2019

COMPUTATIONAL DRUG REPURPOSING FOR BREAST CANCER SUBTYPES

Roopesh Dhara
University of Windsor

COMPUTATIONAL DRUG REPURPOSING FOR BREAST CANCER SUBTYPES

by

Roopesh Dhara

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2019

COMPUTATIONAL DRUG REPURPOSING FOR BREAST CANCER SUBTYPES

by

Roopesh Dhara

APPROVED BY:

H. Wu

Department of Electrical and Computer Engineering

L. Rueda

School of Computer Science

A. Ngom, Advisor

School of Computer Science

April 30, 2019

Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Breast cancer makes up 25 percent of all new cancer diagnoses globally according to the American Cancer Society(ACS). Developing a highly effective drug can be a time consuming and an expensive ordeal. Drug repurposing is a tremendous approach which takes away some disadvantages of traditional drug development procedures making it both time and cost effective. In this thesis, we are interested in finding good drugs for each of the ten subtypes of breast cancer. Repurposing incorporates identifying unique indications of pre-approved drugs and utilizing them to observe the anti-correlation between the perturbation data and disease data. If anti-correlation, whether it is up-regulation or down-regulation, is detected, it indicates that those drugs cause an effect making them a suitable candidate for drug repurposing. The gene expression data and the discrete copy number variation data will be used to compute z-scores and normalize the data for ten sets of disease subtypes. Gene expression data for ten subtypes was extracted from the METABRIC dataset. We have extracted values corresponding to MCF7 cell line from the pharmacogenomics perturbation data which is the National Institute of Health's (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) dataset. We have used our proposed clustering methods to select the best suited drug candidates per subtype. We have obtained a ranked list of suitable drug repurposing and repositioning candidates for each of the 10 breast cancer subtypes.

Dedication

I would like to dedicate this thesis to my father D. V. C. Prasad, mother D. V. Durga Bai, and brother Rakesh Dhara, who have always encouraged me to take this program. Without their unconditional love and support, I could not have achieved any of this.

Acknowledgements

I would like to sincerely express my most profound gratitude towards my supervisor Dr. Alioune Ngom, for his support, guidance, encouragement and valuable advice. Thank you for being patient and training me to think better.

I would like to thank my internal reader Dr. Luis Rueda and my external reader Dr. Huapeng Wu for all their valuable inputs and suggestions given to me.

I would also like to thank my friends Sindhuja, Amangel, Emamuzo, Satya, Sowndarya, Parth, Priya, and more each of whom helped me in their own way during my thesis.

Finally, I would like to thank my parents D.V.C. Prasad, D.V. Durga Bai and my brother Rakesh Dhara, for their immense support from the very beginning.

Collectively, all of their support and guidance has enabled me to successfully complete my masters program.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Drug Repurposing

1.1.1 What is a drug?

A drug is any chemical substance which is administered to living organisms to generate a biological effect. The structure of these chemical substances is known. There are a wide variety of drugs, each capable of causing different physiological and sometimes psychological effect to living beings induced with the drug [31].

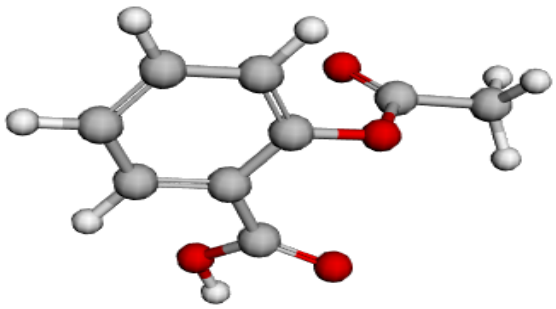

Figure 1.1: 3D molecular structure of Aspirin

In other words, a drug is used to cure a disease and alleviate any symptoms of illnesses. Figure 1.1 shows the 3D molecular structure of a drug known and sold as Aspirin [12], the most common drug in the world. There are some drugs which are not used to specifically treat a particular disease but act as a psychoactive chemical substance influencing a better mood by impacting the central nervous system.

1.1.2 Traditional drug discovery and development process

Traditional drug discovery and development procedures can be highly time consuming and come at exorbitant development costs. For instance, it would take 15 years on average and billions of dollars usually put into the various steps needed to successfully discover and develop an effective drug [17].

Figure 1.2 shows the steps involved in traditional drug discovery and development procedure. Step 1, disease related genomics, involves studying those genes or set of genes which are responsible for causing a particular disease. Step 2, target identification and validation, deals with identifying the target. A target is a pathogen on which the drug is meant to cause an effect on. Step 3, lead discovery and optimization, is one of the initial stages of drug discovery process where the small molecules (drugs) are carefully vetted to observe traces of lead compound, a pharmacological chemical. Furthermore, these undergo thorough optimization before making it to the pre-clinical trials. Step 4, pre-clinical trials, comprises of scrutinizing the dosage level of drugs thereby ensuring that the drug is safe. This phase is essential before proceeding to clinical trials. Since a drug cannot be used on humans without having the knowledge of whether it is safe to consume or not, these trials are conducted on other species that have genetics resembling human genetics. Step 5, clinical trials are where drugs are tested on humans to study their effect before making it available for purchase.

Figure 1.2: Traditional drug development process

To minimize the time and costs associated with traditional drug discovery process drug repurposing is a preferred alternative.

1.1.3 What is drug repurposing?

Drug repurposing, is a technique that makes use of unapproved drugs which have passed the initial phases of drug discovery process and are categorized as approved, experimental, investigational, withdrawn, unknown, illicit, vet approved, or nutraceutical and perform clinical trials in them only if they have been observed to share similarities with approved drugs intended to treat a specific disease [23].

Drug repositioning, is a technique in which approved drugs currently being used to treat another disease could be used to treat a different target disease [32].

The drugs used in our research fall under five of these categories as listed below.

- Approved
- Experimental
- Investigational
- Withdrawn
- Unknown

Approved drugs are those that have passed clinical trials. Experimental drugs are those that have shown to bind proteins in mammals or bacteria. Investigational drugs are at one of the phases of drug design process in one jurisdiction or more. Withdrawn drugs are those that were once approved but have lost their approval status for any reason. There is not enough data on unknown drugs as these are in the preliminary stages of drug discovery. When referring to drugs from various drug categories such as experimental, investigational, withdrawn, unknown etc., we shall collectively call these 'unapproved drugs'.

Upon performing experiments using various methods, we observe which unapproved drugs display similar properties to those of the approved ones. The unap-

proved drugs which closely resemble the properties of approved drugs intended to treat breast cancer shall be selected as suitable candidates for drug repurposing.

1.1.4 What is a gene?

Figure 1.3 shows a gene [22] which is a unit of DNA responsible for relaying genetic traits. Every data point generated by a DNA microarray experiment denotes the ratio of expression levels. The results from one experiment with **n** number of genes on one test subject denotes a series of expression levels. In each of these ratios, the numerator represents expression level of the gene in a varying condition and the denominator denotes the expression level of the gene in a reference condition. Data compiled together to form **m** such experiments presents a gene expression matrix. The gene expression value will be positive if the production of that gene is increased in that particular test case and will be negative if the generation of that gene is decreased instead. [2]

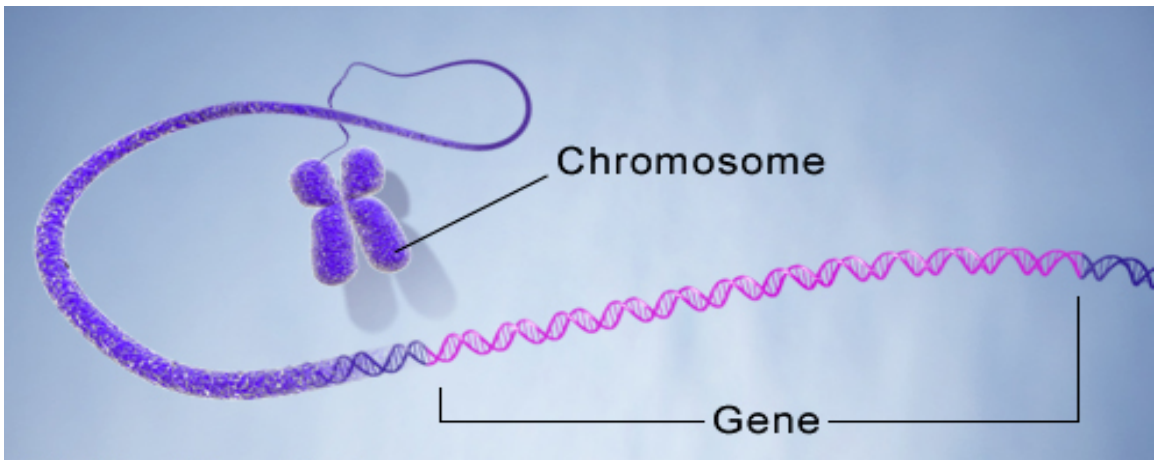

Figure 1.3: Gene

1.2 Breast Cancer

1.2.1 What is breast cancer?

Breast Cancer occurs when the cells in the breast begin to grow out of control and form a tumour that can be seen on an x-ray or be observable as a lump by touch. The tumour is malignant (cancer) if the cells can grow into and invade the surrounding tissues or spread (metastasize to distant areas of the body. Breast cancer occurs almost entirely in women, but men can get it too. Overall, there are ten subtypes of breast cancer. In this thesis, we obtain suitable drug candidates for each of these ten subtypes. [29]

1.2.2 How is breast cancer affecting women globally?

Breast Cancer makes up 25% of all new cancer diagnoses in women across the globe according to the American Cancer Society (ACS). [1]

In Canada:-

- 26,300 women were diagnosed with breast cancer, which represents 25% of all new cancer cases in women in 2017 [FIG 1 INSERT from PPT]
- 5000 women died from breast cancer. This represents 13% of all cancer deaths in women in 2017 [FIG 2 INSERT from PPT]
- On average, 72 Canadian women were diagnosed with breast cancer everyday
- On average, 14 Canadian women died from breast cancer everyday

1.3 Problem Definition

Given drug perturbation data and gene expression data for all ten subtypes of breast cancer, we aim to obtain a ranked list of drugs which would make suitable candidates for drug repurposing and drug repositioning for all ten breast cancer subtypes. We achieve this by performing preprocessing steps such as calculation of z-scores in the METABRIC dataset and calculation of p-value and q-value in the LINCS dataset to filter the differentially expressed genes and the drugs respectively. We then make use of two machine learning methods such as a centroid based clustering model, K-means clustering and a connectivity based clustering model, agglomerative clustering. Then using Euclidean distance we are able to provide a ranked list of good drug repurposing and repositioning candidates for each of the ten subtypes.

1.4 Thesis Motivation

Researching the repurposing of unapproved drugs sharing similarities with approved drugs intended to treat breast cancer would help speed up the drug design process involving drug discovery and development phases. As a result, years of time and billions of dollars will have been conserved in an effort to help cure breast cancer disease. Most importantly, this thesis does its part in helping us move one step closer to acquiring suitable drugs to tackle breast cancer.

1.5 Contribution

In this thesis, we have proposed application of existing preprocessing and clustering methods on all ten breast cancer subtypes to obtain a ranked list of suitable drug repurposing and repositioning candidates for each of the 10 subtypes.

1.6 Thesis Organization

The rest of the thesis/ research work is organized in the following manner.

- In Chapter 2, we discuss literature review in the area of drug repurposing using computational approaches
- In Chapter 3, we introduce our proposed approach and explain all the techniques used to obtain suitable drug repurposing candidates for each of the ten subtypes of breast cancer
- In Chapter 4, we present the experimental results and perform an analysis of those results.
- Chapter 5 concludes the research by explaining insights received during the work and setting up the field of opportunities for possible future work

Chapter 2

Literature Review

This chapter consists of some literature regarding computational drug repurposing using cancer data.

2.1 Existing approaches on drug repurposing

There have been several researchers whose contribution to drug repurposing is worth noting. We discuss some of those works below.

2.1.1 Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures

The authors of this paper, Lee et al. [20] have developed a series of seven classifiers using logistic regression to predict drug repurposing candidates for the treating of glioblastoma, lung cancer, and breast cancer. Their method makes use of three types of signatures obtained from the chemical structure (S), drug-target relation (T), and gene expression data (E). Suitable drug repurposing candidates were predicted on the basis of similarity of the aforementioned signatures between the compounds and disease or known its drugs. The authors have carefully observed the prediction perfor-

mance in a completely unbiased way. The observations were conducted in three ways, (i) using a cross-validation scheme using known drugs as a benchmark, (ii) 29 anti-cancer HTS datasets for 11,000 to 40,000 compounds, and (iii) assays of glioblastoma cancer cell lines and patient-derived primary cells.

2.1.2 Integrative cancer pharmacogenomics to infer large-scale drug taxonomy

In this paper, the authors, Haibe-Kains et al. [15] have developed an integrative taxonomy inference approach, Drug Network Fusion (DNF), making use of pharmacological phenotypes and transcriptional perturbation profiles. The authors of this paper used DNF to perform a comparison between their integrative taxonomy, single-layer drug taxonomies and other published methods used to predict drug targets. Their results showcase that DNF is superior towards drug classification while highlighting singular data types pivotal for predicting drug groups in terms of anatomical classification as well as drug-target interactions. The results produced by DNF indicate that drug-drug relationships serve as a good way to predict new drug mechanism of actions (MoA) which are uncharacterized compounds representing a challenge in drug development.

2.1.3 A novel computational approach for drug repurposing using systems biology

The authors of this paper, Draghici et al. [24] built a global network (GN) which is the union of all KEGG human signaling pathways. They have extracted a subgraph of GN for each drug-disease pair and termed it drug-disease network (DDN) comprising of the shortest paths between two sets of disease related genes and drug targets. The authors have applied a system level analysis on the gene expression signatures of drug-

disease pairs to generate gene perturbation signatures in the drug-disease network. They have further assigned a repurposing score on each drug-disease pair then finally obtained a ranked drug list with potential therapeutic effects for the given disease on the basis of the aforementioned repurposing scores.

2.1.4 Breaking the paradigm: Dr Insight empowers signature free, enhanced drug repurposing

The authors of this paper, Gu et al. [7], have worked to overcome the limitations of existing computational frameworks by developing Dr. Insight, which offers signature-free, optimal drug repurposing based on gene expression data. It takes into account the dysregulation of gene expression from both disease and drug-perturbed data simultaneously, which renders the CEG's as optimal features to investigate the connections among diseases, drugs and genes. Dr. Insight has broken the computational bottleneck for transcriptome-based drug discovery, which provides an unbiased first look from novel redirections of existing drugs towards a systematic understanding of disease-specific drug mode of actions at a molecular level.

2.1.5 A new computational drug repurposing method using established disease-drug pair knowledge

In this paper, the authors, Draghici et al. [28] have worked towards obtaining drug repurposing candidates for three diseases. They have used GEO disease data for breast cancer, CMAP data for rheumatoid arthritis, and LINCS for idiopathic pulmonary fibrosis. Their workflow consists of transforming the input matrix into a lower dimensionality matrix by incorporating dimensionality methods such as principal component analysis (PCA) or Locally Linear Embedding (LLE). Then the authors have used leveraged the known relationship between disease and its FDA approved

drugs into a transformed space using distance metric learning. In this space, the clinically relevant drugs get close to the disease so their euclidean distance can be computed and ranked from the closest to farthest drugs from the disease. The authors have used five algorithms on each of these datasets per disease and performed a comparative analysis of their results.

Chapter 3

Proposed Methods

In this chapter, we discuss the datasets, preprocessing steps taken, and machine learning techniques used in this thesis.

A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**. This is a modern definition of machine learning [33]. There are two different types of commonly known learning algorithms as listed below.

- Supervised Learning
    - Classification
    - Regression
- Unsupervised Learning
    - Clustering

Supervised learning mainly is subdivided into classification and regression whereas unsupervised learning comprises of cluster analysis also known as clustering.

Classification algorithms have a simple task of classifying objects and assigning them into one of the categories. Perfect classification is nearly impossible to achieve and there are almost always some objects that are misclassified. Classification takes place based on the features provided in the dataset after performing feature extraction. [14] Figure 3.1 displays an example of classification using support vector machine (SVM) classifier.

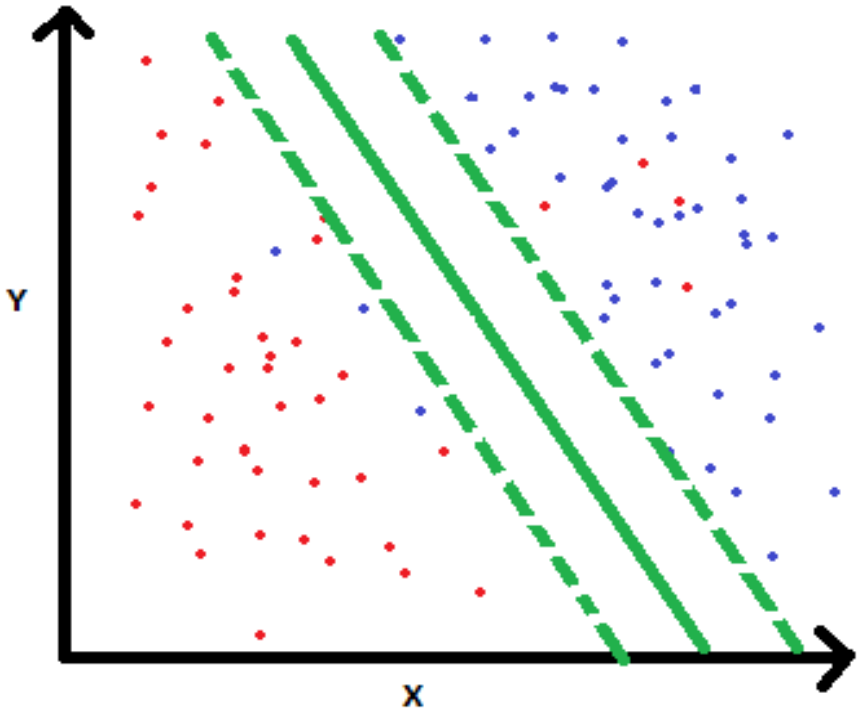

Figure 3.1: Classification example

Regression analysis is a procedure in which a set of statistical processes provide an estimate of the given variables. Generally speaking, in a set of many variables, we may predict the dependent variables' future value by changing the independent variables. The dependent variable is the output variable whose value is being predicted based on the changes in independent variables. [11] Figure 3.2 showcases an example of simple linear regression output.

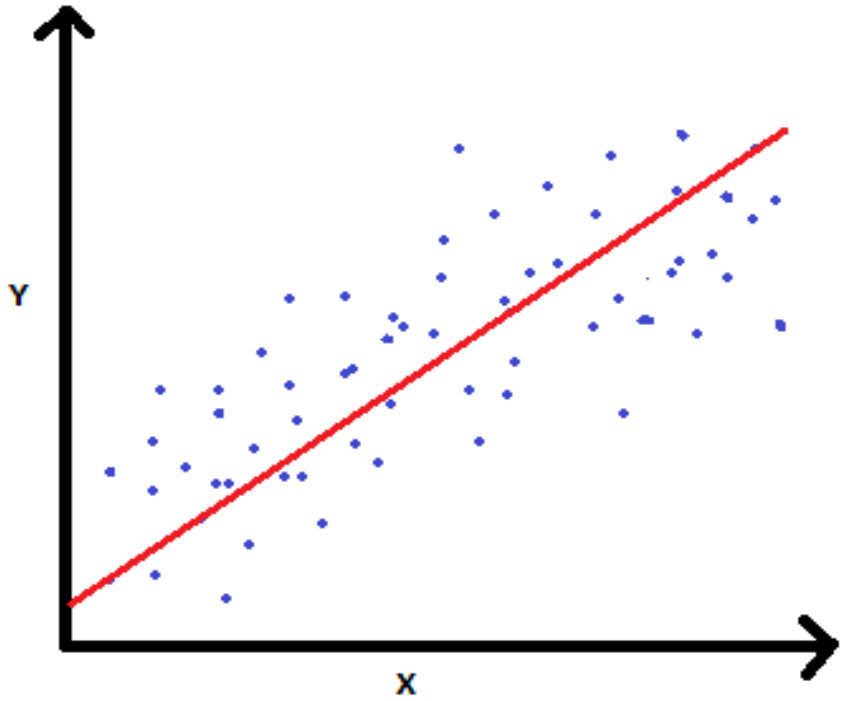

Figure 3.2: Regression example

Clustering is a procedure in which a sizeable group of objects are distinguished and brought together to be categorized in multiple clusters of data. Each of those clusters or categories of data would comprise of objects of the same group that are similar in terms of their properties.[18] Figure 3.3 displays an example of a simple clustering outcome where objects have been grouped into three separate clusters showing that the objects belonging to each one of those clusters share similarities with the rest of them within the same cluster or group.

We will see more about clustering and two of the clustering algorithms used in this thesis later in this chapter. The two clustering algorithms used are as follows:-

- K-Means clustering
- Agglomerative clustering

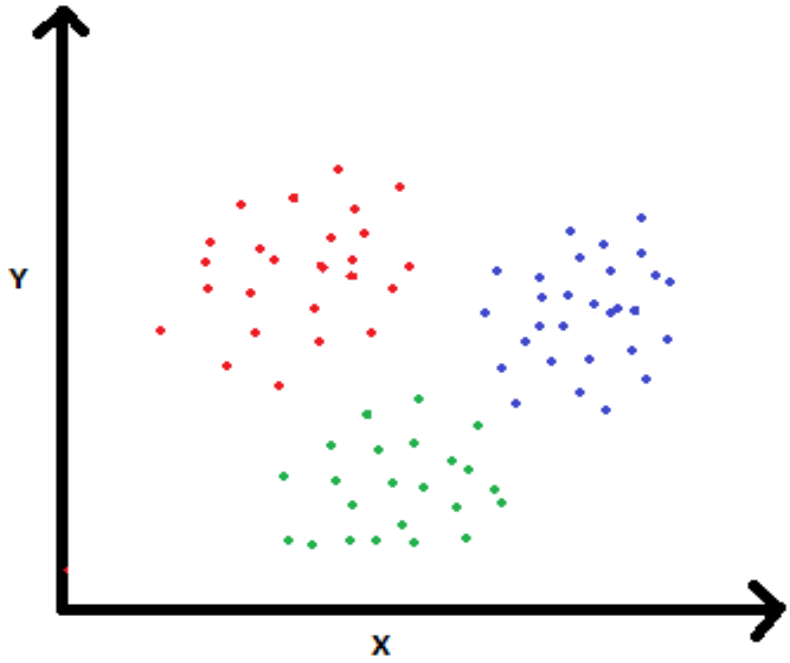

Figure 3.3: Clustering example

3.1 Datasets

3.1.1 METABRIC

METABRIC is an abbreviation for Molecular Taxonomy of Breast Cancer International Consortium. This dataset consists of gene expression data for a large pool of breast cancer genes. These expression values are arranged for each test subject column wise. It has 1904 test subjects as columns and 24,368 genes as its rows.

Studies conducted on a large cis-associated gene pool of breast cancer genes show that biological subtypes were found using joint clustering of copy number abberations (CNAs) and gene expression data. 10 groups were suggested based on Dunn's index. [3] Breast cancer is essentially 10 different diseases where each of them contains a different molecular fingerprint. [10] Cis-regulatory elements (CREs) are those regions

in DNA which are responsible for regulation of transcription of neighbouring genes. [34]

List of 10 Subtypes of Breast Cancer

Integrative subtypes have shown to occur at various frequencies and so in concentrating sequencing efforts on these subtypes could prove to benefit those working on a resolving it at a sequence-level [10].

Knowing that there are 10 subtypes of breast cancer means that patients can get treatment based on the specific genetic fingerprints of their tumors.

The 10 subtypes are listed as follows:- [10]

- Subtype 1: ER+, luminal B tumours
- Subtype 2: ER+, luminal tumours
- Subtype 3: Luminal A tumours
- Subtype 4: CNA-devoid (mixed subgroups, both ER+ and ER-)
- Subtype 5: ER-, HER2-enriched and ER+, luminal tumours
- Subtype 6: ER+, luminal tumours
- Subtype 7: Luminal A 16p gain/16q loss
- Subtype 8: Luminal A 1q gain/16q loss
- Subtype 9: ER+, mixed subgroup
- Subtype 10: Basal-like tumours

Network Biomarkers

Network biomarkers provide us with an interaction value indicating the protein-protein interaction level within the genes belonging to each subtype. In this thesis, we have extracted all gene pairs whose interaction level was between 50 to 100 (both inclusive). We have conducted this process for each of the 10 subtypes. Interaction level indicates that the presence of one gene in a gene pair is dependent on the other within its pair. The higher the interaction level value, the greater the dependency.

Interaction networks usually are comprised of gene regulatory network, protein-protein interaction network, RNA network, etc.[35] They can give information about the models of cellular networks on the basis of large and heterogeneous dataset integration [16].

Table 3.1: Gene count per subtype

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
73	50	129	147	61	155	96	94	95	76

Table 3.1 shows the number of differentially expressed genes used in each of the ten subtypes labelled as S1, S2, ..., S10.

CNA

Copy number aberrations (CNAs) are changes in copy number that have occurred in somatic tissue, for instance, in a tumor. [25] In other words, copy number aberration (CNA) means that a chromosome has a duplicated section of DNA instead of having one section.

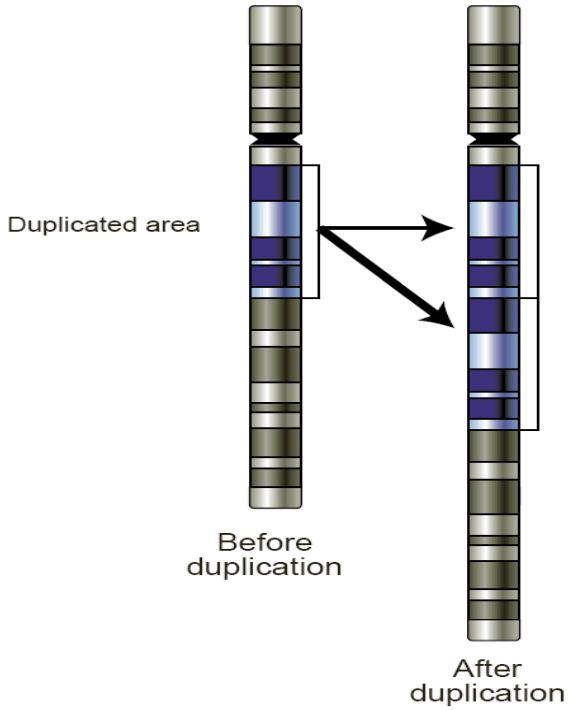

Figure 3.4: Gene duplication

The gene duplication occurring in this process is a mechanism in molecular evolution wherein a new genetic material is created as shown in figure 3.4. The CNA file is comprised of 0's and 1's where 0 indicates that the gene is diploid. These expression values are the only ones which we use while normalizing the expression data. We will look into normalization in the preprocessing section.

3.1.2 LINCS

The drug data was extracted from the pharmacogenomics perturbation data which is the National Institute of Healths (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) dataset. This dataset consists of 21,567 breast cancer drugs in the columns and 12,328 genes in the rows. This dataset consists of normalized z-score values as it is a level 5 LINCS dataset. Level 4 LINCS data consists of two sets of data, before administration of drugs and after administration of drugs onto the genes in the dataset. These expression values from both the level 4 datasets are normalized

to form the level 5 LINCS dataset. Here is a list of the 5 levels of LINCS data.

- Level 1: Raw data
- Level 2: Processed raw data
- Level 3: Normalized per sample data
- Level 4: Signatures (used for connecting perturbation data)
- Level 5: Perturbation data

We go through the preprocessing steps showing how we generated the drug disease combination matrices and the anti-correlation matrices in the next section.

3.2 Preprocessing of datasets to be used in machine learning techniques

The preprocessing pipeline used on the datasets to be used in the methods in this thesis is explained in the following subsections.

3.2.1 METABRIC dataset

We have extracted differentially expressed genes for each of the 10 subtypes from the METABRIC disease dataset based on whether the interaction level between gene pairs was 50 or greater on a scale of 1 to 100. We have removed all other gene pairs whose interaction level was known to be below 50.

Z-score normalization indicates that the data is linearly transformed and allows for comparison of two scores which originate from different normal distributions. In statistics, this score is regarded as a common standard.

Using the copy number aberration (CNA) data, we are able to identify all diploid cases in the disease gene expression data for a given set of samples. We generate z score normalized expression values for each gene. The mean and standard deviation of the gene expression values for all the patient samples are calculated in which the gene is diploid. A gene is considered to be diploid if its copy number aberration value is 0. If any of the genes have no diploids in the entire sample set, then the normalized value is denoted by NA. On the other hand, the formula used to compute the normalized expression value is:

$$(r - mu)/sigma \tag{3.1}$$

where,

r is the expression value

mu is the mean of all diploid values in the dataset

sigma is the standard deviation of all diploid values in the dataset

The following is the algorithm to calculate z-score normalized values per gene for each of the 10 subtypes [6].

Algorithm 1: Z-score calculation

```
for each gene {
    detect diploid cases using CNA
    compute mean and standard deviation of expression values of diploid cases
    for each case {
        z-score ← (expressionvalue − mean)/standarddeviation
        }
    }
```

The average of all the z-scores per gene are computed and one vector with a z-score

per gene in each of the 10 subtypes is obtained as shown in figure 3.5.

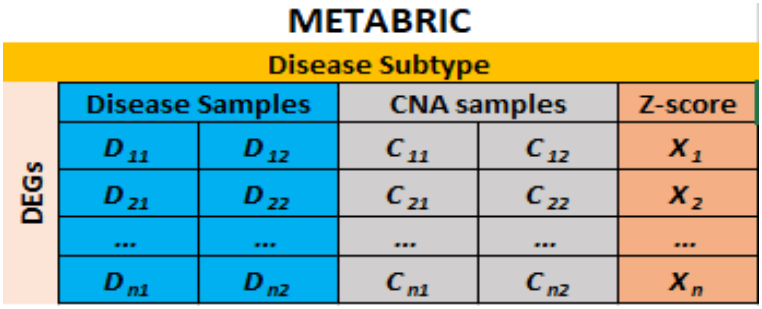

Figure 3.5: METABRIC

3.2.2 LINCS drug perturbation data

We compute p-values for each gene per drug profile to select the statistically significant values. [27] The cut off set here is 0.05 which means that there is a 5% chance that we are choosing a false positive among the list of DE genes. Having the possibility of there being a large number of false positives is not statistically good and so we calculate the q-values using the false discovery rate (FDR) approach. The older approaches reduced the number of false positives while also reducing the number of true positives which is not optimal. This newer FDR approach gives us adjusted p-values in every test case. In simpler terms, p-value predicts that there could be 5% false positives in the entire list of DE genes whereas q-value (FDR-adjusted p-value) predicts that there could be 5% false positives in the significant tests. Significant tests are those values that are deemed to be true positives based on the p-value.

Figure 3.6 shows the LINCS dataset before computing the p-values. We compute the p-value per gene based on the z-score as shown in figure 3.7. As shown in figure 3.8, we proceed to calculating the q-value per gene and discard those drug profiles which contain less than 1% DE (differentially expressed) genes. Here, a DE gene is defined by a gene that has a q-value that is less than 0.05. So in our dataset, out of 12,328 genes, we checked if there are less than 123 DE genes in a drug profile or not. If a drug profile meets this criterion, we discard it. However, if it has 123 or more DE

genes, we extracted those drugs to include in our drug disease combination matrix which we will see in the next subsection.

LINCS

	$Drug_1$	$Drug_2$	...	$Drug_m$
Genes	Z-score	Z-score	...	Z-score
G_1	Y_{11}	Y_{12}	...	Y_{1m}
G_2	Y_{21}	Y_{22}	...	Y_{2m}
...	...	...	...	...
G_n	Y_{n1}	Y_{n2}	...	Y_{nm}

Figure 3.6: LINCS

LINCS

	$Drug_1$		$Drug_2$		...		$Drug_m$	
Genes	Z-score	P-value	Z-score	P-value	...	...	Z-score	P-value
G_1	Y_{11}	P_{11}	Y_{12}	P_{12}	...	...	Y_{1m}	P_{1m}
G_2	Y_{21}	P_{21}	Y_{22}	P_{22}	...	...	Y_{2m}	P_{2m}
...	...	...	...	...	...	...	...	...
G_n	Y_{n1}	P_{n1}	Y_{n2}	P_{n2}	...	...	Y_{nm}	P_{nm}

Figure 3.7: LINCS: P-value

LINCS

	$Drug_1$			$Drug_2$			...			$Drug_m$		
Genes	Z-score	P-value	q-value	Z-score	P-value	q-value	...	...	...	Z-score	P-value	q-value
G_1	Y_{11}	P_{11}	q_{11}	Y_{12}	P_{12}	q_{12}	...	...	...	Y_{1m}	P_{1m}	q_{1m}
G_2	Y_{21}	P_{21}	q_{21}	Y_{22}	P_{22}	q_{22}	...	...	...	Y_{2m}	P_{2m}	q_{2m}
...	...	...	...	...	...	...	...	...	...	...	...	...
G_n	Y_{n1}	P_{n1}	q_{n1}	Y_{n2}	P_{n2}	q_{n2}	...	...	...	Y_{nm}	P_{nm}	q_{nm}

Figure 3.8: LINCS: q-value

Out of 7 cell lines, we have extracted drugs belonging to the cell line "MCF7". This way we have multiple entries of most drugs so we have filtered them based on the dosage and time under administration. Within this cell line, we have filtered drugs whose dosage was 1.11 um and whose time under administration was 24 h. This step

has enabled us to select unique instances of all drugs fitting our criteria. We have extracted a total of 177 drugs based on these filters.

3.2.3 Drug Disease Combination Matrix

As shown in figure 3.9, the resulting matrix from the previous steps contains the reversed z-score vector of all the DE genes from the disease subtype as the first row and all other rows comprise of z-scores of those very DE genes from each of the drug profiles. In this thesis, we call this matrix, drug disease combination matrix (DDCM). We generate 10 such drug disease combination matrices, one for each subtype. Upon applying anti-correlation on all 10 DDCMs, we arrive at a reduced list of potential drug repurposing candidates per disease subtype.

	DEGs			
Disease Subtype	$-(X1)$	$-(X_2)$	...	$-(X_n)$
$Drug_1$	Y_{11}	Y_{12}	...	Y_{1n}
$Drug_2$	Y_{21}	Y_{22}	...	Y_{2n}
...	...	...	...	...
$Drug_m$	Y_{m1}	Y_{m2}	...	Y_{mn}

Figure 3.9: DDCM

3.2.4 Anti-correlation Matrix

After generating the drug disease combination matrices, we proceed to making the anti-correlation matrices in an effort to detect potential drugs which can be used in the experiments after this. In figure 3.10, 'A' value here can be '0', '+1', or '-1'. For instance, if the disease subtype's z-score value is negative for a DE gene, and if drug m's z-score value for the same DE gene is moving in the opposite direction then we assign it a '+1'. This is because, a negative z-score indicates that a particular gene is being down-regulated and a reverse change means the application of this drug would

cause up-regulation. Similarly, if the disease subtype's z-score value is positive, and if drug1's z-score value is moving in the opposite direction then we assign it a '-1'. This is because, a positive z-score indicates that a particular gene is being up-regulated and a reverse change means the application of this drug would cause down-regulation. If this moves in the same direction, we assign a '0'.

Down-regulation indicates a decrease in the production of that gene as an effect of the disease. Up-regulation indicates a increase in the production of that particular gene as an effect of the disease.

	DEGs			
Disease Subtype	-(X1)	$-(X_2)$	...	$-(X_n)$
$Drug_1$	A_{11}	A_{12}	...	A_{1n}
$Drug_2$	A_{21}	A_{22}	...	A_{2n}
...	...	...	...	...
$Drug_m$	A_{m1}	A_{m2}	...	A_{mn}

Figure 3.10: Anti-correlation Matrix

If there are 50% or less number of 0's in any drug profile, then we keep the drug otherwise we discard this drug as it does not show much potential in reversing the up-regulation or down-regulation caused by the disease genes.

This method allows us to filter out drugs for each subtype which leaves us with a different set of drugs from the 177 in each of the 10 subtypes.

FDA Status

Now that we have the processed, filtered list of drugs, we have used online drug databases such as DrugBank [13], Kegg [19] etc.[26][8][30][21][4][5] to obtain each drugs' FDA status. All of the drug databases used to search for FDA status are listed in bibliography.

3.3 Machine learning techniques

In this thesis, we have used clustering methods followed by a distance measure to rank the drugs in order of drugs that appeared closest to farthest from the disease subtype. Upon performing clustering on the DDCMs, we select the cluster which grouped a set of drugs along with the disease subtype. All drugs within this cluster are chosen and we compute Euclidean distance for all of those from the disease subtype. We now have a list of drugs ranked from potentially best suited drug repurposing candidates for this subtype to potentially less effective drug repurposing candidates.

3.3.1 Cluster analysis

Cluster analysis or, simply put, clustering is a task in which objects are grouped together in clusters wherein all objects present in a particular cluster are deemed to share similar properties with each other and all objects in one cluster are considered to have different properties to those belonging to another cluster. Clustering by itself is a procedure of grouping and dividing objects of similarities and dissimilarities respectively. There are several types of algorithms used for clustering, two of which we have implemented in our thesis.

- K-Means clustering
- Agglomerative clustering

3.3.2 K-Means clustering

This is a centroid-based clustering algorithm where the number of clusters are predefined before running the program. We have taken four different **k=(2,3,4,5)** values in this algorithm where **k** signifies the number of clusters. The algorithm finds the **k** cluster centers so that it can assign the drugs to their nearest cluster center while

making sure that the squared distances from the cluster are diminished. This is an NP-hard optimization problem indicating that it should search for approximate solutions only. K-means algorithm finds a local optimum. We set the initialization parameter to 10 so k-means algorithm will run 10 times with different centroids each time randomly picked and the final result will be the best outcome of these 10 initializations.

We chose these **k** values because our filtered drug data consisted of 5 categories of drugs (approved, experimental, investigational, unknown, withdrawn). One of the biggest drawbacks of this algorithm is that the number of clusters need to be defined in advance. In this algorithm, each observation (drug) belongs to at least one cluster. At the same time, no observation belongs to more than one cluster. [9]

Some of the advantages of this algorithm are that it functions very efficiently when presented with large datasets as the input and it usually produces tighter clusters.

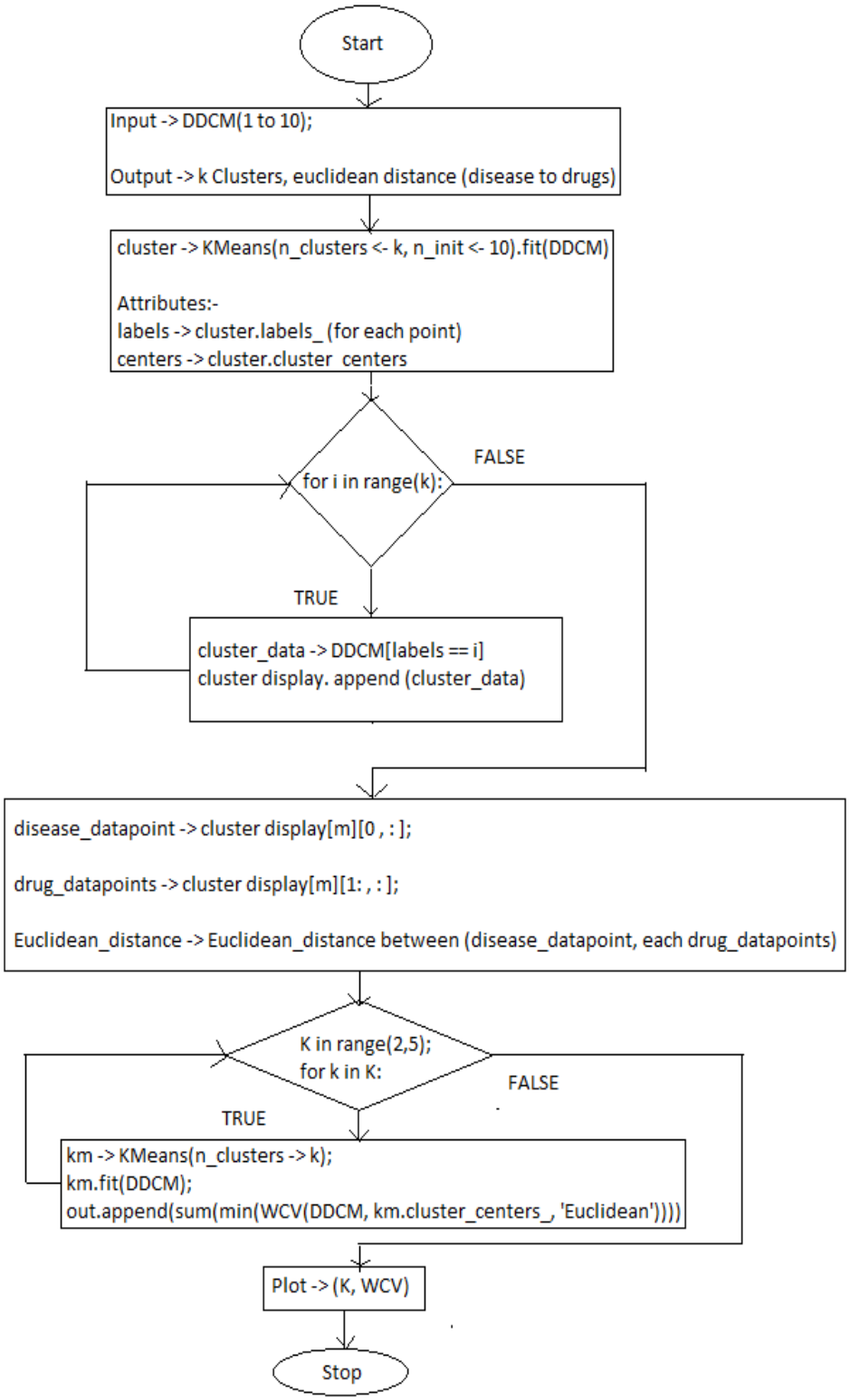

Figure 3.11: Proposed K-Means Working

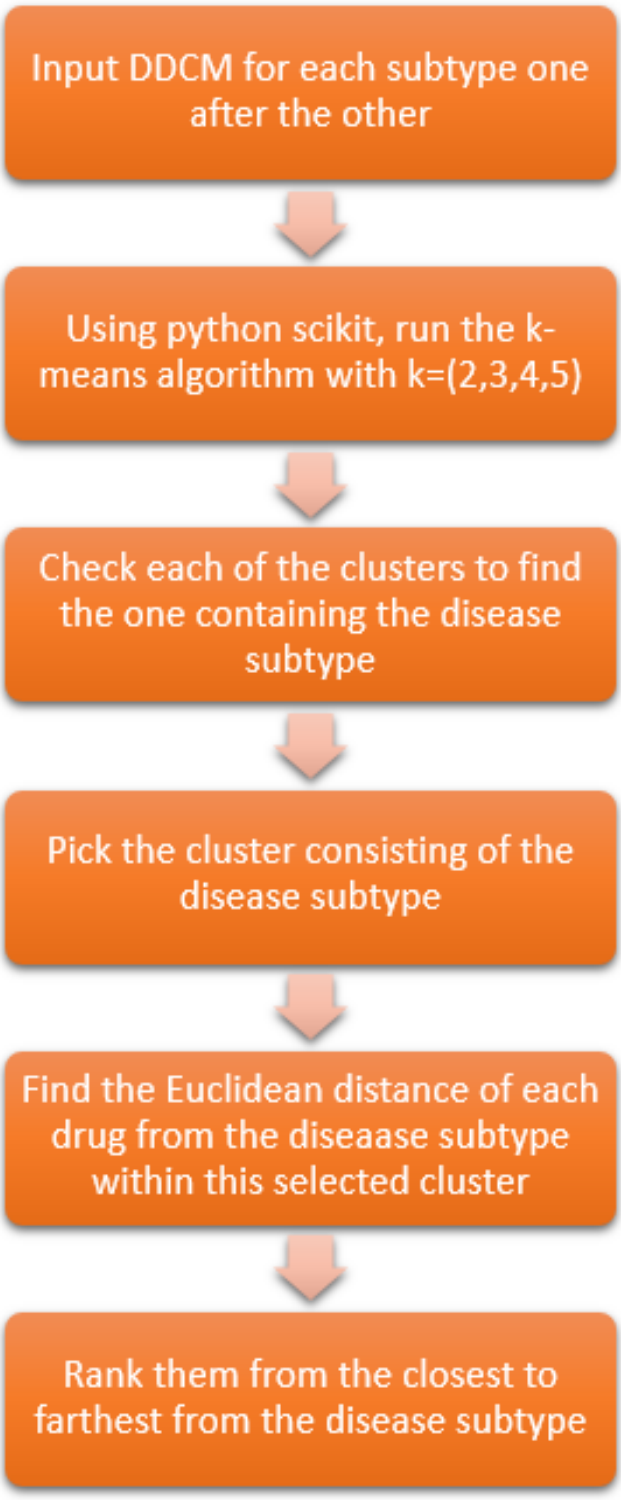

Figure 3.12: K-Means Process Workflow

We now take the DDCM obtained towards the end of our preprocessing as the

input for this algorithm. Once the algorithm generates the clusters based on the four **k** values used, we select the cluster containing the disease subtype. Then all drug data points within this cluster are chosen to compute the Euclidean distance between a drug and the subtype. We then rank all of these drugs from the closest to farthest from the disease subtype. All drugs grouped together in the same cluster as the disease subtype indicate that they share similar properties. Thus, only those drugs within the same cluster as the disease subtype were chosen for computing the Euclidean distance. The drugs that are found to be closest to the subtype are considered better drug repurposing candidates when compared to drugs that are found farther away.

Algorithm 2: K-Means Clustering

Input: DDCM for subtype $S_n(n = 1\ to\ 10)$ *with* $k = (2, 3, 4, 5)$ $n\_init = 10$

Output: A set of k clusters

Method: *Randomly pick k objects from DDCM as cluster centers*

Repeat:

1. *Assign every object to the cluster with which it shares most similarities on the basis of its mean value of objects within the cluster*
2. *Calculate new updated mean for each cluster*

Until:

Convergence is met

The minimization formula (3.2) shows that we aim to partition the observations into **k** clusters while making sure the total within-cluster variation (WCV) summed up over all clusters is as low as possible. [36]

$$Minimize\ C_1\ to\ C_k \left\{ \sum_{k=1}^{K} WCV(C_k) \right\} \tag{3.2}$$

where,

C is a cluster

k is the cluster number

WCV is within-cluster variance

In order to define within-cluster variance, we use the following formula.

$$WCV = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (X_{ij} - X_{i'j})^2 \tag{3.3}$$

Combining equation 3.2 and 3.3 gives us the optimization problem defining k-means clustering.

$$Minimize\ C_1\ to\ C_k \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (X_{ij} - X_{i'j})^2 \right\} \tag{3.4}$$

3.3.3 Agglomerative clustering

This is a bottom up hierarchical clustering algorithm where we start by assuming all data points given as input to be clusters. This algorithm merges clusters based on their proximity to each other. The clusters which are closest to each other would be the ones that will have been merged. Then we repeat this process until all data points are merged into one single cluster.

Before the clustering is performed, determining the proximity of each clusters is usually computed. There are a few methods to use such as 'single linkage', 'complete linkage', 'average linkage', 'ward linkage' and more.

In single linkage, the distance between each cluster is noted as the closest distance between two points, one from each cluster. In complete linkage, the distance between each cluster is noted as the farthest distance between two points, one from each

cluster. In average linkage, it is defined as the average distance between each point in one cluster to every other point to the other cluster. In ward linkage, it minimizes the variance of the clusters being merged by finding the error function which is the average RMS distance of each data point in a cluster.

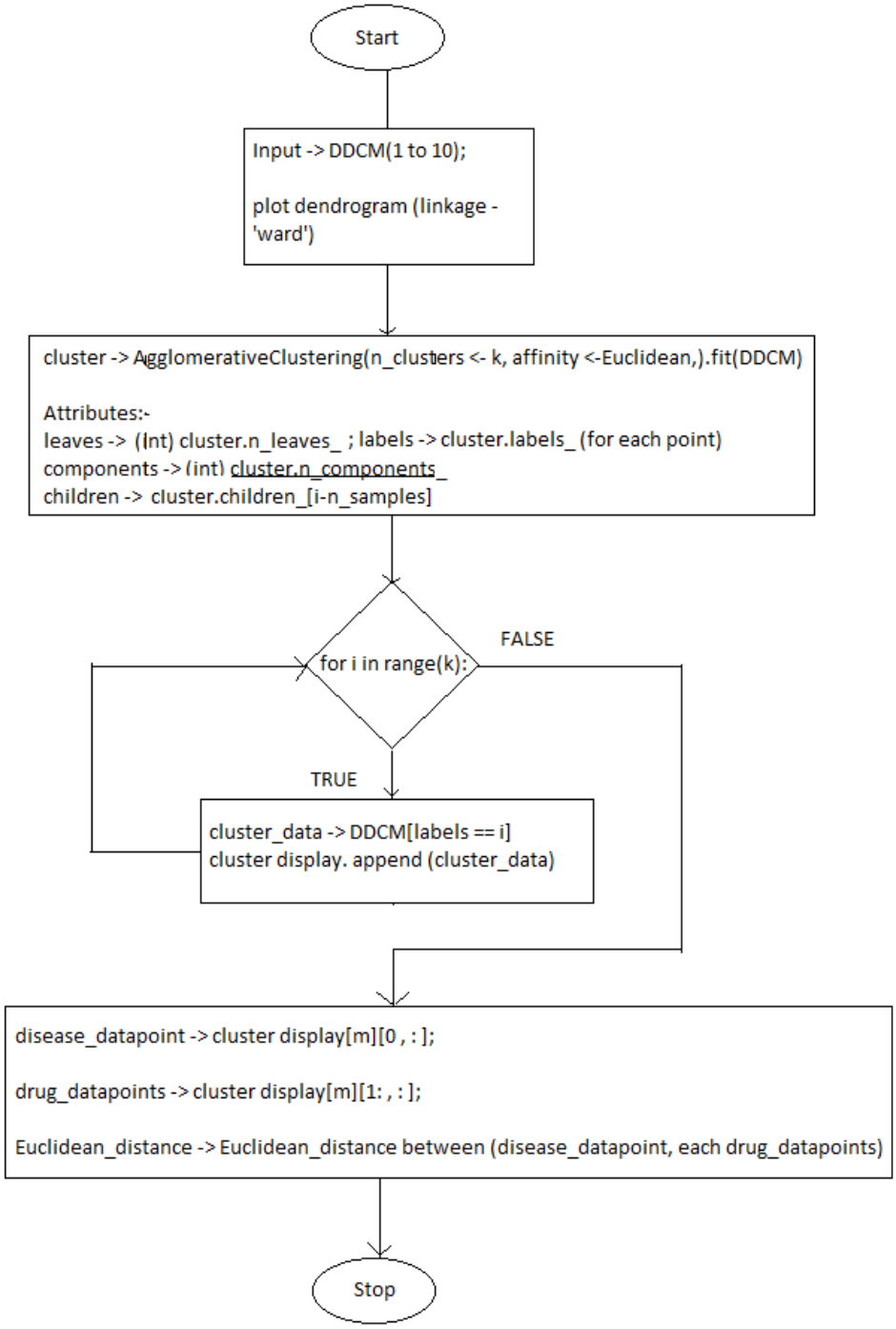

Figure 3.13: Proposed Agglomerative Working

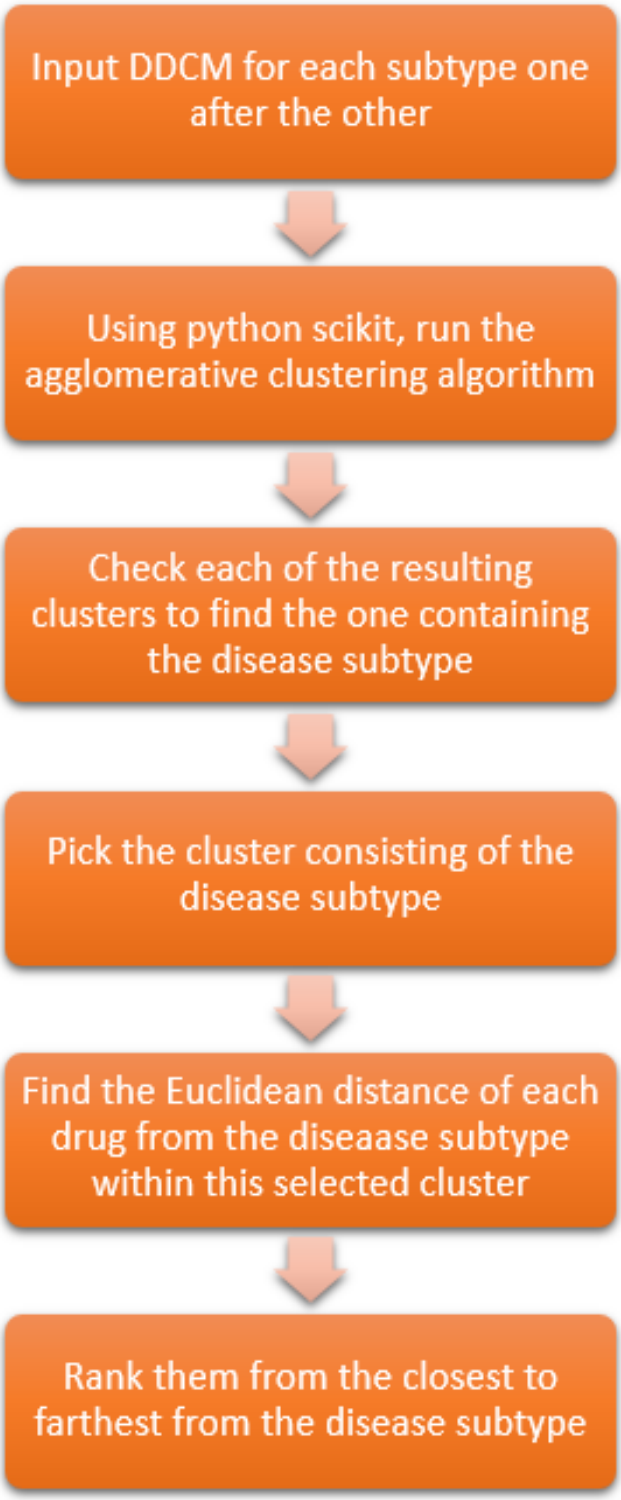

Figure 3.14: Agglomerative Process Workflow

The DDCM obtained towards the end of our preprocessing is taken as the input for this algorithm. After generating the output clusters, we then choose the cluster that contains the disease subtype. Then each drug data point within this cluster are picked to calculate the Euclidean distance between a drug and the disease subtype. Next step is to rank all of these drugs from the closest to farthest from the disease subtype. Each and every drug grouped together in the same cluster as the disease subtype indicate that they have similar properties. Hence, only the drugs within the same cluster as the disease subtype were chosen for calculating the Euclidean distance. Those drugs that are found to be closest to the subtype are considered better drug repurposing candidates when compared to drugs that are found farther away.

Algorithm 3: Agglomerative Clustering

Input: DDCM for subtype $S_n(n = 1\ to\ 10)$

Output: A set of k clusters

Method: *Start by assuming all data points to be clusters*

Repeat:

Merge the two closest clusters

Until:

All data points are in a single cluster

3.3.4 Euclidean distance

Euclidean distance is defined as the ordinary straight line distance between two points 'X' and 'Y'. Here, we compute pairwise Euclidean distance matrix using equation 3.6.

$$(x1 - y1)2 + (x2 - y2)2 + \ + (xn - yn)2 \tag{3.5}$$

For efficiency, the Euclidean distance between a row vector X and Y is computed as follows.

$$dist(x, y) = sqrt(dot(x, x) - 2 * dot(x, y) + dot(y, y)) \tag{3.6}$$

This formula is advantageous as it is computationally more efficient when sparse data is considered.

Chapter 4

Results

In this chapter, we shall go through the results of clustering algorithms per for each subtype and compare the results obtained using both k-means clustering and agglomerative clustering. The results showcase several unapproved drugs alongside approved drugs closest to the subtype indicating that the unapproved drugs share similarities with the approved drugs which means that they are worth pursuing for repurposing. All the following tables showing top ten drugs per subtype show us that since these drugs appear to be the closest to their particular disease subtype, the generated lists contain best suited drug repurposing candidates.

4.1 SubType 1

In this section, we observe the results obtained for disease subtype 1. Here, we notice that both algorithms have produced a decent set of results showing the top drugs best suited for potential drug repurposing candidates for this subtype.

4.1.1 K-Means clustering

Using this algorithm, we obtained the same drugs as in the top ten closest to the disease subtype for each of the four k values. We observe no difference in this result list for each of the four k values as shown in table 4.1. The list comprises of three approved drugs, three experimental drugs, three unknown drugs and one investigational drug. The optimal k value was found to be 5.

Table 4.1: SubType 1 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4, k=5)

Rank	Drug Name	FDA Status
1	Trimethobenzamide	Approved
2	KIN001-266	Unknown
3	Agomelatine	Approved
4	Clinofibrate	Experimental
5	Dopamine	Approved
6	AS-601245	Experimental
7	YM-976	Experimental
8	ZK-200775	Investigational
9	JTE-907	Unknown
10	Dacinostat	Unknown

4.1.2 Agglomerative clustering

When we obtained the results for this subtype using agglomerative clustering, we observe that the 7 out of the top ten drugs are the same as what we observed in the k-means result for this subtype. Although only the top four drugs on these lists retain the same position in both sets of results as shown in table 4.2. The list is comprised of four approved drugs, one unknown drug and five experimental drugs. The optimal k value was found to be 2.

Table 4.2: SubType 1 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	Trimethobenzamide	Approved
2	KIN001-266	Unknown
3	Agomelatine	Approved
4	Clinofibrate	Experimental
5	CHIR-99021	Experimental
6	Tozasertib	Experimental
7	Dopamine	Approved
8	AS-601245	Experimental
9	Amfepramone	Approved
10	YM-976	Experimental

4.2 SubType 2

In this section, we look at the results obtained for disease subtype 2. We notice that both algorithms have produced a decent set of results showing the top drugs best suited for potential drug repurposing candidates for this subtype.

4.2.1 K-Means clustering

The results obtained using this algorithm show that we have some differences in the top ten drug list in each of the four k values. For k=2, we see that there are three approved drugs, three unknown drugs, two experimental drugs and two investigational drugs as shown in table 4.3. The optimal k value was found to be 5.

For k=3, we observe that the top ten best suited drugs include three approved drugs, three unknown drugs, two experimental drugs and two investigational drugs as shown in table 4.4.

For k=4, we observe that the top ten best suited drugs include three approved drugs, two unknown drugs, two experimental drugs and three investigational drugs

Table 4.3: SubType 2 ranked list of drugs: K-Means Clustering (k=2)

Rank	Drug Name	FDA Status
1	CEP-37440	Investigational
2	tioconazole	Approved
3	KIN001-266	Unknown
4	agomelatine	Approved
5	HC-030031	Unknown
6	clinofibrate	Experimental
7	climbazole	Unknown
8	tozasertib	Experimental
9	BMS-777607	Investigational
10	quetiapine	Approved

Table 4.4: SubType 2 ranked list of drugs: K-Means Clustering (k=3)

Rank	Drug Name	FDA Status
1	CEP-37440	Investigational
2	tioconazole	Approved
3	KIN001-266	Unknown
4	agomelatine	Approved
5	HC-030031	Unknown
6	clinofibrate	Experimental
7	climbazole	Unknown
8	tozasertib	Experimental
9	quetiapine	Approved
10	indibulin	Investigational

as shown in table 4.5.

For k=5, we observe that the top ten best suited drugs include four approved drugs, two unknown drugs, two experimental drugs and two investigational drugs as shown in table 4.6.

Table 4.5: SubType 2 ranked list of drugs: K-Means Clustering (k=4)

Rank	Drug Name	FDA Status
1	CEP-37440	Investigational
2	tioconazole	Approved
3	agomelatine	Approved
4	HC-030031	Unknown
5	clinofibrate	Experimental
6	climbazole	Unknown
7	tozasertib	Experimental
8	BMS-777607	Investigational
9	quetiapine	Approved
10	indibulin	Investigational

Table 4.6: SubType 2 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	CEP-37440	Investigational
2	tioconazole	Approved
3	KIN001-266	Unknown
4	agomelatine	Approved
5	HC-030031	Unknown
6	clinofibrate	Experimental
7	tozasertib	Experimental
8	quetiapine	Approved
9	olaparib	Approved
10	gatifloxacin	Approved

4.2.2 Agglomerative clustering

Using this method, we observe that the results obtained show three approved, three unknown, two experimental and two investigational drugs in the top ten best suited drugs for repurposing in this subtype as shown in table 4.7. We observe that the list obtained with this algorithm is the same set of results as that of drug list generated by k=2. The optimal k value was found to be 2.

Table 4.7: SubType 2 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	CEP-37440	Investigational
2	tioconazole	Approved
3	KIN001-266	Unknown
4	agomelatine	Approved
5	HC-030031	Unknown
6	clinofibrate	Experimental
7	climbazole	Unknown
8	tozasertib	Experimental
9	BMS-777607	Investigational
10	quetiapine	Approved

4.3 SubType 3

Here in subtype 3, we observe good results obtained by both algorithms.

4.3.1 K-Means clustering

We can infer from the following tables that k=(2,3,5) have produced the same list of top ten drugs. From table 4.8, we can observe that the list comprises of six approved drugs, two experimental and two unknown drugs. The optimal k value was found to be 5.

For k=4, we infer from table 4.9 that there are seven approved drugs along with two unknown and one experimental drug. There are three different drugs from this list and table 4.8.

4.3.2 Agglomerative clustering

Using this method, we infer from table 4.10 that these results match the results obtained using k-means (k=2,3,5) in this subtype indicating the top ten drugs to

Table 4.8: SubType 3 ranked list of drugs: K-Means Clustering (k=2, k=3, k=5)

Rank	Drug Name	FDA Status
1	climbazole	Unknown
2	CHIR-99021	Experimental
3	Agomelatine	Approved
4	folic-acid	Approved
5	amfepramone	Approved
6	YM-976	Experimental
7	desvenlafaxine	Approved
8	felbamate	Approved
9	KIN001-266	Unknown
10	fluspirilene	Approved

Table 4.9: SubType 3 ranked list of drugs: K-Means Clustering (k=4)

Rank	Drug Name	FDA Status
1	climbazole	Unknown
2	CHIR-99021	Experimental
3	Agomelatine	Approved
4	folic-acid	Approved
5	desvenlafaxine	Approved
6	KIN001-266	Unknown
7	fluspirilene	Approved
8	sacubitril	Approved
9	fluticasone-propionate	Approved
10	trimethobenzamide	Approved

comprise of six approved, two unknown and two experimental drugs. The optimal k value was found to be 3.

4.4 SubType 4

In subtype 4, we observe that the results obtained by both algorithms fare well. The top ten drug lists contain majority approved drugs and the unapproved drugs that share similarities with them.

Table 4.10: SubType 3 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	climbazole	Unknown
2	CHIR-99021	Experimental
3	Agomelatine	Approved
4	folic-acid	Approved
5	amfepramone	Approved
6	YM-976	Experimental
7	desvenlafaxine	Approved
8	felbamate	Approved
9	KIN001-266	Unknown
10	fluspirilene	Approved

4.4.1 K-Means clustering

In this algorithm, we notice from table 4.11 that one experimental, two unknown and seven approved drugs appear to be the best suited drugs. The optimal k value was found to be 4

Table 4.11: SubType 4 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4)

Rank	Drug Name	FDA Status
1	Trimethobenzamide	Approved
2	Dopamine	Approved
3	KIN001-266	Unknown
4	Clinofibrate	Experimental
5	Agomelatine	Approved
6	dacinostat	Unknown
7	folic-acid	Approved
8	aurora-a-inhibitor-I	Approved
9	sacubitril	Approved
10	gatifloxacin	Approved

From table 4.12, we observe that using k-means (k=5), we obtain six approved, one investigational, one experimental and two unknown drugs are present in the top ten drugs.

Table 4.12: SubType 4 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	Trimethobenzamide	Approved
2	Dopamine	Approved
3	KIN001-266	Unknown
4	Clinofibrate	Experimental
5	dacinostat	Unknown
6	aurora-a-inhibitor-I	Approved
7	sacubitril	Approved
8	gatifloxacin	Approved
9	piribedil	Investigational
10	fluticasone-propionate	Approved

4.4.2 Agglomerative clustering

From table 4.13 we infer that using this method the top ten drugs appear to consist of seven approved drugs, two unknown and one experimental drug. These results appear to be the same as the ones obtained by k-means (k=2,3,4) in this subtype. The optimal k value was found to be 2.

Table 4.13: SubType 4 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	Trimethobenzamide	Approved
2	Dopamine	Approved
3	KIN001-266	Unknown
4	Clinofibrate	Experimental
5	Agomelatine	Approved
6	dacinostat	Unknown
7	folic-acid	Approved
8	aurora-a-inhibitor-I	Approved
9	sacubitril	Approved
10	gatifloxacin	Approved

4.5 SubType 5

In subtype 5, we observe the results obtained by using two algorithms on our dataset. We notice that both algorithms generated a good set of results.

4.5.1 K-Means clustering

This algorithm shows for k values 2,3, and 4, seven approved, one experimental and two unknown drugs as the top ten best suited drug candidates for repurposing for this subtype as shown in table 4.14. The optimal k value was found to be 3.

Table 4.14: SubType 5 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4)

Rank	Drug Name	FDA Status
1	lisinopril	Approved
2	felbamate	Approved
3	aurora-a-inhibitor-I	Approved
4	agomelatine	Approved
5	clinofibrate	Experimental
6	dacinostat	Unknown
7	cinaciguat	Unknown
8	trifluoperazine	Approved
9	desvenlafaxine	Approved
10	gatifloxacin	Approved

This algorithm shows for k values 5, six approved, one experimental, one investigational, and two unknown drugs as the top ten best suited drug candidates for repurposing for this subtype as shown in table 4.15.

4.5.2 Agglomerative clustering

This method produced the same results as that of (k=2,3,4) in this subtype as shown in table 4.16. It contains seven approved, one experimental and two unknown drugs

Table 4.15: SubType 5 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	lisinopril	Approved
2	aurora-a-inhibitor-I	Approved
3	agomelatine	Approved
4	clinofibrate	Experimental
5	dacinostat	Unknown
6	desvenlafaxine	Approved
7	gatifloxacin	Approved
8	piribedil	Investigational
9	benazepril	Approved
10	JTE-907	Unknown

ranked as the best suited drug candidates for repurposing in this subtype. The optimal k value was found to be 3.

Table 4.16: SubType 5 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	lisinopril	Approved
2	felbamate	Approved
3	aurora-a-inhibitor-I	Approved
4	agomelatine	Approved
5	clinofibrate	Experimental
6	dacinostat	Unknown
7	cinaciguat	Unknown
8	trifluoperazine	Approved
9	desvenlafaxine	Approved
10	gatifloxacin	Approved

4.6 SubType 6

In this section, we look at the results obtained for disease subtype 6. We observe that both algorithms have produced a good set of results showing the top drugs best

suited for potential drug repurposing candidates for this subtype.

4.6.1 K-Means clustering

Using this algorithm we can infer from table 4.17 that the ranked drug list obtained using k-means (k=2,3,4) comprises of two unknown, one experimental and seven approved drugs. The optimal k value was found to be 4.

Table 4.17: SubType 6 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4)

Rank	Drug Name	FDA Status
1	agomelatine	Approved
2	KIN001-266	Unknown
3	CHIR-99021	Experimental
4	JTE-907	Unknown
5	fluticasone-propionate	Approved
6	olaparib	Approved
7	dopamine	Approved
8	lisinopril	Approved
9	trimethobenzamide	Approved
10	tioconazole	Approved

We can infer from table 4.18 that for k=5, we obtain a slightly different set of ranked drugs in which the number of approved, experimental and unknown drugs remain the same as the previous list in this subtype but with one different approved drug made the list in place of another.

4.6.2 Agglomerative clustering

In this method we can observe that the ranked list of drugs are the same as table 4.17. From table 4.19 we infer that there are two unknown, one experimental and seven approved drugs ranked in the top ten. The optimal k value was found to be 4.

Table 4.18: SubType 6 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	agomelatine	Approved
2	KIN001-266	Unknown
3	CHIR-99021	Experimental
4	JTE-907	Unknown
5	olaparib	Approved
6	dopamine	Approved
7	lisinopril	Approved
8	trimethobenzamide	Approved
9	tioconazole	Approved
10	desvenlafaxine	Approved

Table 4.19: SubType 6 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	agomelatine	Approved
2	KIN001-266	Unknown
3	CHIR-99021	Experimental
4	JTE-907	Unknown
5	fluticasone-propionate	Approved
6	olaparib	Approved
7	dopamine	Approved
8	lisinopril	Approved
9	trimethobenzamide	Approved
10	tioconazole	Approved

4.7 SubType 7

In this section, we observe the results obtained by using both clustering methods. The tables below indicate the best suited drug candidates for repurposing for this subtype.

4.7.1 K-Means clustering

Using this method, for the three k values (2,3,4), we notice that two investigational, one unknown and seven approved drugs made the list as shown in table 4.20. The

optimal k value was found to be 4.

Table 4.20: SubType 7 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4)

Rank	Drug Name	FDA Status
1	tioconazole	Approved
2	gatifloxacin	Approved
3	trimethobenzamide	Approved
4	agomelatine	Approved
5	cinaciguat	Unknown
6	ambroxol	Investigational
7	fluticasone-propionate	Approved
8	calcitriol	Approved
9	felbamate	Approved
10	ebselen	Investigational

From table 4.21 we can infer that when the k value was set to 5, we obtained an almost same list of drugs except that one experimental drug made the list in place of an approved drug.

Table 4.21: SubType 7 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	tioconazole	Approved
2	gatifloxacin	Approved
3	trimethobenzamide	Approved
4	agomelatine	Approved
5	cinaciguat	Unknown
6	ambroxol	Investigational
7	fluticasone-propionate	Approved
8	calcitriol	Approved
9	ebselen	Investigational
10	AS-601245	Experimental

4.7.2 Agglomerative clustering

Using this method, we interpret that the ranked list of drugs obtained are the same as they appear in table 4.20. We notice this in the list provided in table 4.22. The optimal k value was found to be 3.

Table 4.22: SubType 7 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	tioconazole	Approved
2	gatifloxacin	Approved
3	trimethobenzamide	Approved
4	agomelatine	Approved
5	cinaciguat	Unknown
6	ambroxol	Investigational
7	fluticasone-propionate	Approved
8	calcitriol	Approved
9	felbamate	Approved
10	ebselen	Investigational

4.8 SubType 8

In this section, we look at the results obtained for disease subtype 8. We notice that both algorithms have produced a decent set of results showing the top drugs best suited for potential drug repurposing candidates for this subtype.

4.8.1 K-Means clustering

Using this method, the ranked list of drugs for k values 2,3, and 4, appear to contain one experimental drug, four unknown drugs and five approved drugs as shown in table 4.23. The optimal k value was found to be 5.

Table 4.23: SubType 8 ranked list of drugs: K-Means Clustering (k=2, k=3, k=4)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	agomelatine	Approved
3	dopamine	Approved
4	trimethobenzamide	Approved
5	tioconazole	Approved
6	climbazole	Unknown
7	gatifloxacin	Approved
8	dacinostat	Unknown
9	YM-976	Experimental
10	JTE-907	Unknown

For k value of 5, we observe that it contains one experimental drug, four unknown drugs and five approved drugs as shown in table 4.24. Although one of the approved drug from the previous table has been replaced by another approved drug as shown in this table.

Table 4.24: SubType 8 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	agomelatine	Approved
3	dopamine	Approved
4	trimethobenzamide	Approved
5	climbazole	Unknown
6	gatifloxacin	Approved
7	dacinostat	Unknown
8	YM-976	Experimental
9	JTE-907	Unknown
10	desvenlafaxine	Approved

4.8.2 Agglomerative clustering

In this method, we notice that the ranked list of drugs is the same as the one generated by k-means with k values of 2,3, and 4 as shown in table 4.25. The optimal k value was found to be 2.

Table 4.25: SubType 8 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	agomelatine	Approved
3	dopamine	Approved
4	trimethobenzamide	Approved
5	tioconazole	Approved
6	climbazole	Unknown
7	gatifloxacin	Approved
8	dacinostat	Unknown
9	YM-976	Experimental
10	JTE-907	Unknown

4.9 SubType 9

Subtype 9's results show differences in the drugs list for both methods. The top two drugs remain the same in every instance in this subtype.

4.9.1 K-Means clustering

In this method, for a k value of 2, we obtain a drug list comprising of five approved drugs, three unknown drugs, one investigational and one experimental drug as we can see in table 4.26. The optimal k value was found to be 5.

From table 4.27, we can infer that six approved, one investigational and three

Table 4.26: SubType 9 ranked list of drugs: K-Means Clustering (k=2)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	JTE-907	Unknown
3	CHIR-99021	Experimental
4	dopamine	Approved
5	gatifloxacin	Approved
6	benazepril	Approved
7	folic-acid	Approved
8	agomelatine	Approved
9	piribedil	Investigational
10	climbazole	Unknown

unknown drugs appear to be the closest to this subtype.

Table 4.27: SubType 9 ranked list of drugs: K-Means Clustering (k=3)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	JTE-907	Unknown
3	dopamine	Approved
4	gatifloxacin	Approved
5	agomelatine	Approved
6	piribedil	Investigational
7	aurora-a-inhibitor-I	Approved
8	donitriptan	Unknown
9	trimethobenzamide	Approved
10	desvenlafaxine	Approved

In table 4.28 where the k value is 4, we can see that there are five approved, three unknown, and two investigational drugs which are closest to this subtype.

From table 4.29, we can notice that taking a k value of 5 placed one experimental, one investigational, three unknown and five approved drugs.

Table 4.28: SubType 9 ranked list of drugs: K-Means Clustering (k=4)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	JTE-907	Unknown
3	dopamine	Approved
4	gatifloxacin	Approved
5	agomelatine	Approved
6	piribedil	Investigational
7	aurora-a-inhibitor-I	Approved
8	donitriptan	Unknown
9	desvenlafaxine	Approved
10	tacedinaline	Investigational

Table 4.29: SubType 9 ranked list of drugs: K-Means Clustering (k=5)

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	JTE-907	Unknown
3	CHIR-99021	Experimental
4	dopamine	Approved
5	gatifloxacin	Approved
6	agomelatine	Approved
7	piribedil	Investigational
8	aurora-a-inhibitor-I	Approved
9	donitriptan	Unknown
10	trimethobenzamide	Approved

4.9.2 Agglomerative clustering

Using this method, we observe that the results produced are the same as the ones generated by k-means with a k value of 2. As we notice in table 4.30, we obtain five approved, one investigational, one experimental and three unknown drugs closest to this subtype. The optimal k value was found to be 2.

Table 4.30: SubType 9 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	KIN001-266	Unknown
2	JTE-907	Unknown
3	CHIR-99021	Experimental
4	dopamine	Approved
5	gatifloxacin	Approved
6	benazepril	Approved
7	folic-acid	Approved
8	agomelatine	Approved
9	piribedil	Investigational
10	climbazole	Unknown

4.10 SubType 10

In this section we look into the results obtained for subtype 10 using both our methods. The three tables show us the best suited drug candidates for repurposing for this subtype.

4.10.1 K-Means clustering

For k values of 2 and 3, we obtain six approved, one experimental, one investigational and two unknown drugs to be considered as good drug repurposing candidates as shown in table 4.31. The optimal k value was found to be 5.

For k values of 4 and 5, we obtain five approved, two experimental, one investigational and two unknown drugs to be considered as good drug repurposing candidates as shown in table 4.32. We observe that except fofr one experimental drug in this table, the first nine drugs remian the same as the list in table 4.31.

Table 4.31: SubType 10 ranked list of drugs: K-Means Clustering (k=2, k=3)

Rank	Drug Name	FDA Status
1	dopamine	Approved
2	KIN001-266	Unknown
3	felbamate	Approved
4	donitriptan	Unknown
5	desvenlafaxine	Approved
6	CHIR-99021	Experimental
7	CEP-37440	Investigational
8	quetiapine	Approved
9	olmesartan-medoxomil	Approved
10	trimethobenzamide	Approved

Table 4.32: SubType 10 ranked list of drugs: K-Means Clustering (k=4, k=5)

Rank	Drug Name	FDA Status
1	dopamine	Approved
2	KIN001-266	Unknown
3	felbamate	Approved
4	donitriptan	Unknown
5	desvenlafaxine	Approved
6	CHIR-99021	Experimental
7	CEP-37440	Investigational
8	quetiapine	Approved
9	olmesartan-medoxomil	Approved
10	clinofibrate	Experimental

4.10.2 Agglomerative clustering

In this method, we notice that with the exception of one drug, every other drug displayed in table 4.33 also appears to be in table 4.31. The optimal k value was found to be 3.

Overall, we have displayed the best suited potential drug repurposing candidates for each of the ten subtypes using the aforementioned methods.

Table 4.33: SubType 10 ranked list of drugs: Agglomerative Clustering

Rank	Drug Name	FDA Status
1	dopamine	Approved
2	KIN001-266	Unknown
3	donitriptan	Unknown
4	desvenlafaxine	Approved
5	CHIR-99021	Experimental
6	CEP-37440	Investigational
7	quetiapine	Approved
8	olmesartan-medoxomil	Approved
9	trimethobenzamide	Approved
10	clinofibrate	Experimental

Chapter 5

Conclusion and Future Work

In this thesis, we aimed to find suitable drug repurposing candidates for each of the ten breast cancer subtypes. We were given METABRIC and LINCS datasets. We performed a series of preprocessing steps on both of these datasets to create a DDCM (Drug Disease Combination Matrix) for each of the ten disease subtypes. Then using anti-correlation matrix which we generated, we have obtained a filtered DDCM per subtype. Then we applied k-mean clustering along with agglomerative clustering method. Then we computed the euclidean distances of all the drugs found in the same cluster as the disease subtype, and ranked them closest to farthest from the subtype. We then picked the top ten drugs for each subtype using the aforementioned methods.

5.1 Possible Future Work

Future work that can be conducted includes the following:-

- Using the DDCM we generated, dimensionality reduction and distance metric learning methods can be implemented
- Using side-effect similarity of unapproved drugs with that of approved drugs,

drug repurposing candidates can be obtained

- Our preprocessing steps and methods can be applied on a different cancer dataset such as prostate cancer

These ideas can be an open problem that can be explored in the future.

Bibliography

[1] Breast cancer statistics - canadian cancer society, http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on, 2019. Last accessed March, 2019.

[2] Gene expression data, https://compbio.soe.ucsc.edu/genex/expressdata.html, 2019. Last accessed March, 2019.

[3] James C Bezdek and Nikhil R Pal. Cluster validation with generalized dunn's indices. In *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pages 190–193. IEEE, 1995.

[4] Brenda. Brenda, http://www.brenda-enzymes.info/index.php4, 2019. Last accessed March, 2019.

[5] CancerBrowser. Cancerbrowser, http://www.cancerbrowser.org/drugs, 2019. Last accessed March, 2019.

[6] cBioPortal. Introduction - cbioportal 1.2.2 documentation, https://cbioportal.readthedocs.io/en/latest/Z-Score-normalization-script.html, 2019. Last accessed March, 2019.

[7] Jinyan Chan, Xuan Wang, Jacob A Turner, Nicole E Baldwin, and Jinghua Gu. Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. *Bioinformatics*, 2019.

[8] Chembl. Chembl, `https://www.ebi.ac.uk/chembl/drugstore`, 2019. Last accessed March, 2019.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[10] Sohrab P. Shah Suet-Feung Chin Gulisa Turashvili Oscar M. Rueda Mark J. Dunning Doug Speed et al. Curtis, Christina. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature International Journal of Science*, 486:346–352, 2012.

[11] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 2014.

[12] Drugbank. 3d structure for acetylsalicylic acid (db00945), `https://www.drugbank.ca/structures/small_molecule_drugs/DB00945`, 2019. Last accessed March, 2019.

[13] Drugbank. Drugbank, `https://www.drugbank.ca/`, 2019. Last accessed March, 2019.

[14] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[15] Nehme El-Hachem, Deena MA Gendoo, Laleh Soltan Ghoraie, Zhaleh Safikhani, Petr Smirnov, Christina Chung, Kenan Deng, Ailsa Fang, Erin Birkwood, Chan-

tal Ho, et al. Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer research*, 77(11):3057–3069, 2017.

[16] Janine T Erler and Rune Linding. Network-based drugs and biomarkers. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(2):290–296, 2010.

[17] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[18] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[19] Kegg. Kegg, `https://www.kegg.jp/`, 2019. Last accessed March, 2019.

[20] Haeseung Lee, Seungmin Kang, and Wankyu Kim. Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. *PloS one*, 11(3):e0150460, 2016.

[21] Metador. Metador, `http://matador.embl.de/`, 2019. Last accessed March, 2019.

[22] National Institutes of Health. What is a gene? - genetics home reference - nih, `https://ghr.nlm.nih.gov/primer/basics/gene`, 2019. Last accessed March, 2019.

[23] Tudor I Oprea, Julie E Bauman, Cristian G Bologa, Tione Buranda, Alexandre Chigaev, Bruce S Edwards, Jonathan W Jarvik, Hattie D Gresham, Mark K Haynes, Brian Hjelle, et al. Drug repurposing from an academic perspective. *Drug Discovery Today: Therapeutic Strategies*, 8(3-4):61–69, 2011.

[24] Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, 34(16):2817–2825, 2018.

[25] Richard Segraves Damir Sudar-Steven Clark Ian Poole David Kowbel Colin Collins et al. Pinkel, Daniel. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[26] PubChem. Pubchem, `https://pubchem.ncbi.nlm.nih.gov/`, 2019. Last accessed March, 2019.

[27] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[28] Nafiseh Saberian, Azam Peyvandipour, Michele Donato, Sahar Ansari, and Sorin Draghici. A new computational drug repurposing method using established disease-drug pair knowledge. *Bioinformatics*, 2019.

[29] American Cancer Society. What is breast cancer? — breast cancer definition, `https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html`, 2019. Last accessed March, 2019.

[30] Supertarget. Supertarget, `http://bioinf-apache.charite.de/supertargetv2/`, 2019. Last accessed March, 2019.

[31] Wikipedia contributors. Drug — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Drug&oldid=894103002`, 2019. [Online; accessed 1-April-2019].

[32] Wikipedia contributors. Drug repositioning — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Drug_repositioning&oldid=893599472, 2019. [Online; accessed 1-May-2019].

[33] Wikipedia contributors. Machine learning — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=889950774, 2019. [Online; accessed 29-March-2019].

[34] Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59, 2012.

[35] Xiaodan Wu, Luonan Chen, and Xiangdong Wang. Network biomarkers, interaction networks and dynamical network biomarkers in respiratory diseases. *Clinical and translational medicine*, 3(1):16, 2014.

[36] Pei-Yuan Zhou and Keith CC Chan. A model-based multivariate time series clustering algorithm. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 805–817. Springer, 2014.

Appendix A: Clustering Results

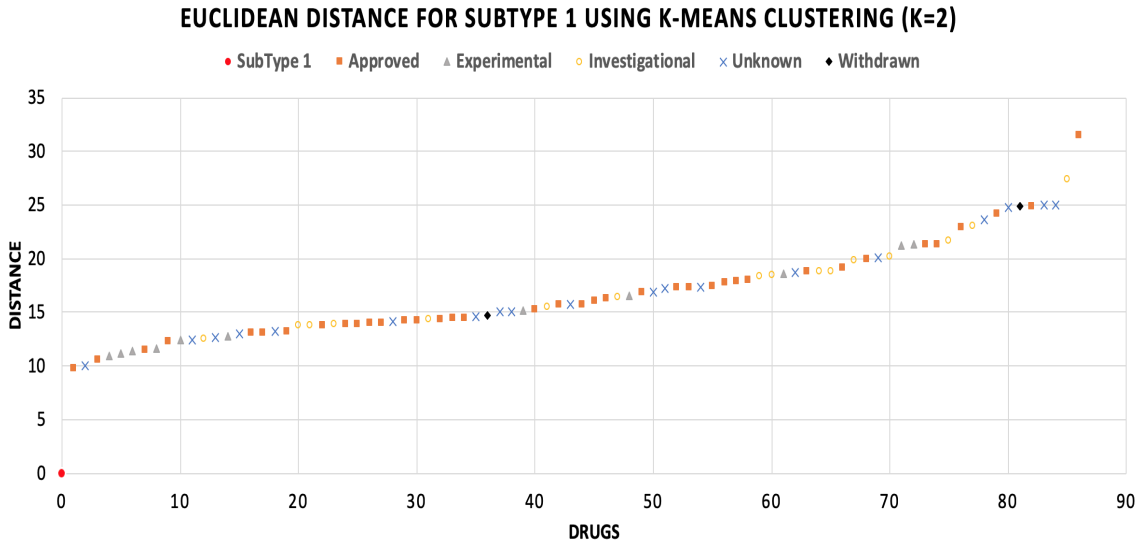

Figure 1: SubType 1 ranked list of drugs: K-Means (K=2)

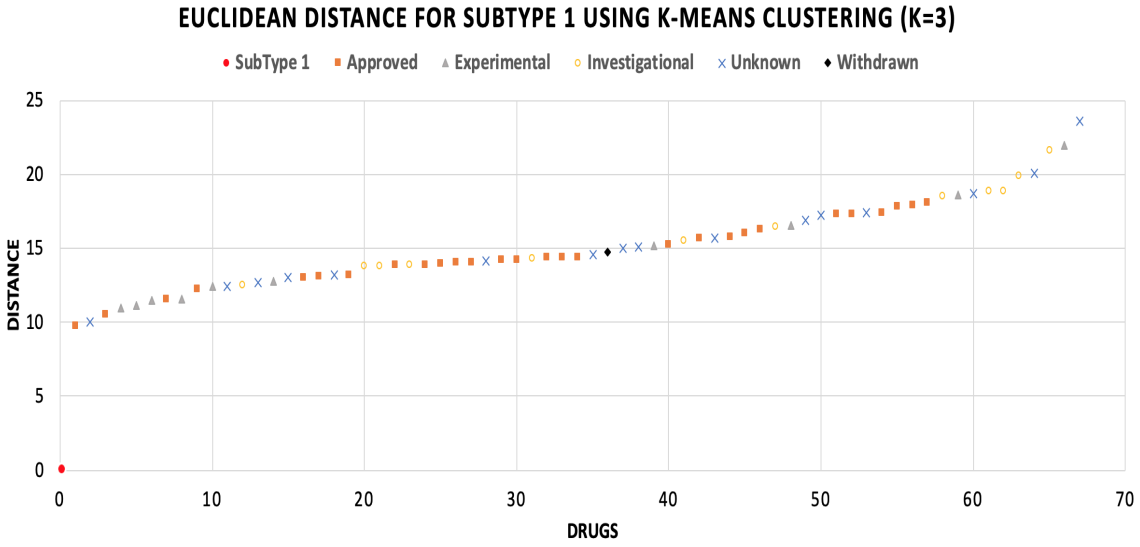

Figure 2: SubType 1 ranked list of drugs: K-Means (K=3)

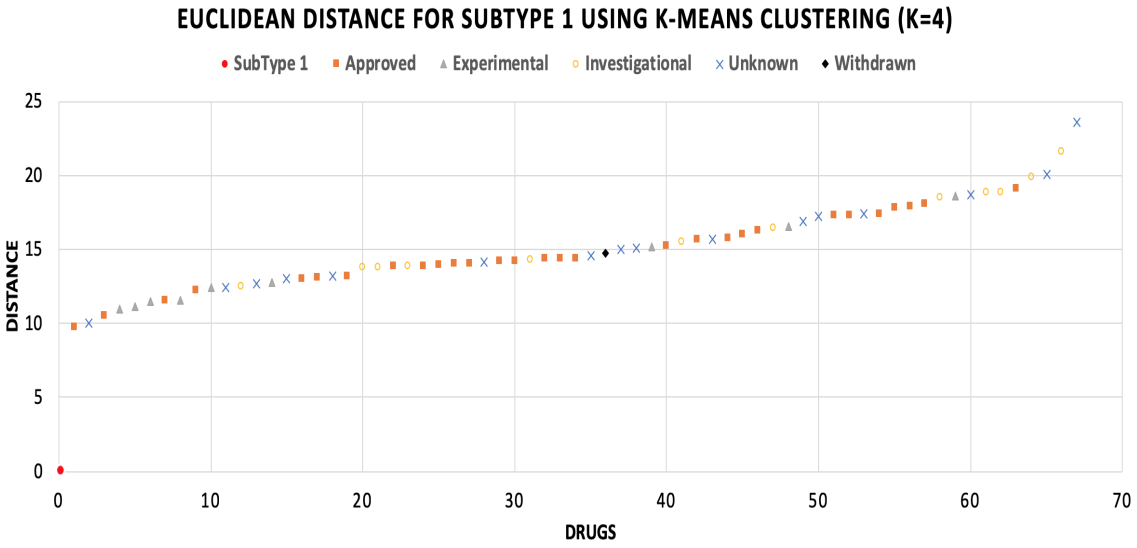

Figure 3: SubType 1 ranked list of drugs: K-Means (K=4)

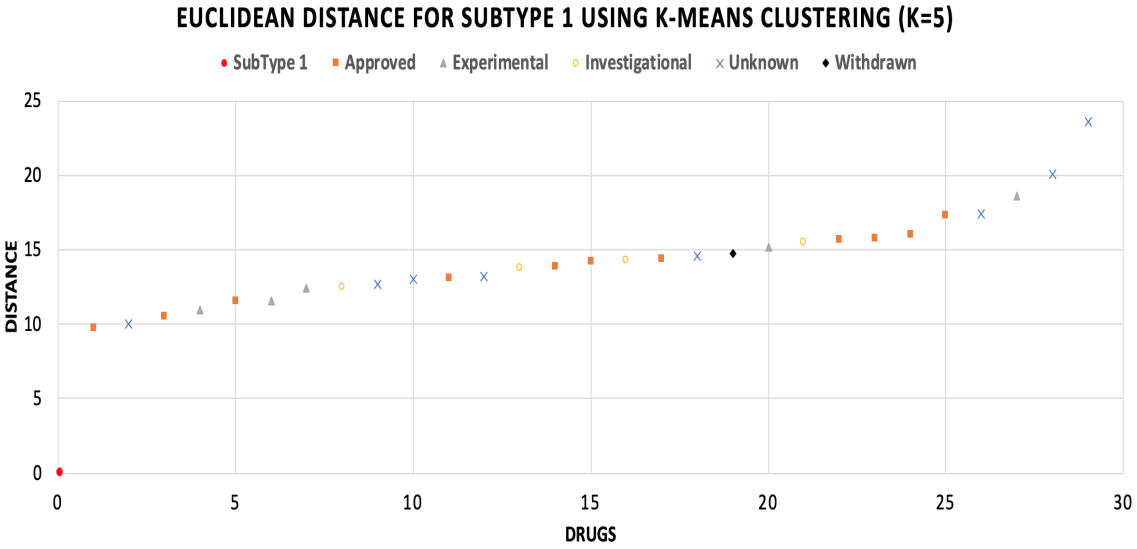

Figure 4: SubType 1 ranked list of drugs: K-Means (K=5)

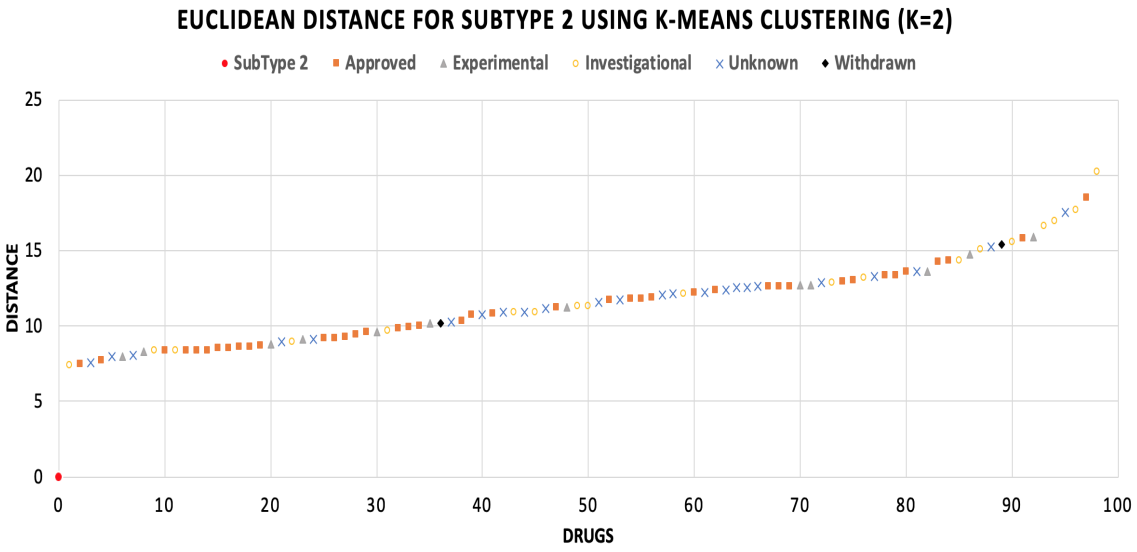

Figure 5: SubType 2 ranked list of drugs: K-Means (K=2)

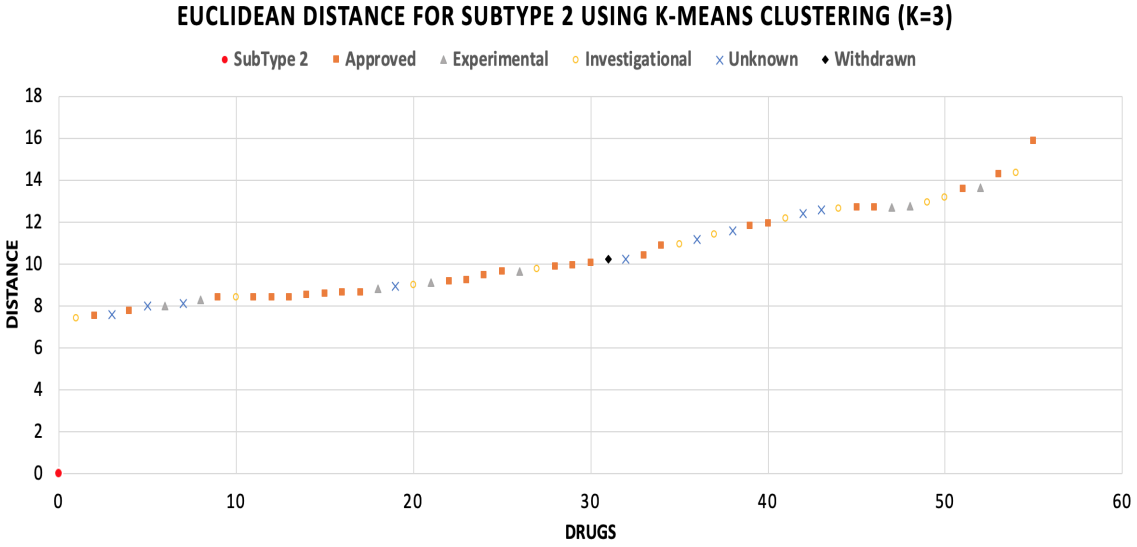

Figure 6: SubType 2 ranked list of drugs: K-Means (K=3)

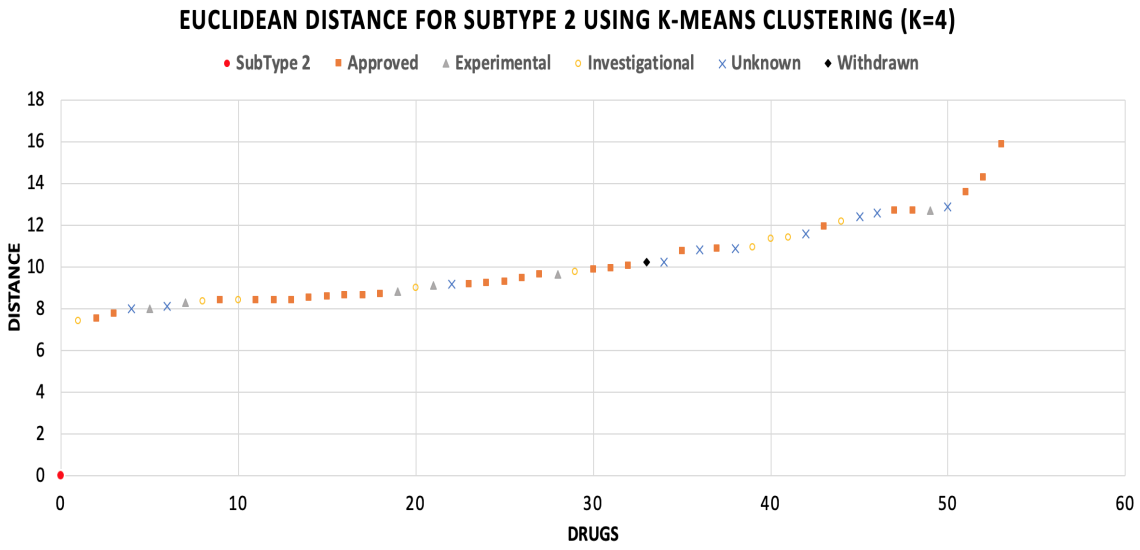

Figure 7: SubType 2 ranked list of drugs: K-Means (K=4)

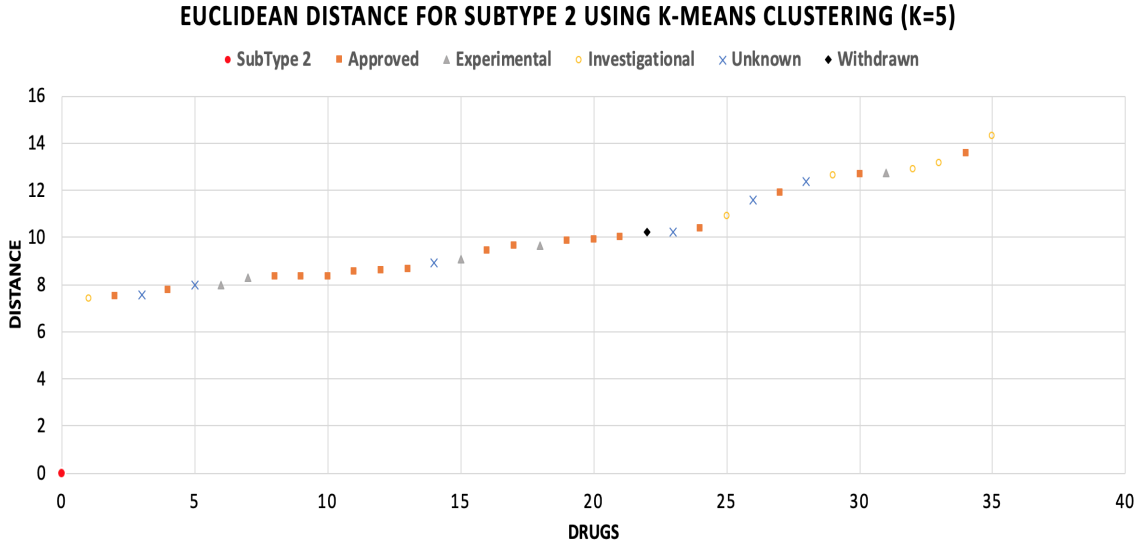

Figure 8: SubType 2 ranked list of drugs: K-Means (K=5)

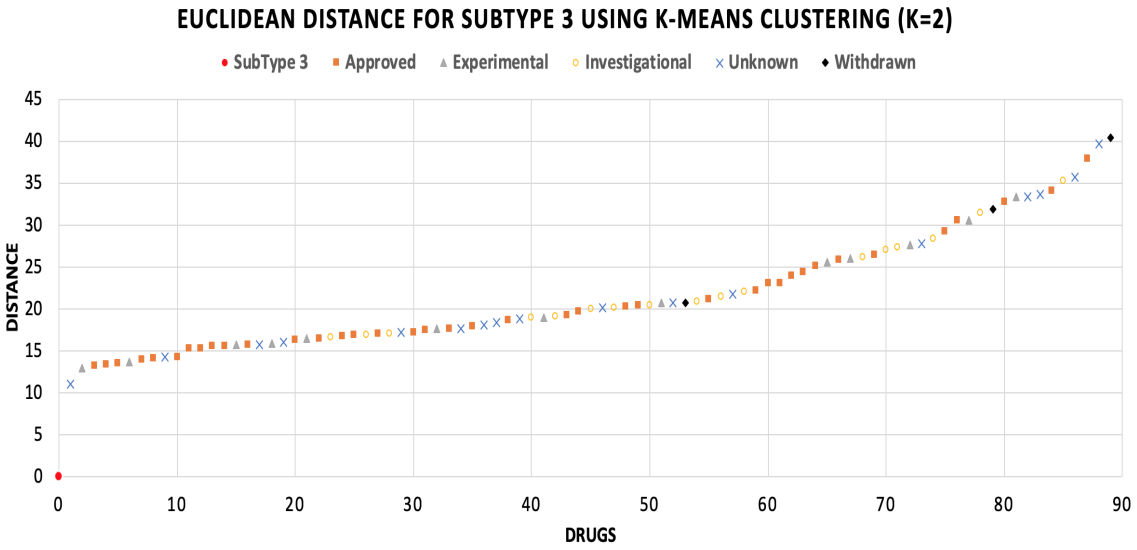

Figure 9: SubType 3 ranked list of drugs: K-Means (K=2)

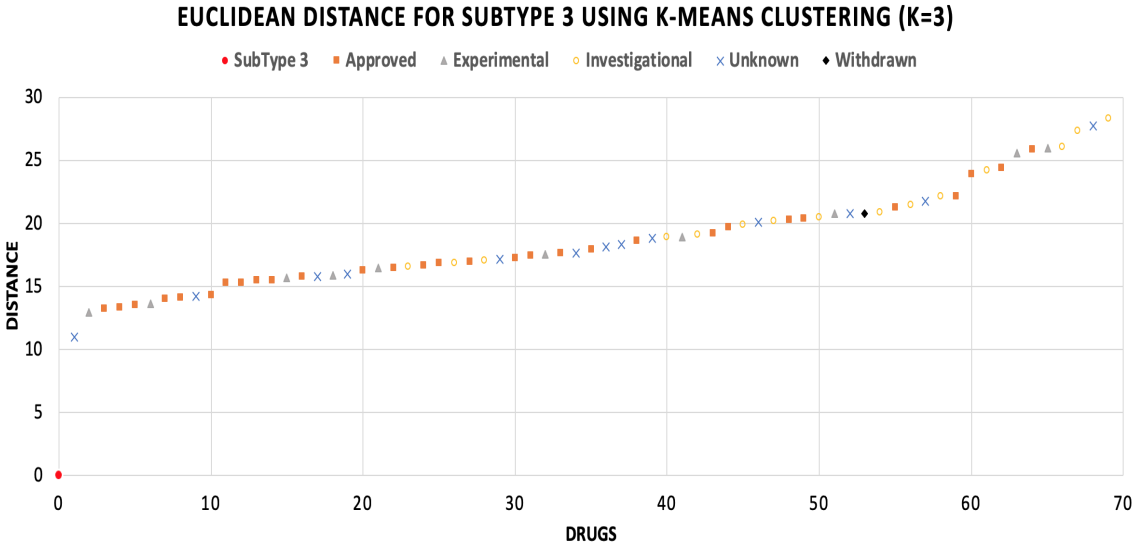

Figure 10: SubType 3 ranked list of drugs: K-Means (K=3)

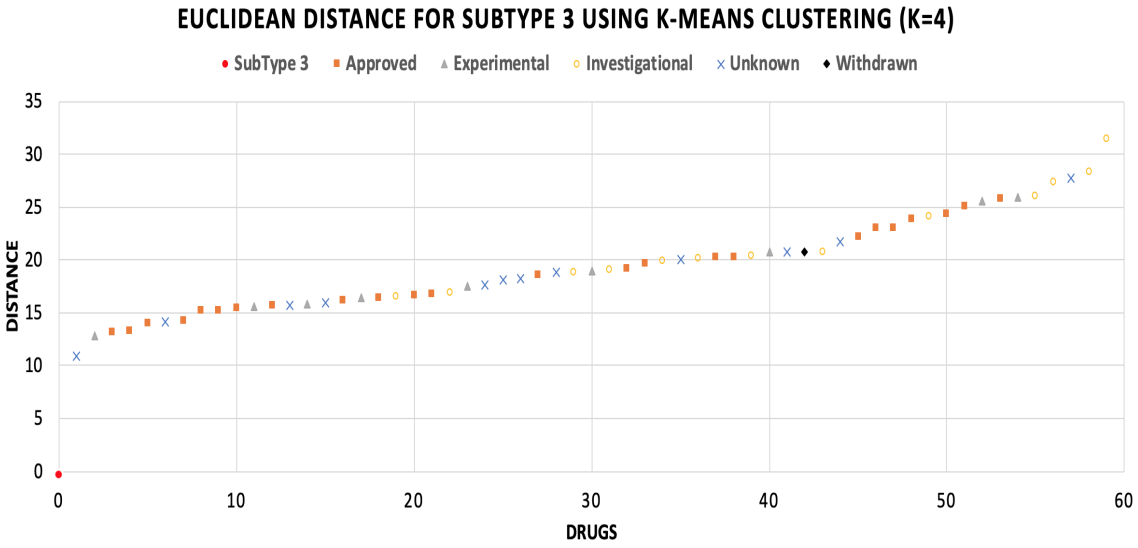

Figure 11: SubType 3 ranked list of drugs: K-Means (K=4)

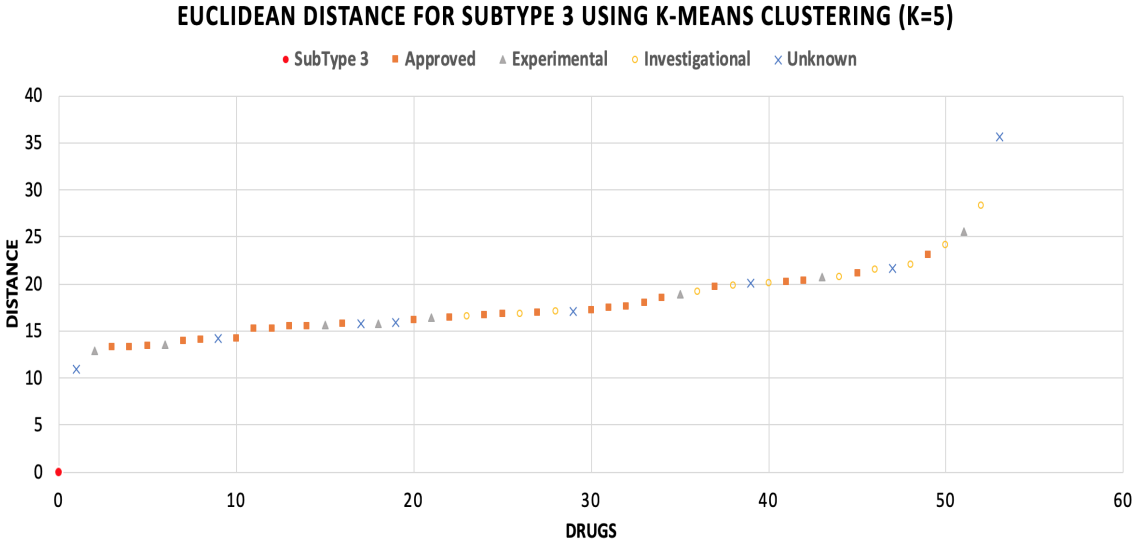

Figure 12: SubType 3 ranked list of drugs: K-Means (K=5)

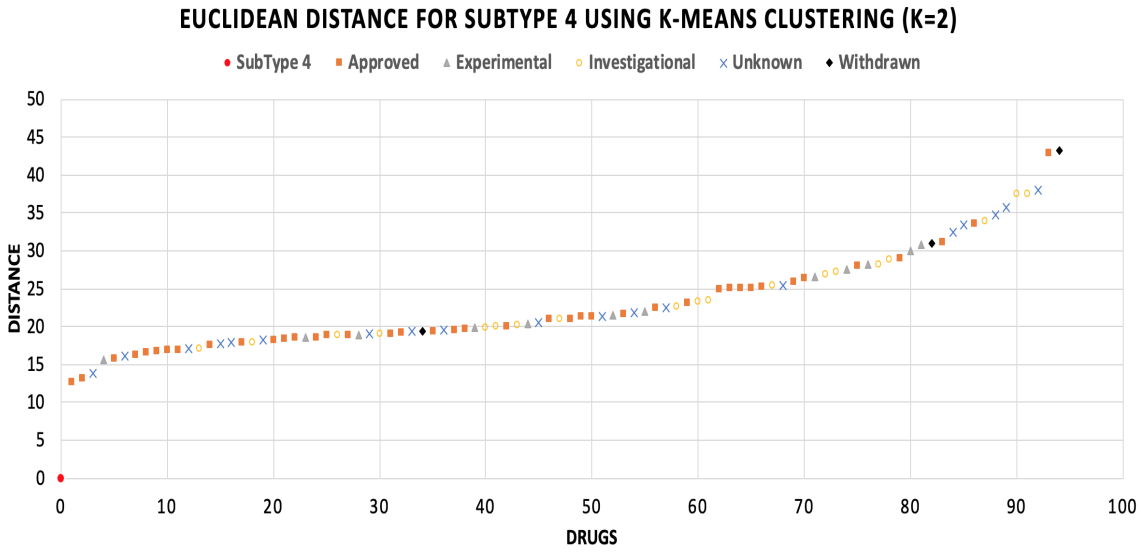

Figure 13: SubType 4 ranked list of drugs: K-Means (K=2)

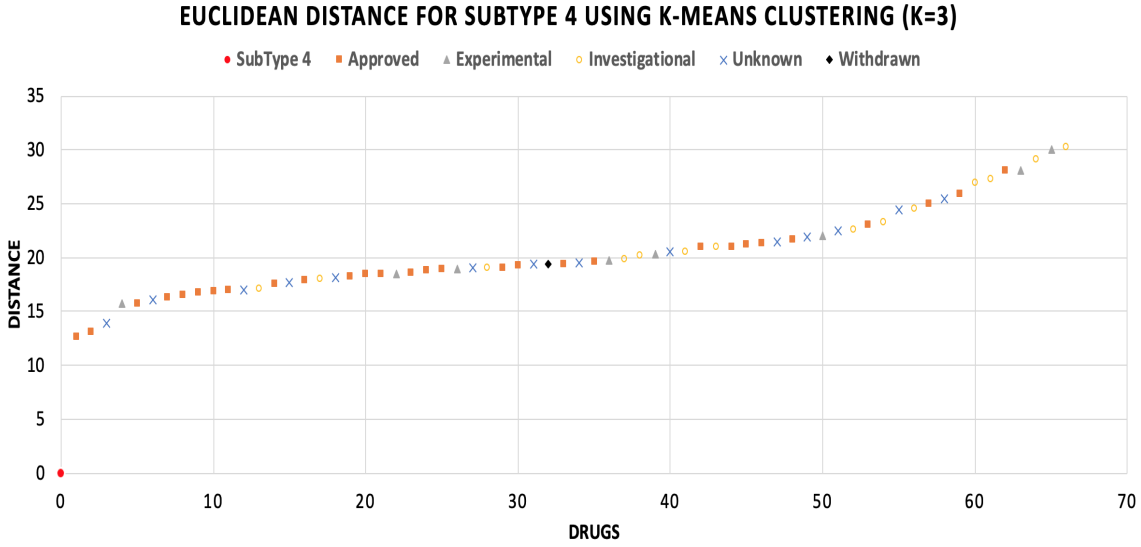

Figure 14: SubType 4 ranked list of drugs: K-Means (K=3)

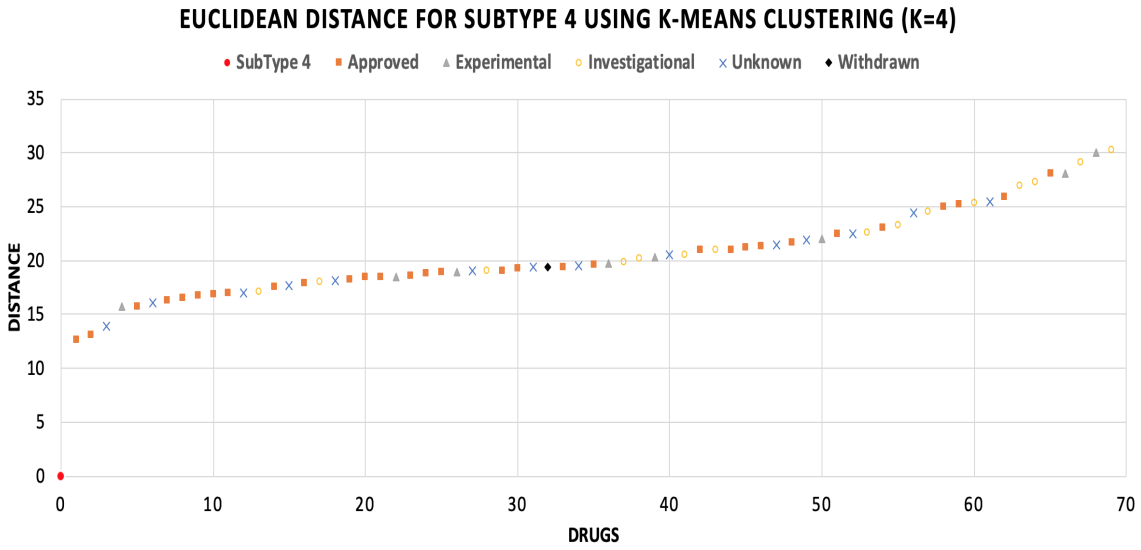

Figure 15: SubType 4 ranked list of drugs: K-Means (K=4)

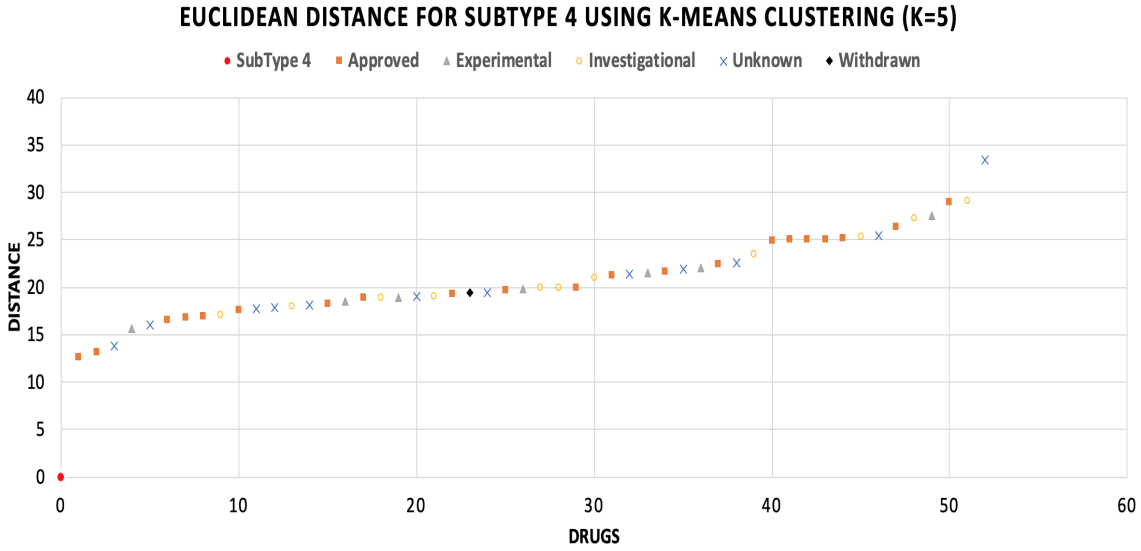

Figure 16: SubType 4 ranked list of drugs: K-Means (K=5)

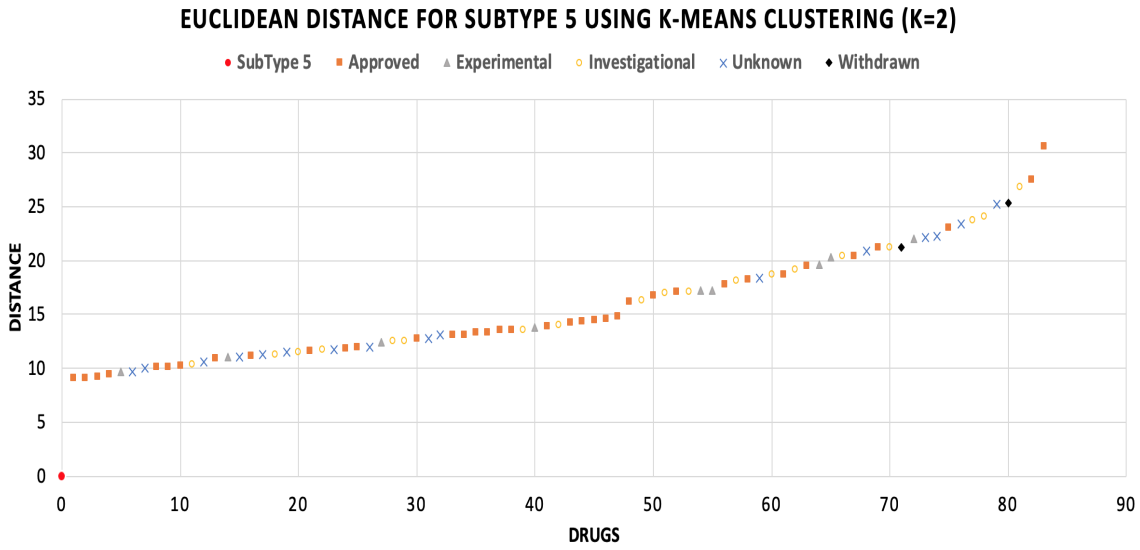

Figure 17: SubType 5 ranked list of drugs: K-Means (K=2)

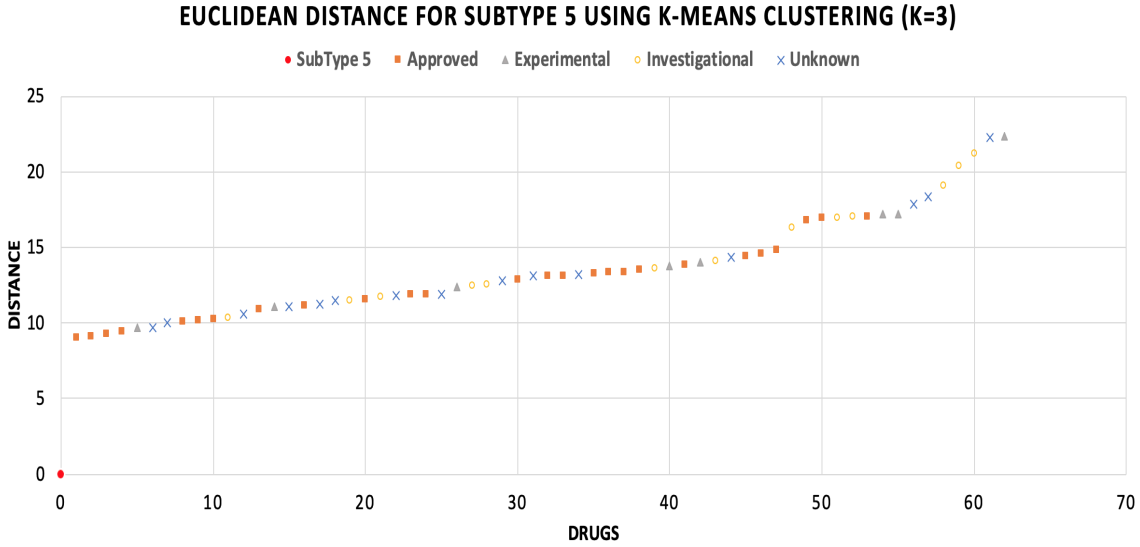

Figure 18: SubType 5 ranked list of drugs: K-Means (K=3)

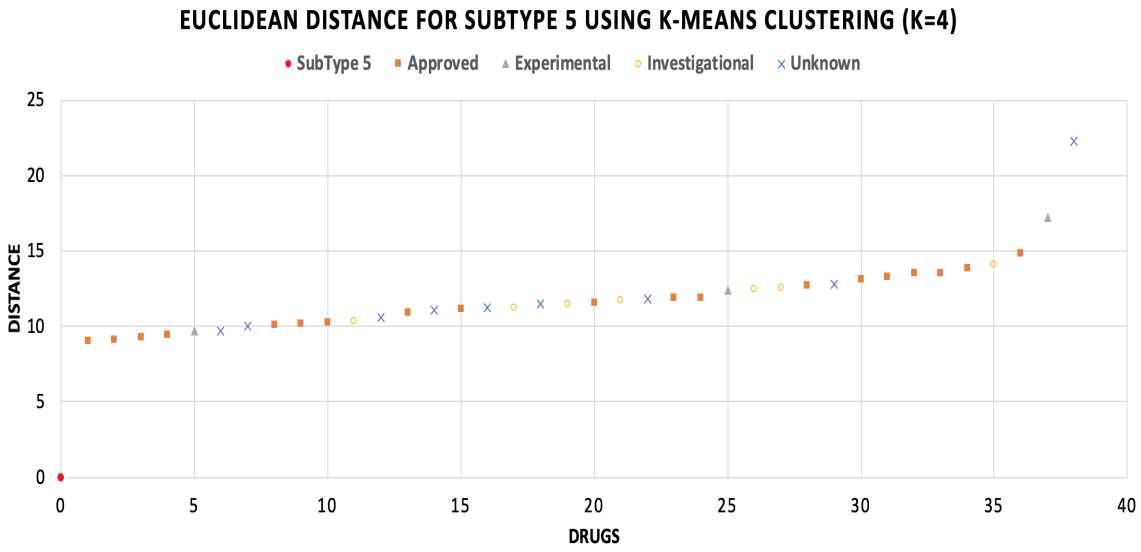

Figure 19: SubType 5 ranked list of drugs: K-Means (K=4)

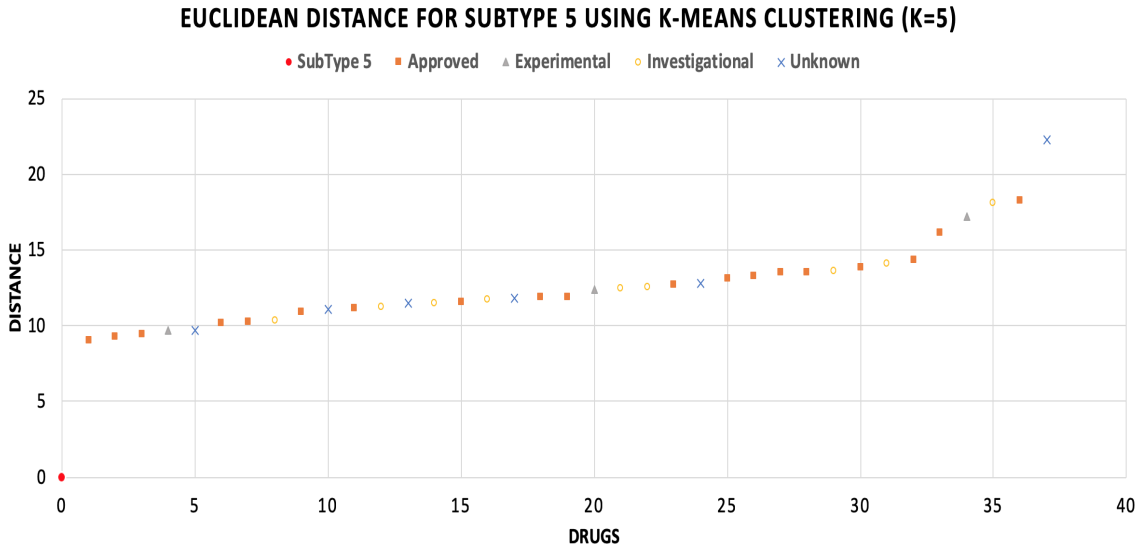

Figure 20: SubType 5 ranked list of drugs: K-Means (K=5)

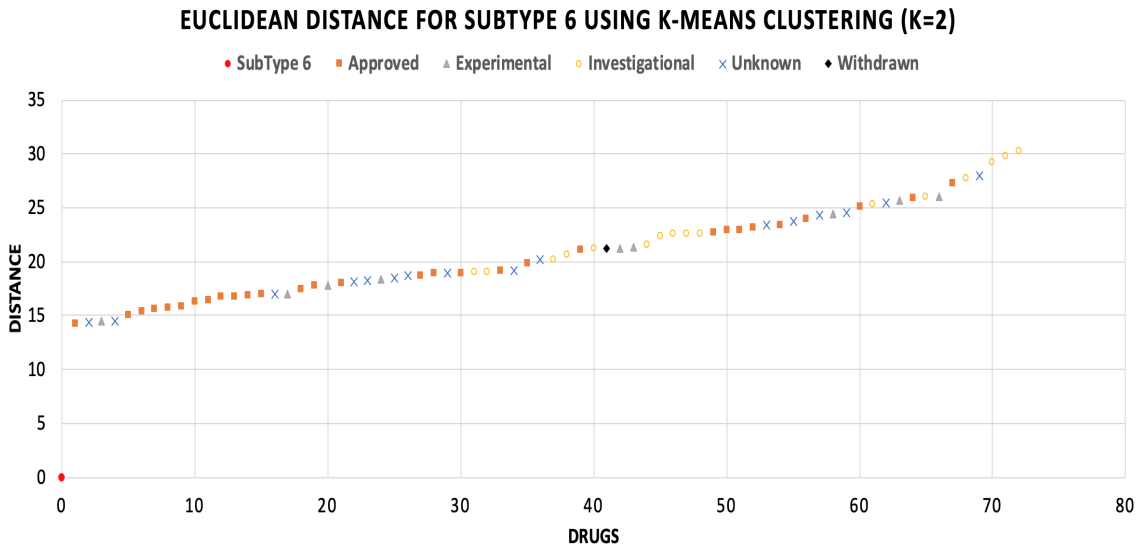

Figure 21: SubType 6 ranked list of drugs: K-Means (K=2)

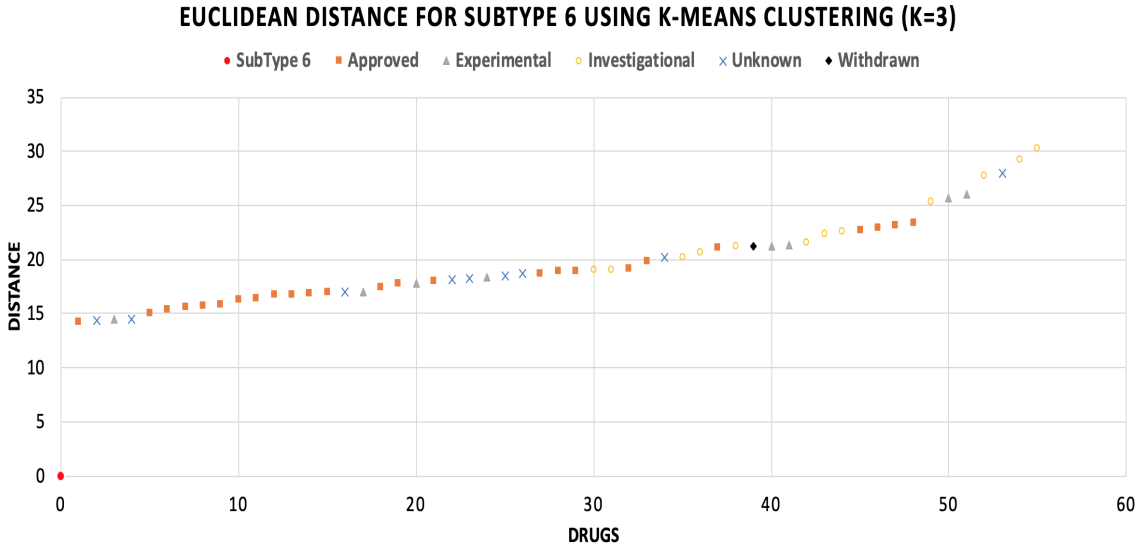

Figure 22: SubType 6 ranked list of drugs: K-Means (K=3)

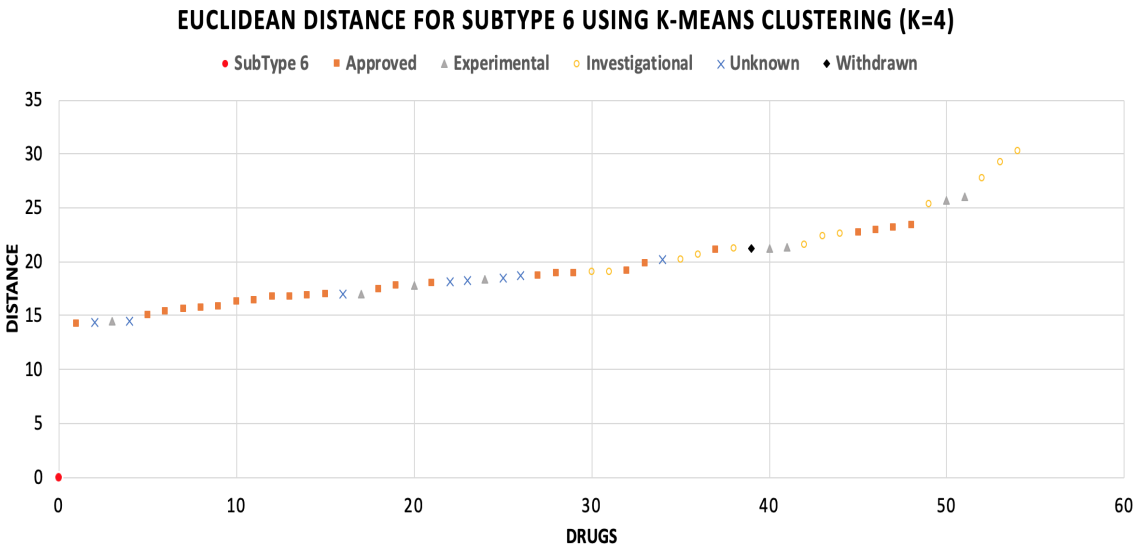

Figure 23: SubType 6 ranked list of drugs: K-Means (K=4)

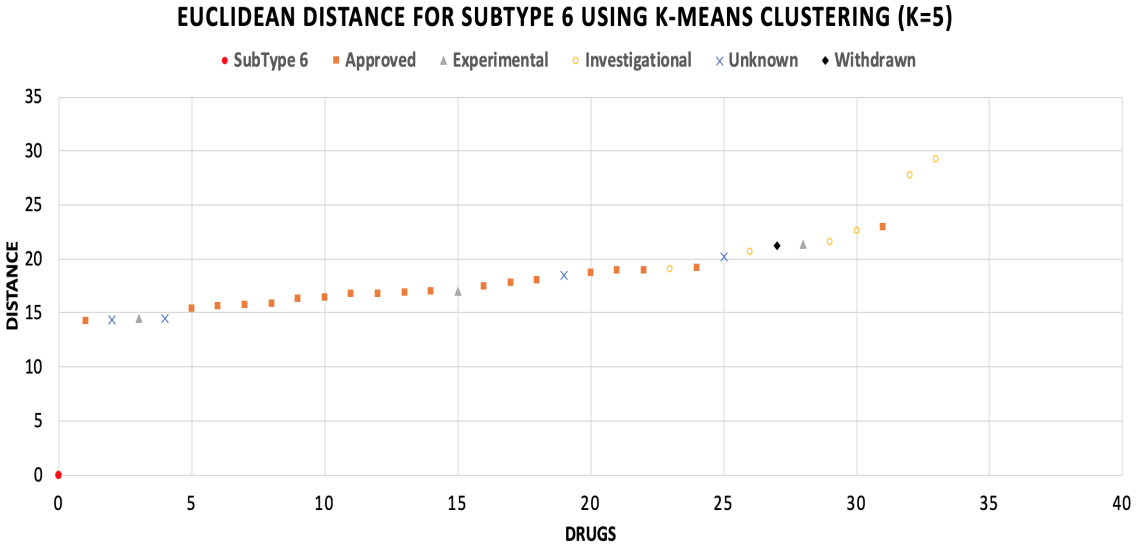

Figure 24: SubType 6 ranked list of drugs: K-Means (K=5)

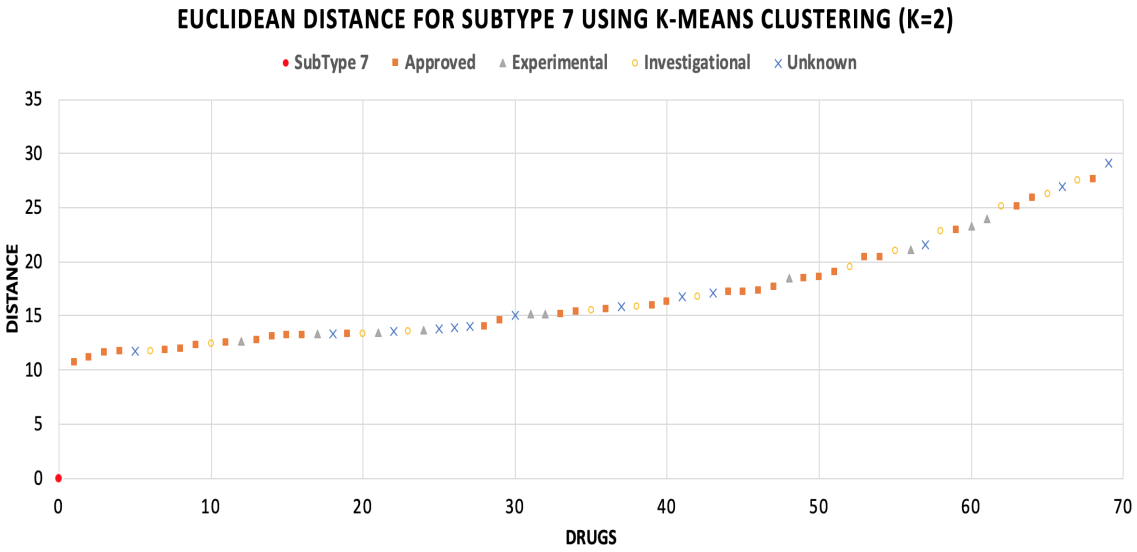

Figure 25: SubType 7 ranked list of drugs: K-Means (K=2)

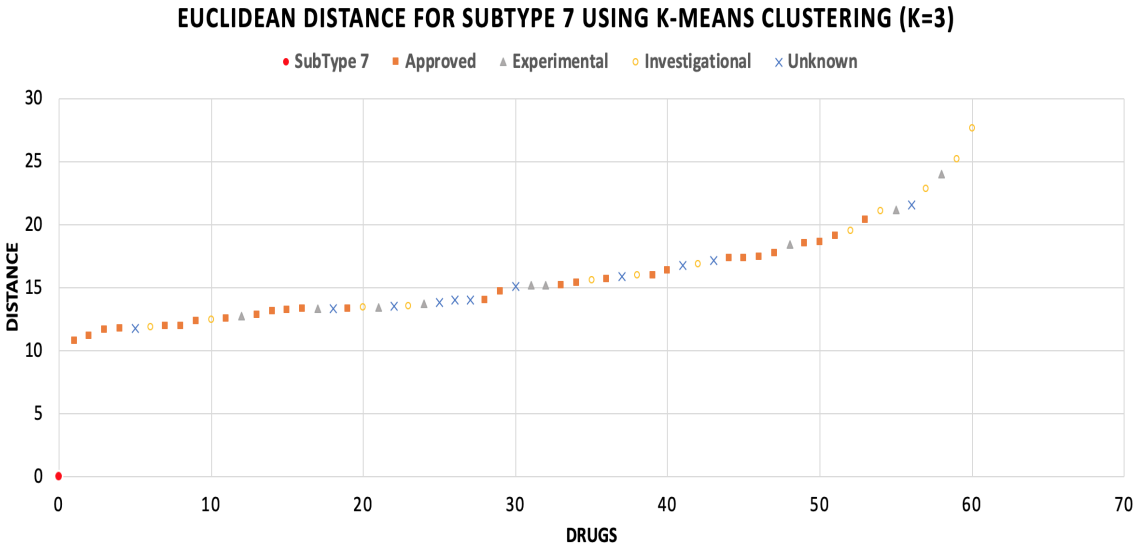

Figure 26: SubType 7 ranked list of drugs: K-Means (K=3)

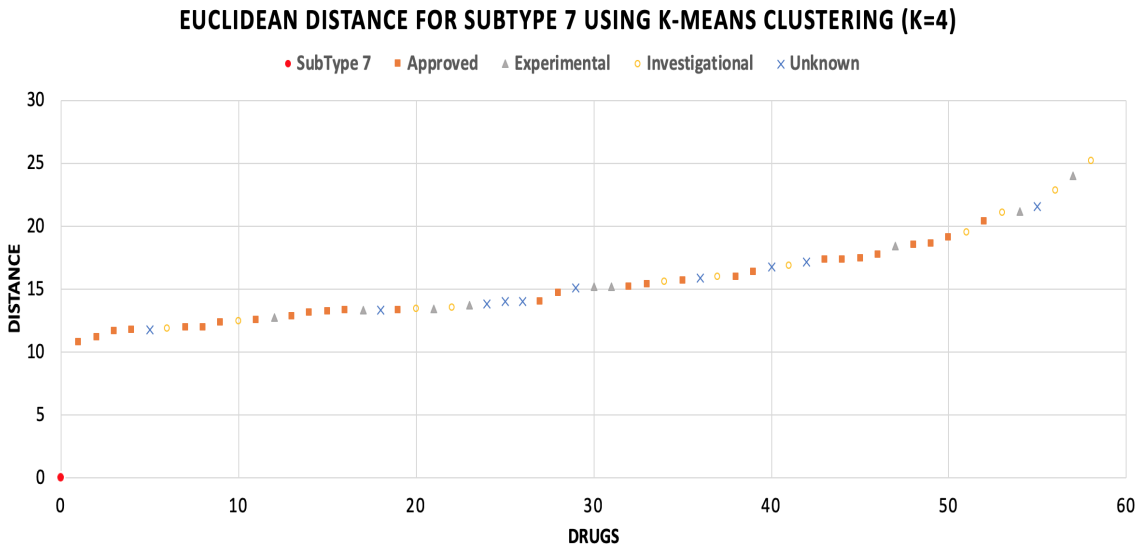

Figure 27: SubType 7 ranked list of drugs: K-Means (K=4)

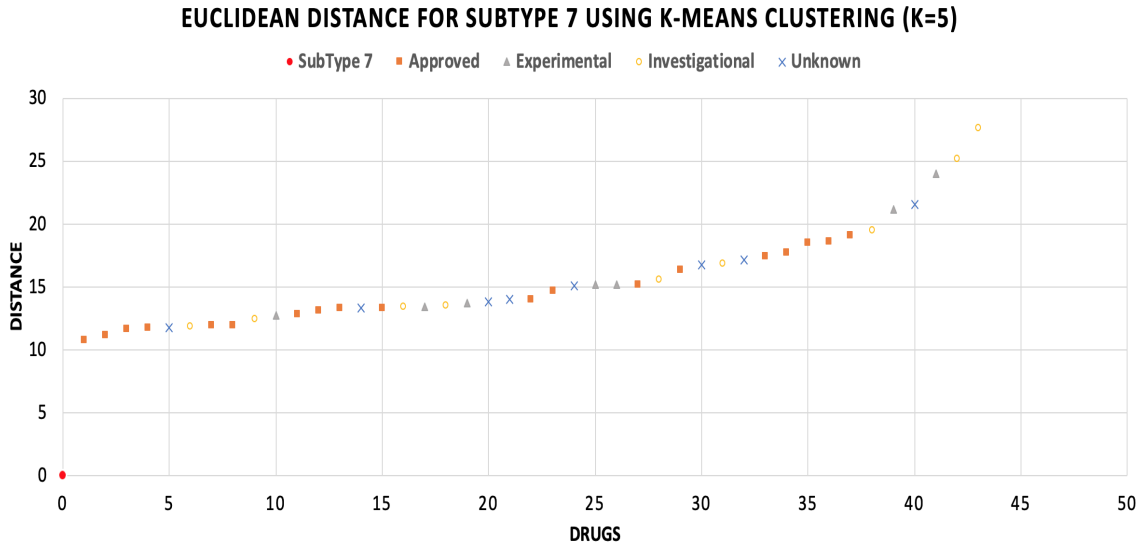

Figure 28: SubType 7 ranked list of drugs: K-Means (K=5)

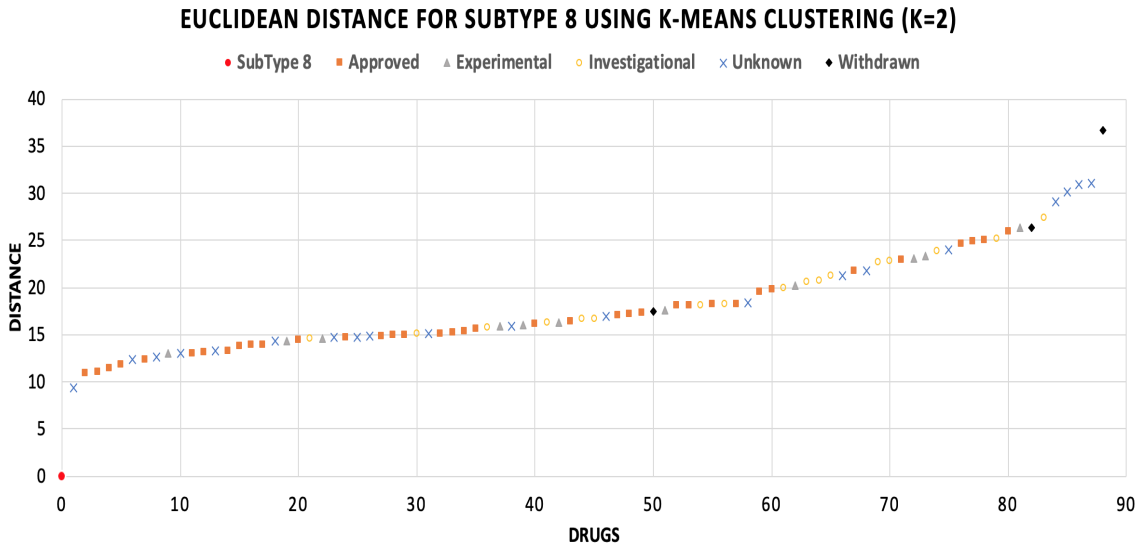

Figure 29: SubType 8 ranked list of drugs: K-Means (K=2)

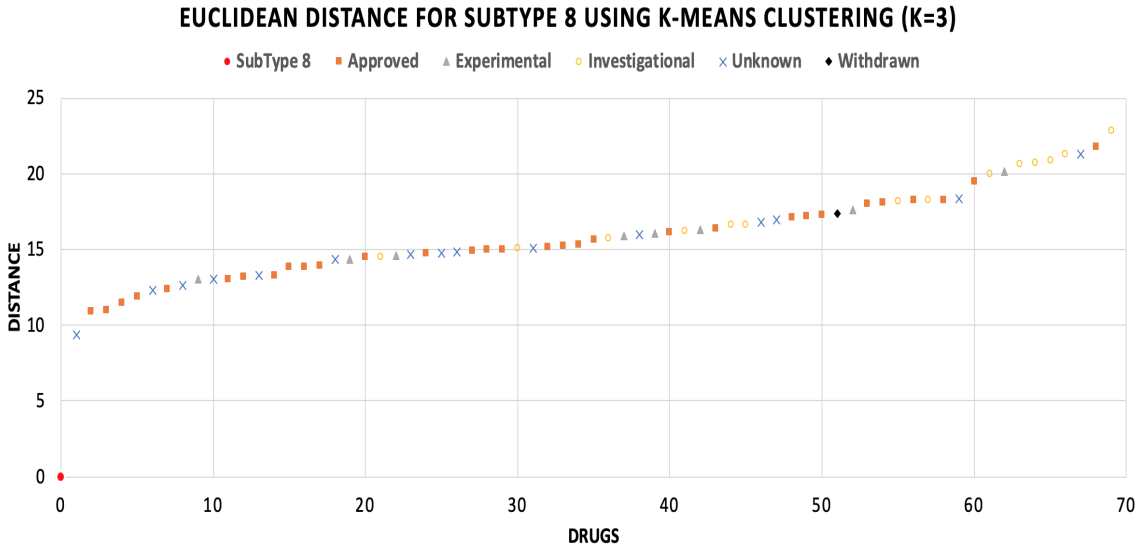

Figure 30: SubType 8 ranked list of drugs: K-Means (K=3)

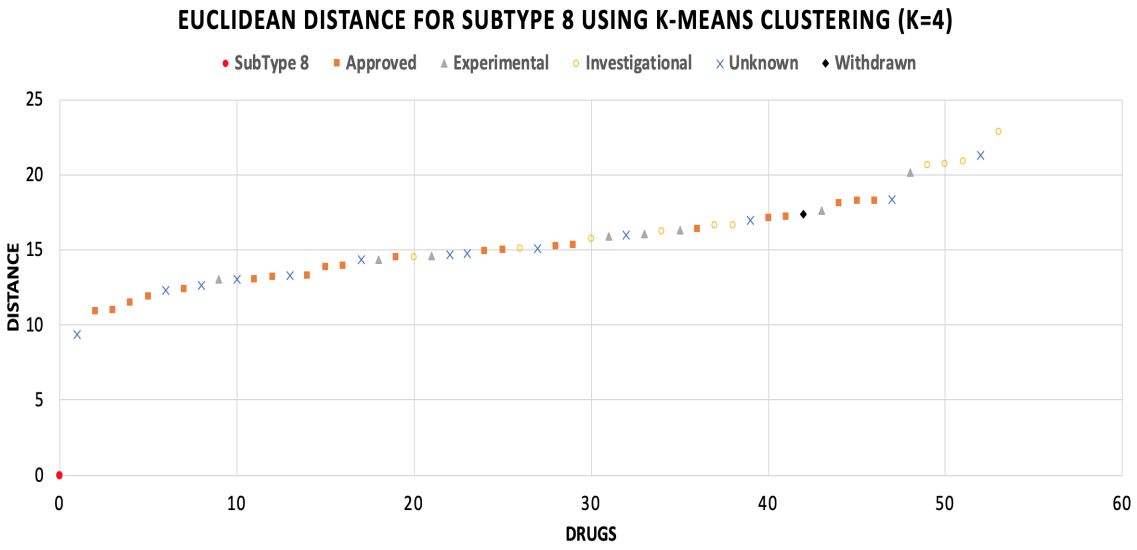

Figure 31: SubType 8 ranked list of drugs: K-Means (K=4)

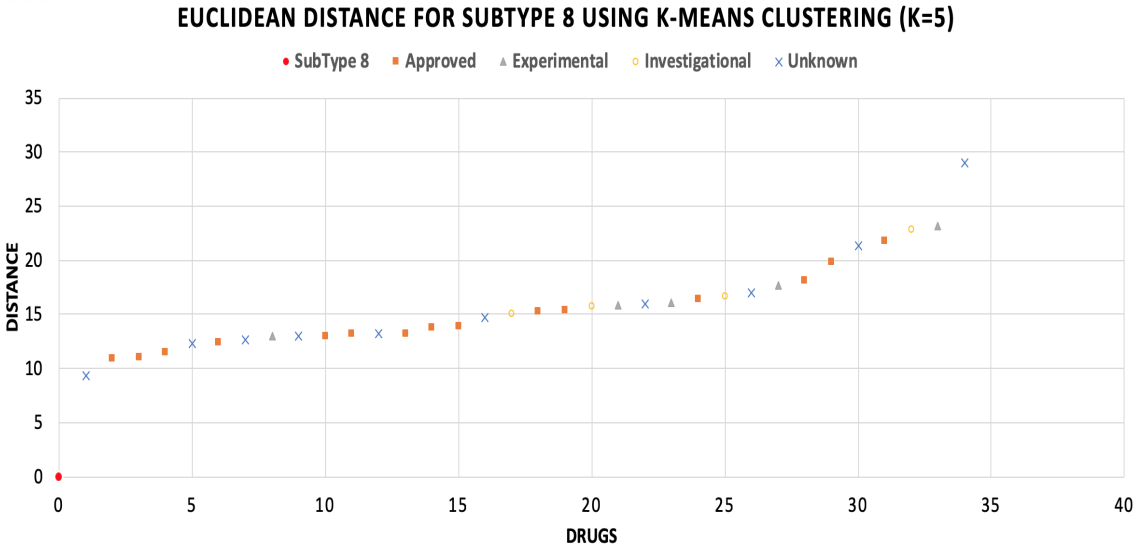

Figure 32: SubType 8 ranked list of drugs: K-Means (K=5)

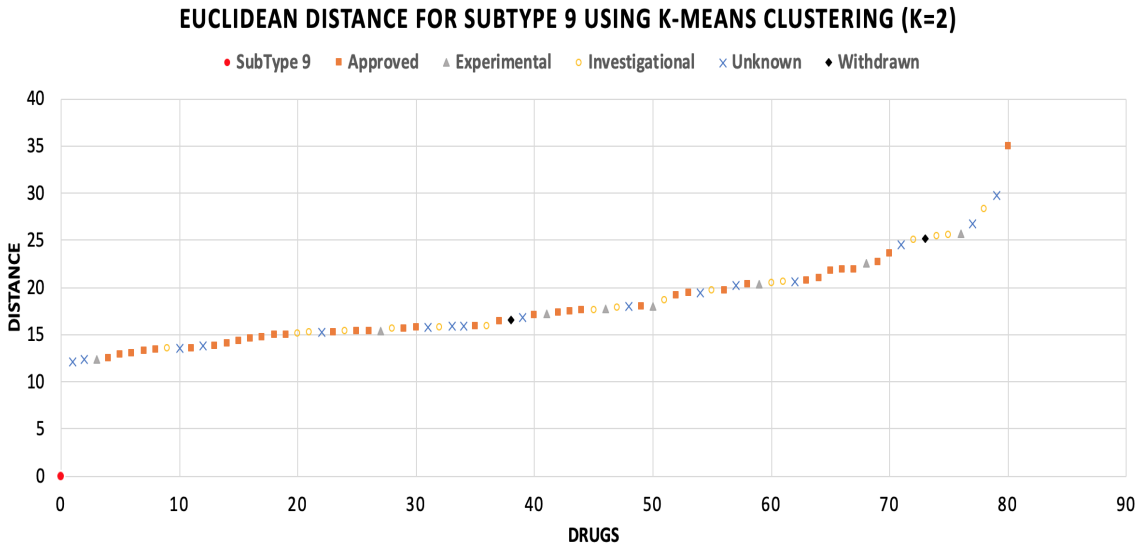

Figure 33: SubType 9 ranked list of drugs: K-Means (K=2)

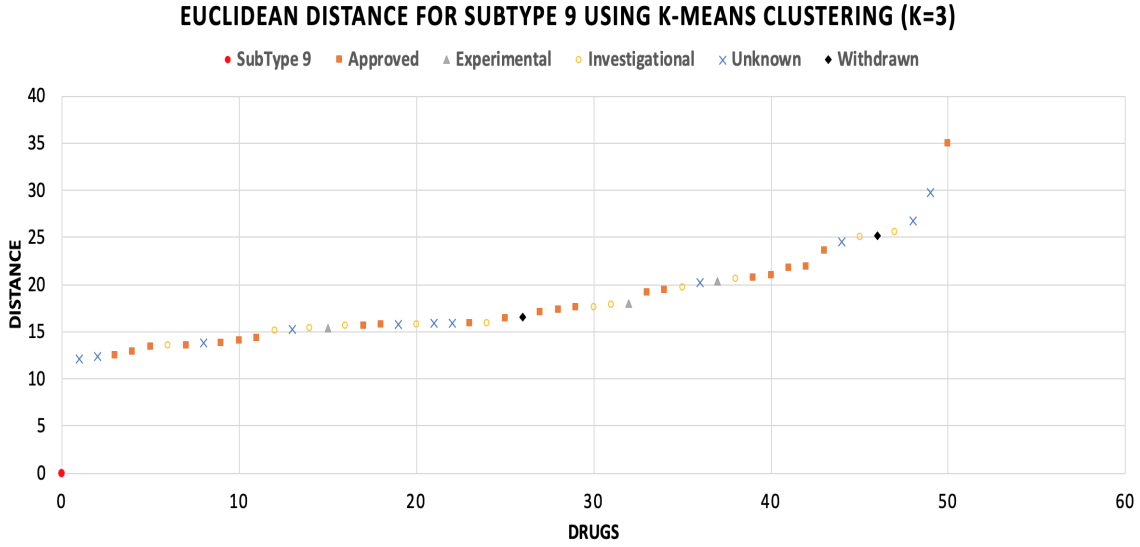

Figure 34: SubType 9 ranked list of drugs: K-Means (K=3)

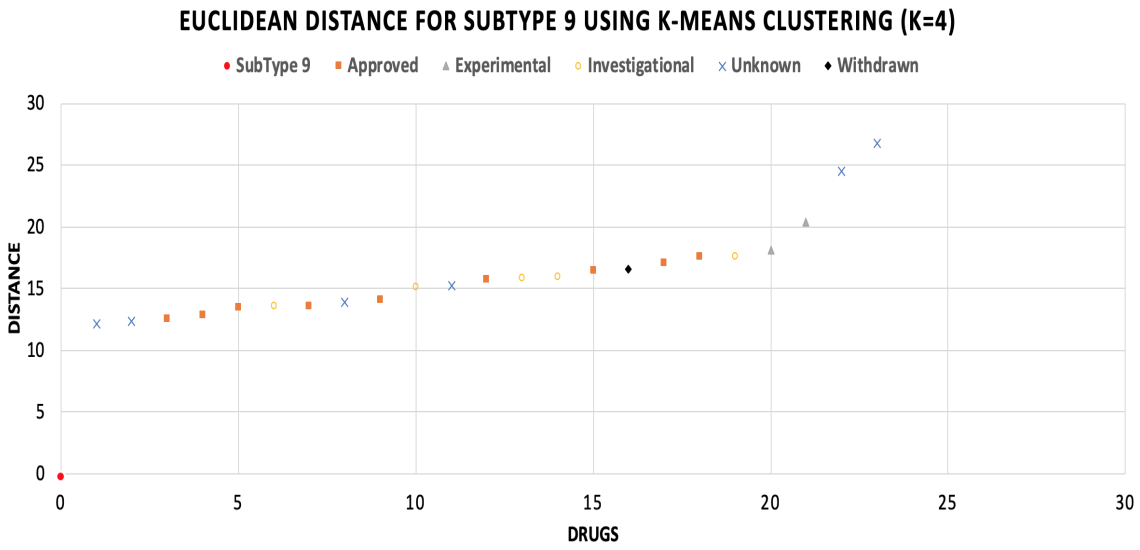

Figure 35: SubType 9 ranked list of drugs: K-Means (K=4)

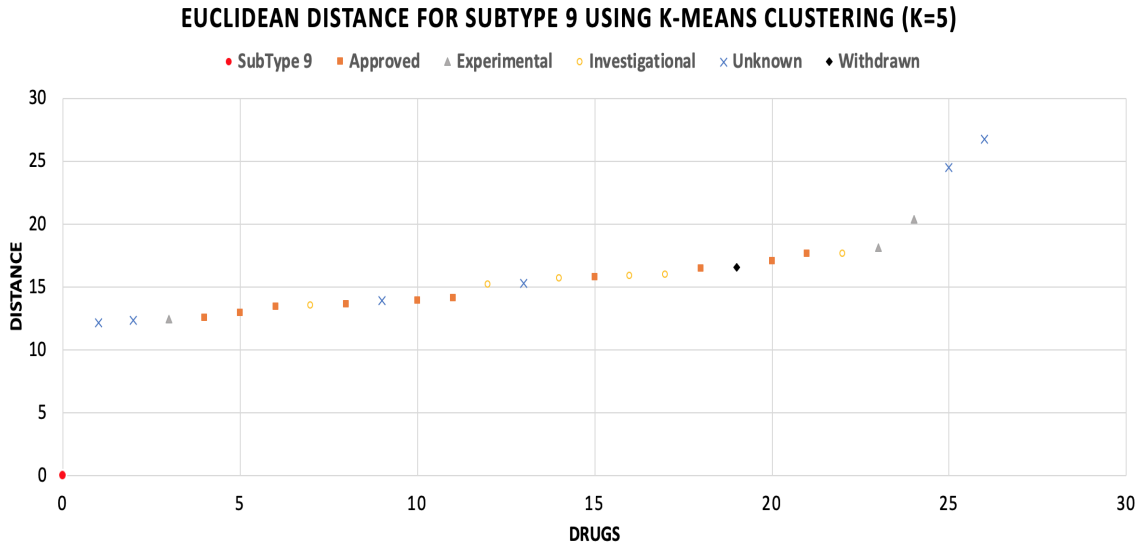

Figure 36: SubType 9 ranked list of drugs: K-Means (K=5)

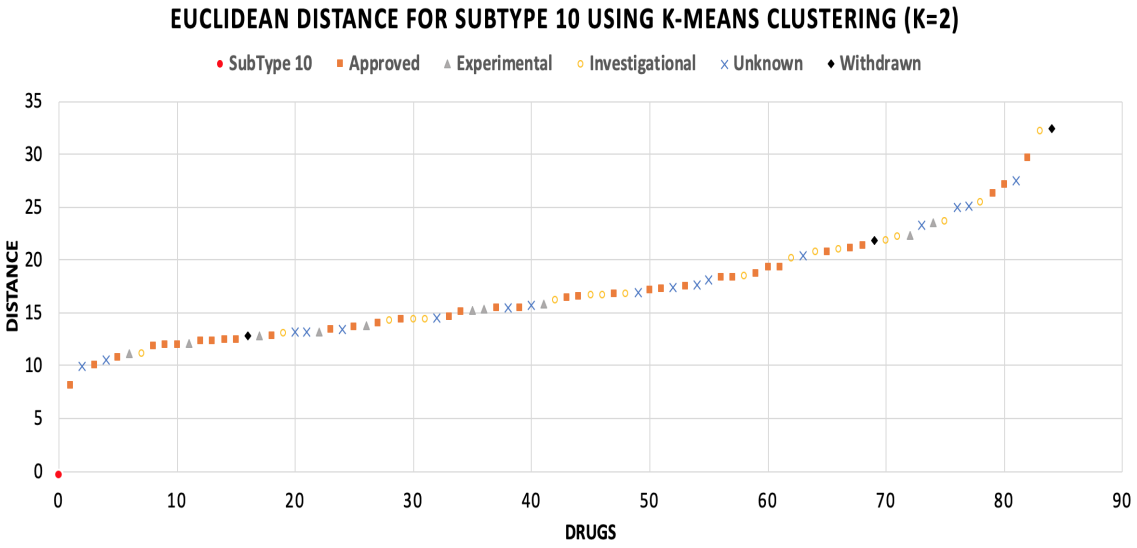

Figure 37: SubType 10 ranked list of drugs: K-Means (K=2)

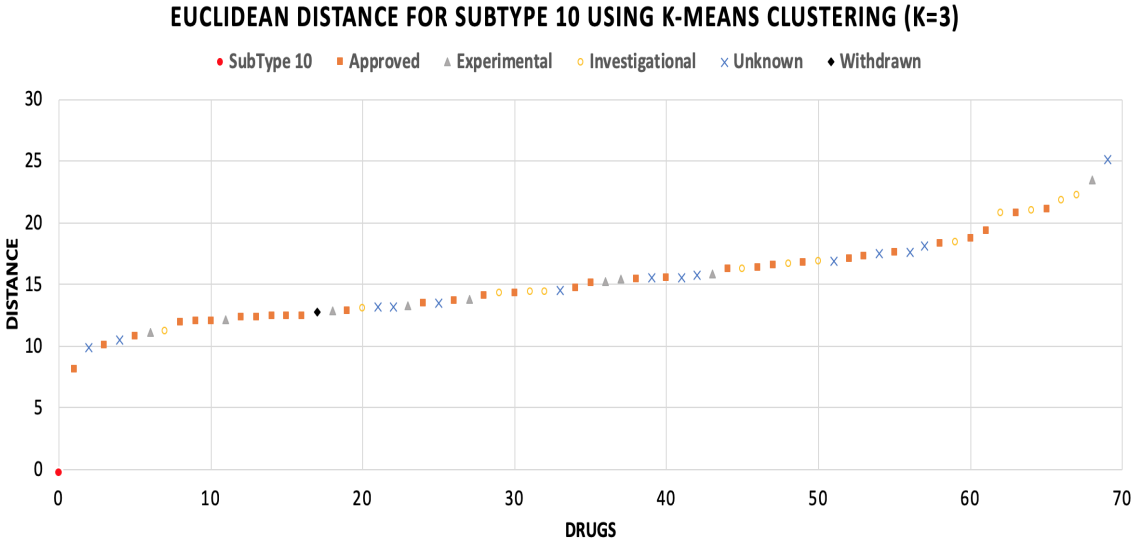

Figure 38: SubType 10 ranked list of drugs: K-Means (K=3)

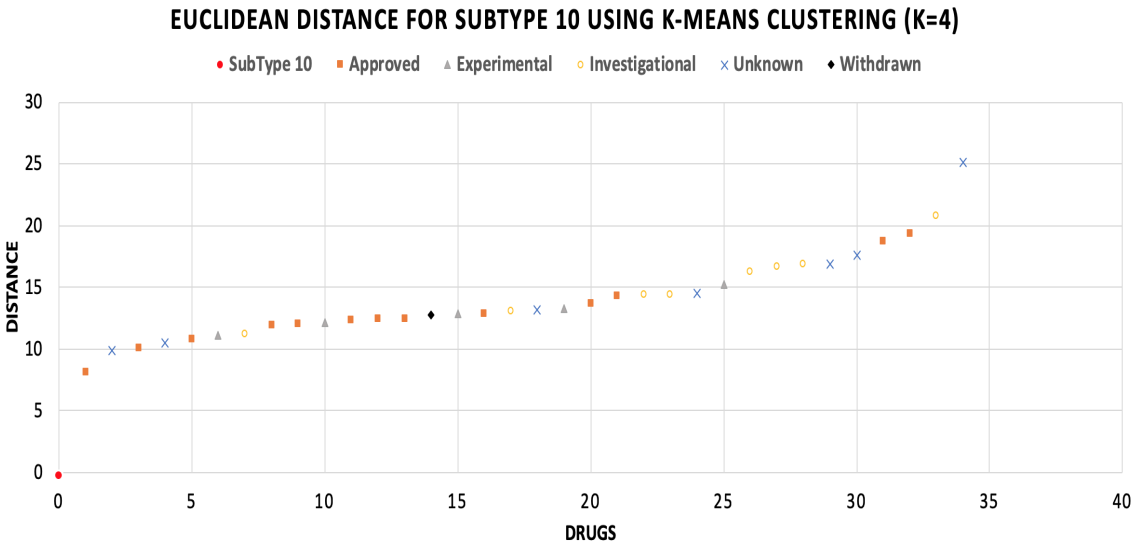

Figure 39: SubType 10 ranked list of drugs: K-Means (K=4)

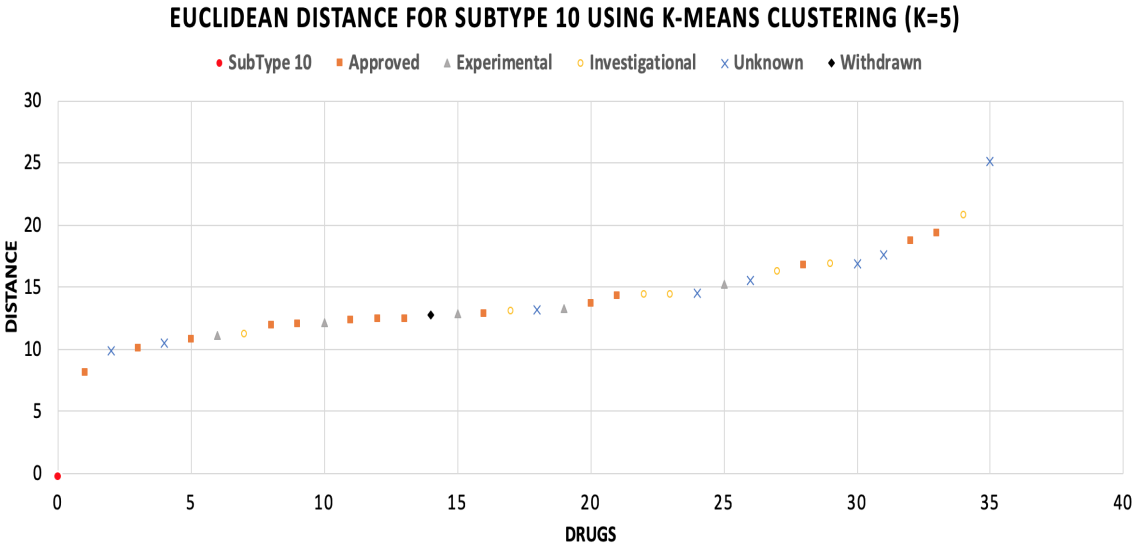

Figure 40: SubType 10 ranked list of drugs: K-Means (K=5)

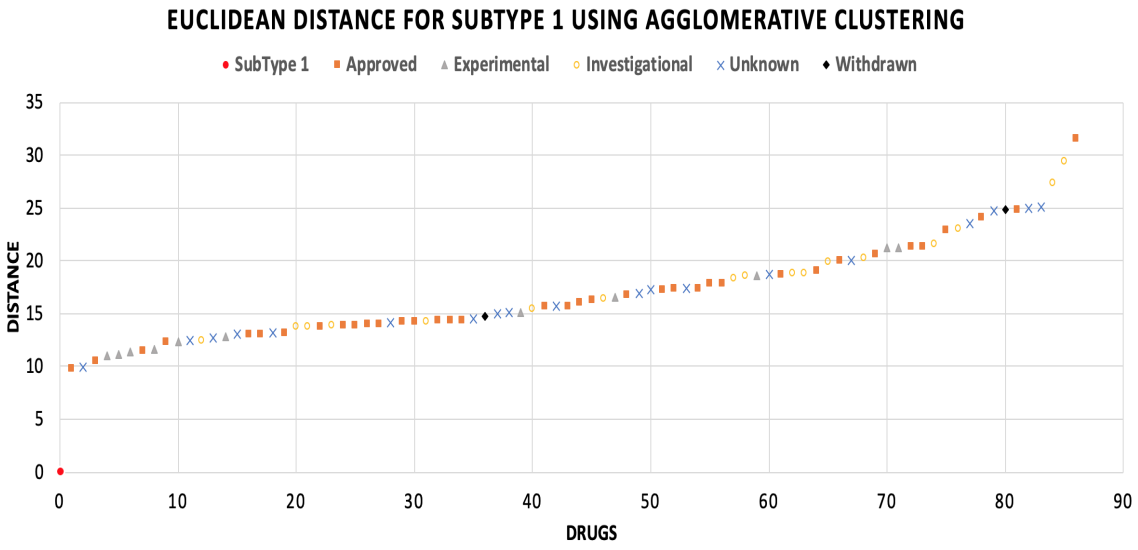

Figure 41: SubType 1 ranked list of drugs: Agglomerative

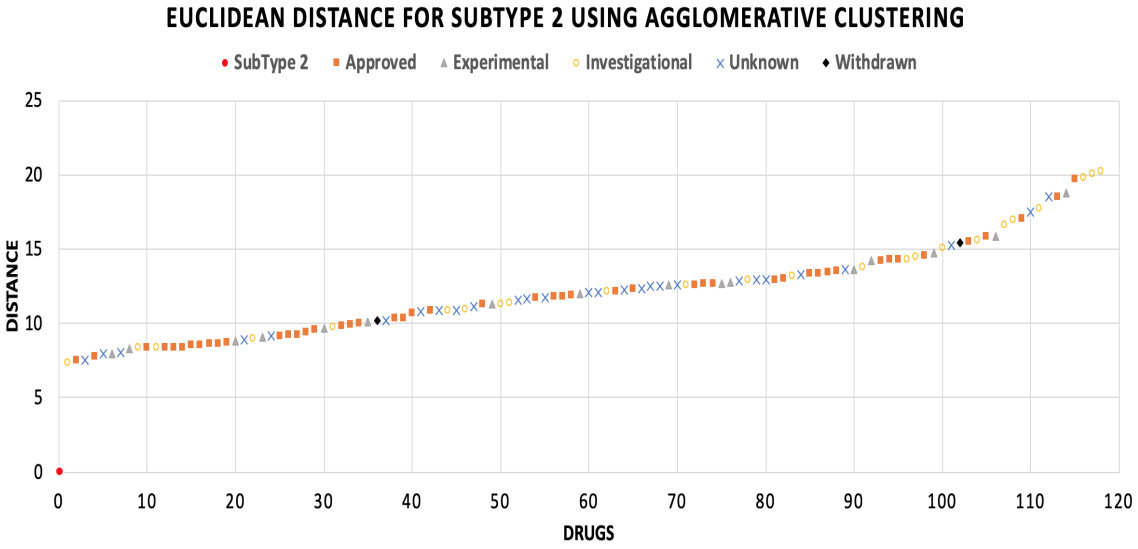

Figure 42: SubType 2 ranked list of drugs: Agglomerative

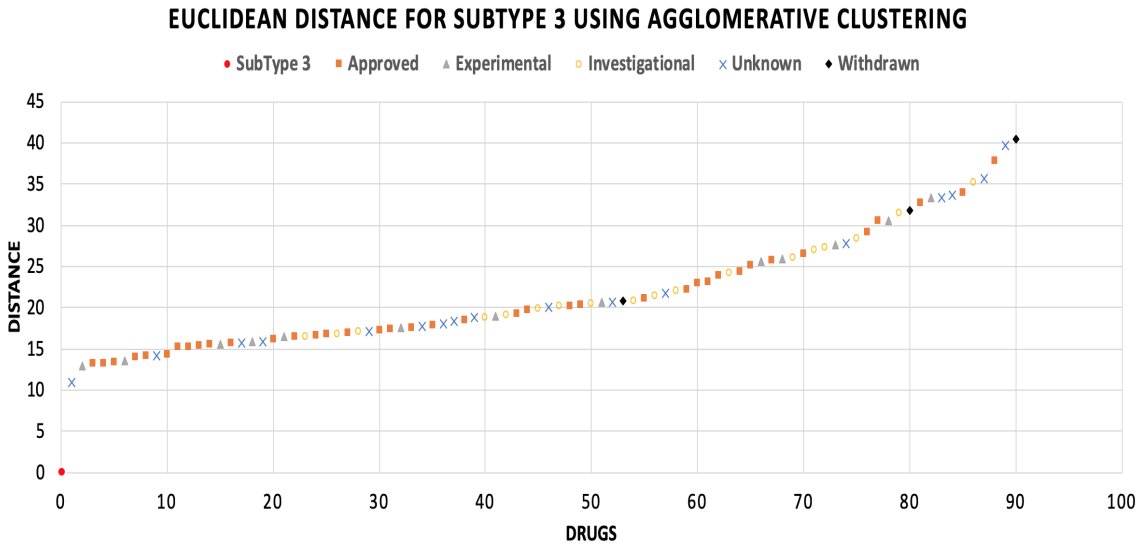

Figure 43: SubType 3 ranked list of drugs: Agglomerative

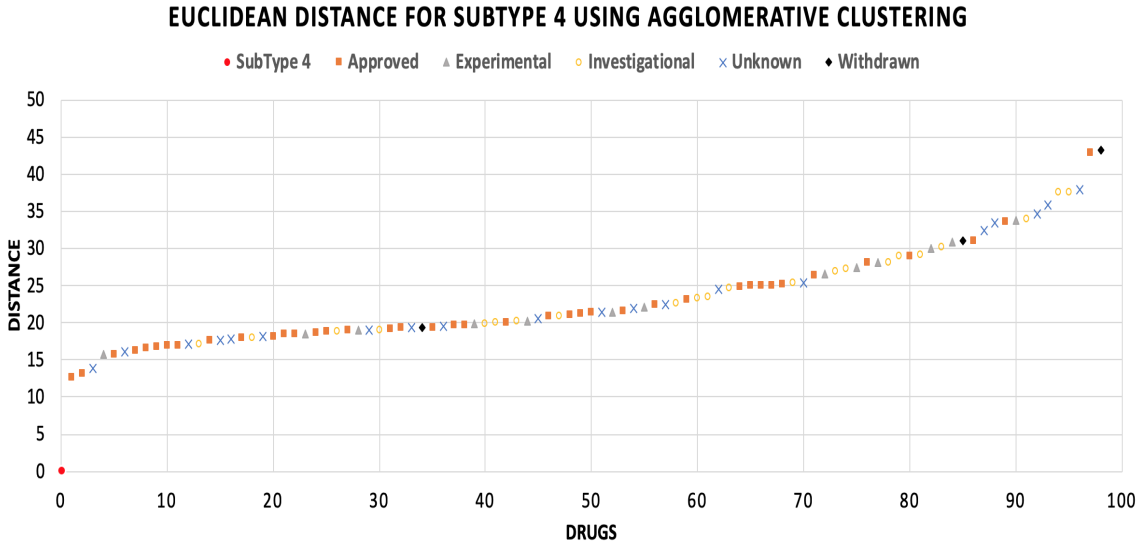

Figure 44: SubType 4 ranked list of drugs: Agglomerative

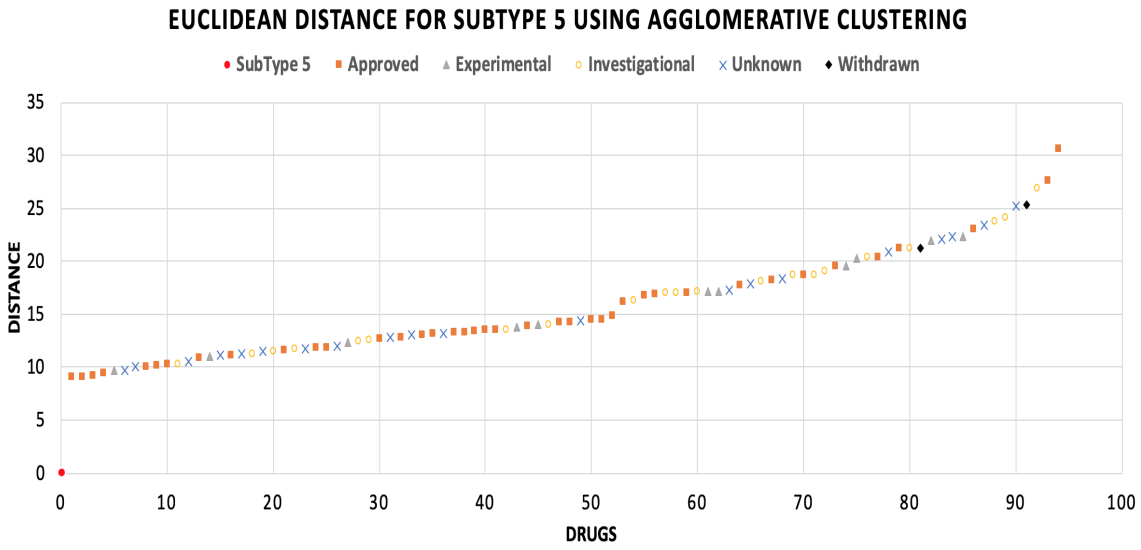

Figure 45: SubType 5 ranked list of drugs: Agglomerative

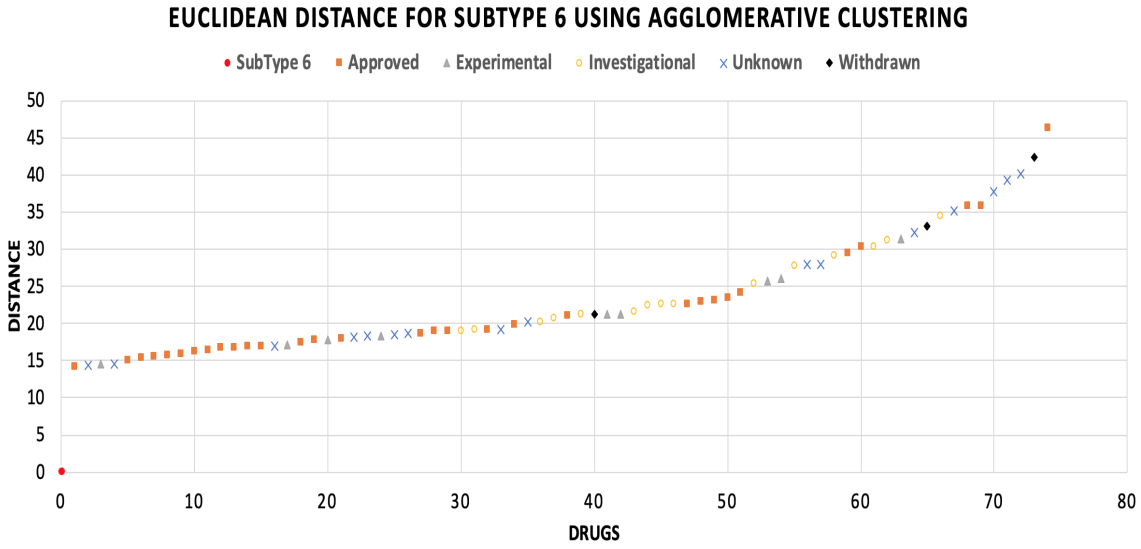

Figure 46: SubType 6 ranked list of drugs: Agglomerative

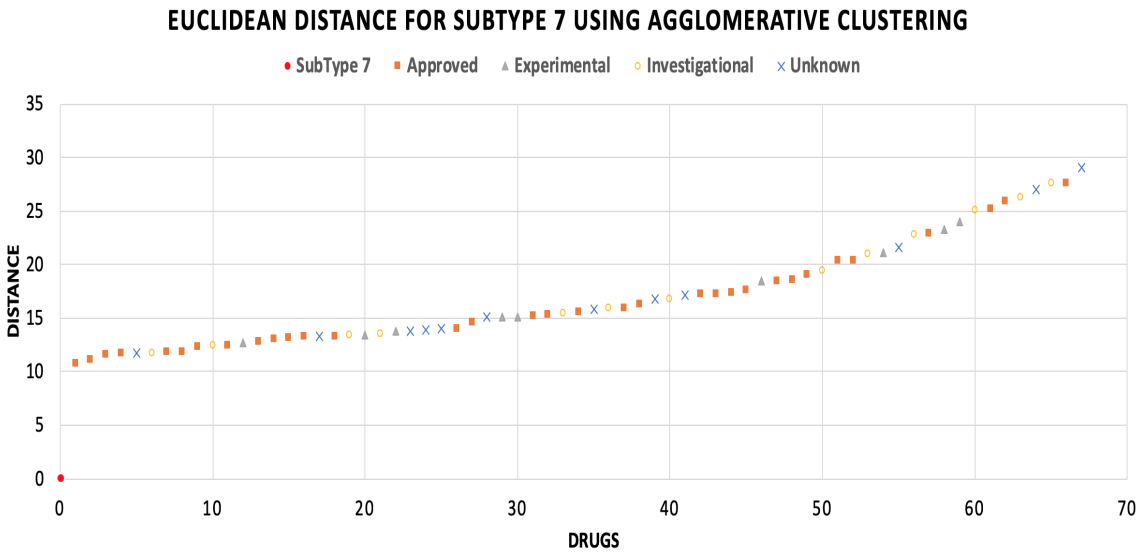

Figure 47: SubType 7 ranked list of drugs: Agglomerative

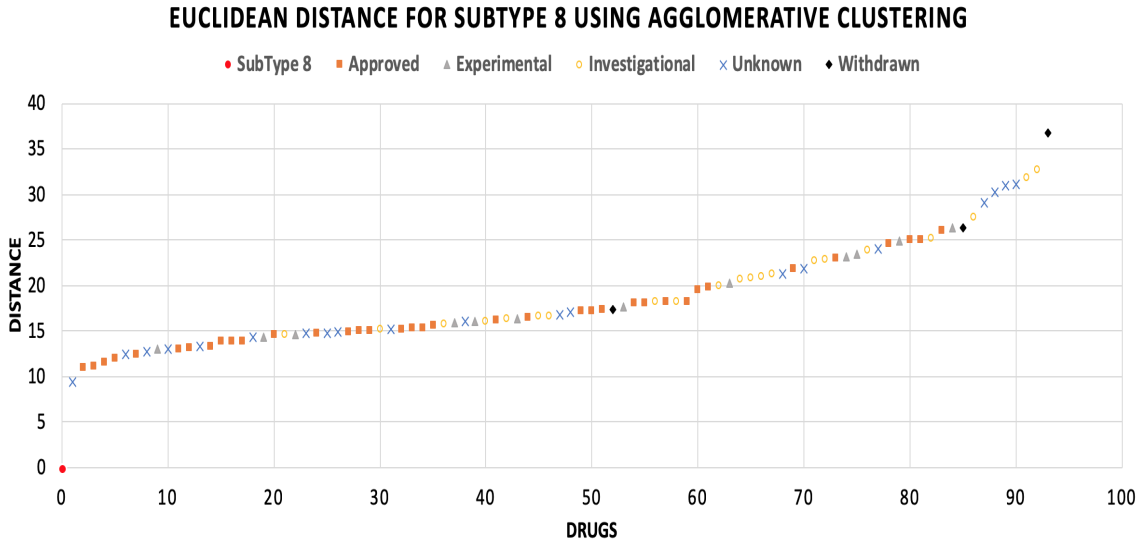

Figure 48: SubType 8 ranked list of drugs: Agglomerative

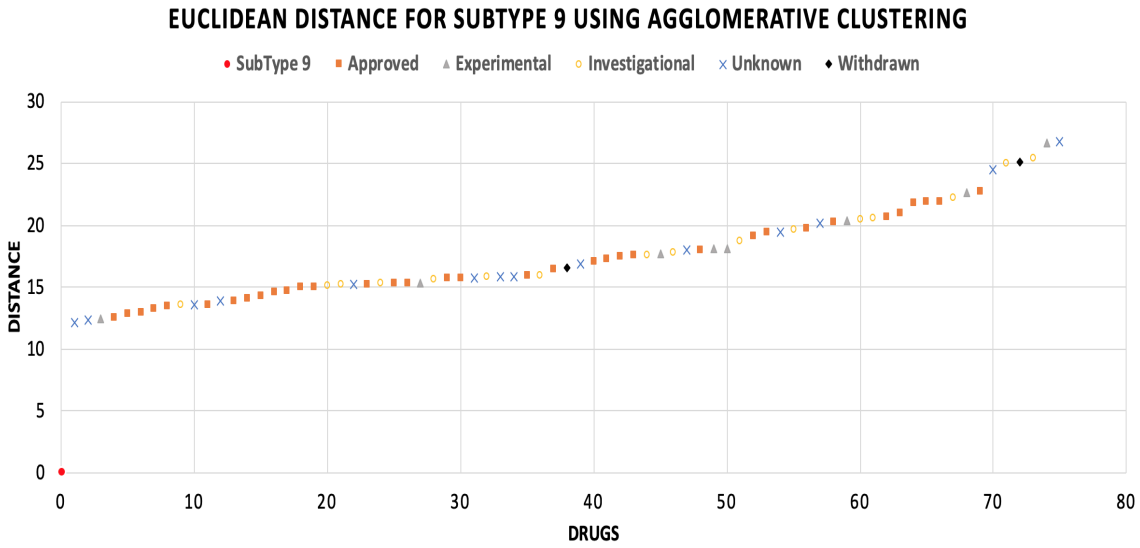

Figure 49: SubType 9 ranked list of drugs: Agglomerative

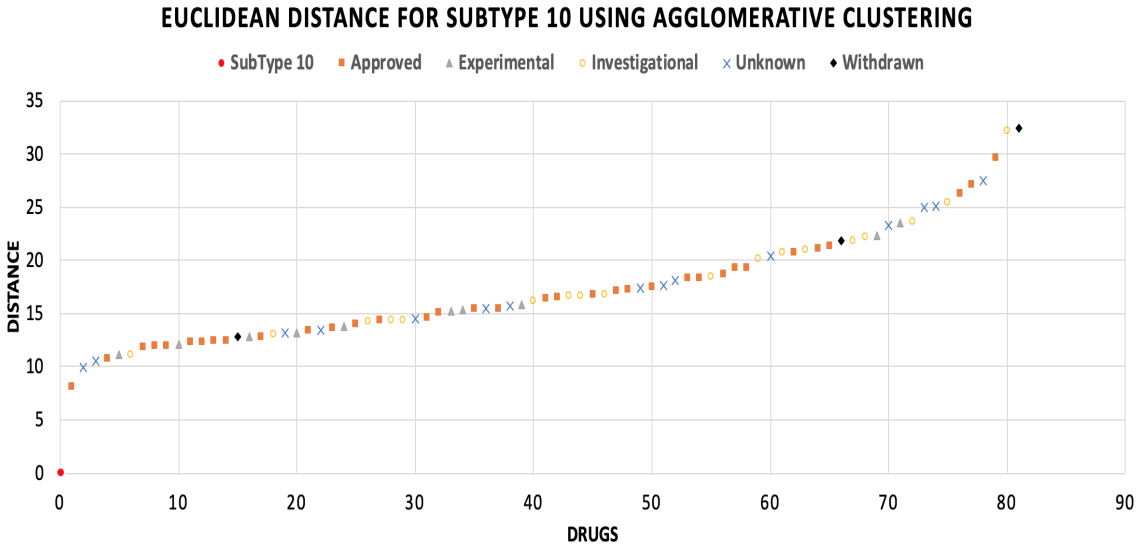

Figure 50: SubType 10 ranked list of drugs: Agglomerative

Vita Auctoris

NAME:	Roopesh Dhara
PLACE OF BIRTH:	Visakhapatnam, India
EDUCATION:	Bachelor of Technology in Computer Science and Engineering, SRM University, Kattankulathur, Tamil Nadu, India, 2016.
	Master of Science in Computer Science, University of Windsor, Windsor, Ontario, Canada, 2019.