals, ultimately improving diagnosis and treatment recommendations.

Logistic Regression: Logistic Regression is a supervised machine learning algorithm used for disease prediction by classifying patient data into distinct categories, such as the presence or absence of a disease. It estimates the probability of a condition based on input features (symptoms) using the logistic function, which maps values between 0 and 1. The algorithm assigns weights to features and applies a decision boundary to differentiate between health conditions. It is particularly useful for binary classification problems, such as predicting whether a patient has a specific disease. Logistic Regression is efficient, interpretable, and widely used in medical decision support systems.

SVM: Support Vector Machine (SVM) is a powerful supervised learning algorithm used for disease prediction by classifying patient data based on symptoms. It works by finding the optimal hyperplane that best separates different disease categories in a high-dimensional space. SVM uses kernel functions to handle complex, non-linear relationships between symptoms and diseases, making it highly effective for medical diagnosis. By maximizing the margin between data points of different classes, SVM ensures accurate predictions, even with limited data. Its robustness and ability to handle high-dimensional data make it a reliable choice for disease classification tasks.

4. EXPERIMENTAL RESULTS

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its

reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

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

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. The formula is used to calculate precision:

Precision = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: The ability of a model to identify all pertinent instances of a class is assessed by machine learning recall. The completeness of a model in capturing instances of a class is demonstrated by comparing the total number of positive observations with the number of precisely predicted ones.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

mAP: Assessing the level of quality Precision on Average (MAP). The position on the list and the number of pertinent recommendations are taken into account. The Mean Absolute Precision (MAP) at K is the sum of all users' or enquiries' Average Precision (AP) at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

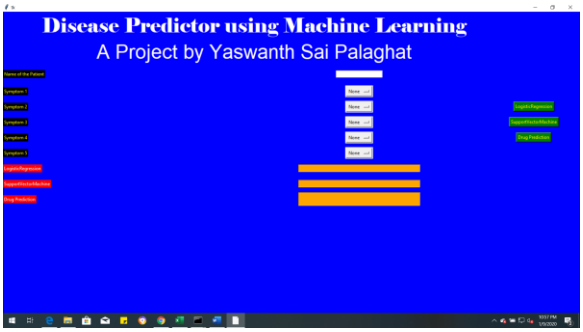


Fig 2 home page

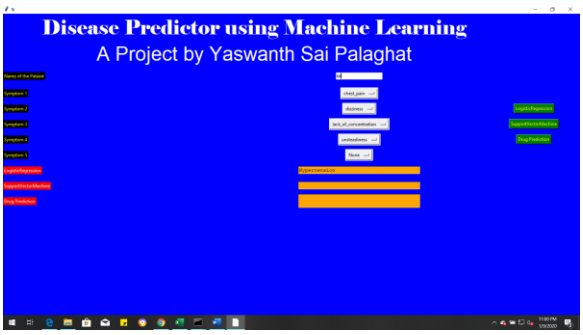


Fig 3 input page

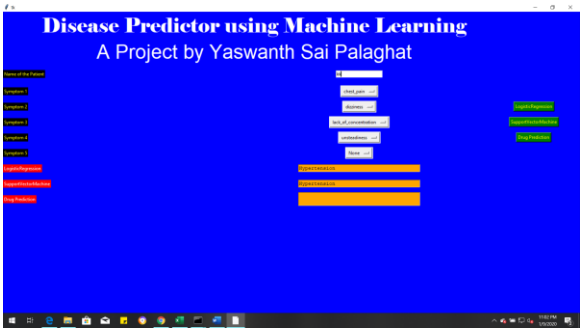


Fig 4 input page

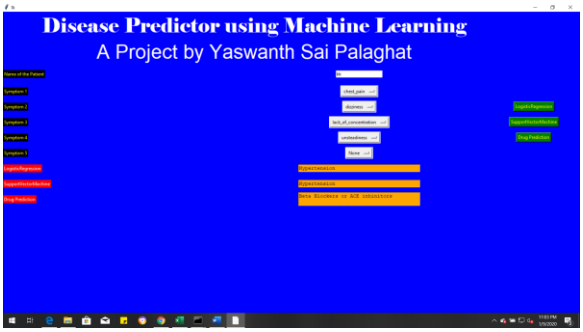


Fig 5 results

5. CONCLUSION

To confirm their theories for predicting medication indications, researchers have taken advantage of publicly available statistics. Nevertheless, the datasets are heterogeneous and dynamic, potentially leading to disparate findings for the same hypothesis. We represented, linked, and accessed pharmacological and illness data from the Bio2RDF project using Semantic Web technologies, particularly Linked Data. To train classifiers, we get medication and illness characteristics via SPARQL queries. It will be feasible to run the queries again and get fresh, updated data in the event that the data version changes. We gathered a larger set of data that included 1393 disorders and their characteristics, along with 816 medications. Predictions for gold standard data that were produced by merging data sources for various medication indications were assessed. We tested our approach using a different dataset that was generated using [23], which demonstrated to us that our method is predictable regardless of the data compiled. Failure to take into account the paired nature of inputs is a critical flaw in a standard assessment system for drug indication predictions that would result in wildly inaccurate predictions [15]. As recommended in [14] for drug target interaction prediction, we divided the data into separate train and test sets where not only pairings but also drugs/diseases do not overlap. We evaluated a number of classifiers using various cross-validation approaches and contrasted our strategy with the state-of-the-art techniques, including PREDICT and SLAMS. In disjoint cross-validation situations, we found that our prediction performance outperformed that of the SLAMS and the PREDICT.

6. FUTURE SCOPE

The future scope of disease-drug prediction using machine learning is vast and promising. With advancements in AI, deep learning, and big data analytics, the accuracy of disease diagnosis and drug recommendations will improve significantly. Integration with electronic health records (EHRs) and wearable health devices will enable real-time disease

monitoring and personalized treatment plans. The use of federated learning will enhance data privacy while improving predictive models. Additionally, explainable AI (XAI) will help in making machine learning models more interpretable for doctors. Future developments may also include AI-driven drug discovery, reducing the time and cost required for new drug development.

REFERENCES

1. Brown, A.S., Patel, C.J.: A standard database for drug repositioning. ScienticData 4, 170029 (2017)

2. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2rdf release 2: improvedcoverage, interoperability and provenance of life science linked data. In:Extended Semantic Web Conference. pp. 200{212. Springer (2013)

3. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P.: Drug target identi-cation using side-eect similarity. Science 321(5886), 263{266 (2008)

4. Chiang, A.P., Butte, A.J.: Systematic evaluation of drug{disease relationships toidentify leads for novel drug uses. Clinical Pharmacology & Therapeutics 86(5),507{510 (2009)

5. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferringnovel drug indications with application to personalized medicine. Molecularsystems biology 7(1), 496 (2011)

6. Guney, E.: Reproducible drug repurposing: When similarity does not suce. In:PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. pp. 132{143 (2017)

7. Hay, P.J., Claudino, A.M.: Bulimia nervosa: online interventions. BMJ clinicalevidence 2015 (2015)

8. Hu, G., Agarwal, P.: Human disease-drug network based on genomic expressionproles. PloS one 4(8), e6536 (2009)

9. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and sideeects. Nucleic acids research 44(D1), D1075{D1079 (2015)

10. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C.,Wrobel, M.J., Lerner,J., Brunet, J.P., Subramanian, A., Ross, K.N., et al.: The connectivity map: usinggene-expression signatures to connect small molecules, genes, and disease. science313(5795), 1929{1935 (2006)

11. Larrosa, O., de la Llave, Y., Barrio, S., Granizo, J.J., Garcia-Borreguero, D.: Stimulantand anticataplectic eects of reboxetine in patients with narcolepsy: a pilotstudy. Sleep 24(3), 282{285 (2001)

12. Lemke, M.R.: Eect of reboxetine on depression in parkinson's disease patients.The Journal of clinical psychiatry 63(4), 300{304 (2002)

13. Melville, J.L., Hirst, J.D.: TMACC: Interpretable Correlation Descriptors forQuantitative StructureActivity Relationships. J. Chem. Inf. Model. 47(2), 626{634 (Mar 2007), http://dx.doi.org/10.1021/ci6004178

14. Pahikkala, T., Airola, A., Pietil□a, S., Shakyawar, S., Szwajda, A., Tang, J., Aittokallio,T.: Toward more realistic drug{target interaction predictions. Briengs inbioinformatics 16(2), 325{337 (2014)

15. Park, Y., Marcotte, E.M.: Flaws in evaluation schemes for pair-input computationalpredictions. Nature methods 9(12), 1134{1136 (2012)

16. Ratner, S., Laor, N., Bronstein, Y.,Weizman, A., Toren, P.: Six-week open-label reboxetinetreatment in children and adolescents with attention-decit/hyperactivitydisorder. Journal of the American Academy of Child & Adolescent Psychiatry44(5), 428{433 (2005)

17. Schmidt, C., Leibiger, J., Fendt, M.: The norepinephrine reuptake inhibitor reboxetineis more potent in treating murine narcoleptic episodes than the serotoninreuptake inhibitor escitalopram. Behavioural brain research 308, 205{210 (2016)

18. Silveira, R.O., Zanatto, V., Appolinario, J., Kapczinski, F.: An open trial of reboxetinein obese patients with binge eating disorder. Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity 10(4), e93{e96 (2005)

19. Tehrani-Doost, M., Moallemi, S., Shahrivar, Z.: An open-label trial of reboxetinein children and adolescents with attention-decit/hyperactivity disorder. Journalof child and adolescent psychopharmacology 18(2), 179{184 (2008)

20. Versiani, M., Cassano, G., Perugi, G., Benedetti, A., Mastalli, L., Nardi, A., Savino,M.: Reboxetine, a selective norepinephrine reuptake inhibitor, is an eective andwell-tolerated treatment for panic disorder. The Journal of clinical psychiatry(2002)

21. Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A.,Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes,A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers,R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J.,'t Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons,A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone,S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M.,Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P.,

Wolstencroft, K., Zhao, J., Mons, B.: The fair guiding principles for scientic datamanagement and stewardship. Scientic Data 3 (2016)

22. Yang, L., Agarwal, P.: Systematic drug repositioning based on clinical side-eects.PloS one 6(12), e28025 (2011)

23. Zhang, P., Agarwal, P., Obradovic, Z.: Computational drug repositioning by rankingand integrating multiple data sources. In: Joint European Conference on MachineLearning and Knowledge Discovery in Databases. pp. 579{594. Springer

(2013)