**Section 1: Drug list curation procedure**

The following steps are performed for drug list preparation using OpenFDA data, the lines of Python code used are indicated in bold font

1. Obtain data from OpenFDA website [Data Accessed: 6th May,2021,[18]
2. Unzip file, and load in Python, use following code

**import json**

**f=open(r'filepath\drug-drugsfda-0001-of-0001.json')**

**data = json.load(f)**

1. Separate out the Results from the Metadata using

**res = data['results']**

1. ‘res’ is a list of 24000+ rows which contains details of each and every approved drug
2. Each element in ‘res’ list contains multiple keys, such as 'submissions', 'application\_number', 'sponsor\_name'. The keys ‘products’ and ‘openfda’ may or may not be present.
3. To get a list of unique names of drugs present in this database, we are using the ‘products’ key
4. The element corresponding to ‘product’ key is a list of products, each contains the key ‘active\_ingredients’, which again corresponds to a list of active\_ingredients.
5. From this list of active ingredients each ‘name’ is accessed. If the ‘name’ is not already present in our list it is added.
6. The above process is repeated for each element in ‘res’ list.
7. The code for generating the unique list of active\_ingredients present in the openfda database is given below:

**namelist = list()**

**for result in res:**

**if('products' in result.keys()):**

**prod = result['products']**

**for product in prod:**

**acti = product['active\_ingredients']**

**for active in acti:**

**name = active['name']**

**if (name not in namelist):**

**namelist.append(name)**

1. For matching the names of drugs in our list and those present in Chemical Items file, a list of all unique Substance Names provided in the Chemical Items file needs to be prepared. This file can be prepared using the Chemical Items file downloaded from Medline Repository. The Chemical Items file can be uploaded in R environment and the unique substance names can be extracted using the R code provided below. The list of unique substance names contains 242672 elements. This list is then printed out in a csv file:

**library(data.table)**

**ci2019 <- fread(‘FilePath\\Chemical\_Items’)**

**colnames(ci2019) <- c('1','2','3','4')**

**subs<- unique(ci2019$'1')**

**subframe <- data.frame(subs)**

**write.csv(subframe,file='filepath\chemicalitemssubstancenames.csv')**

Load this csv file in python using the following code:

**import pandas as pd**

**subs =pd.read\_csv(r'FilePath\chemicalitemssubstancenames.csv', engine="python")**

**substances = subs['subs']**

**substances = substances.tolist()**

**substances = [x.lower() for x in substances]**

**namelist = [x.lower() for x in namelist]**

1. Match the name list extracted from openfda with the list of substances using full length matching and save in fullmatch list. Use the following code in Python

**fullmatch = list()**

**fullmatch = [name for name in namelist if name in substances]**

[Length of namelist = 2487, substances = 242672, fullmatch = 1161]

1. Create unique list of first names of all the names extracted from openfda using the following code in Python

**firstnames=list()**

**firstnames = [name.split(" ")[0] for name in namelist]**

**import pandas as pd**

**firstnames = pd.unique(firstnames).tolist()**

1. Match the list of first names with the names of chemical items substances

**firstnamematched = [name for name in firstnames if name in substances]**

1. Since there would be drugs present among fullmatch list, who would have also matched by firstnames, we need to identify those drugs which have matched only by full name and not first name. Use the following code in Python:

**allonly=list()**

**for full in fullmatch:**

**first = full.split(" ")[0]**

**if first not in firstnamematched:**

**allonly.append(full)**

1. Combine the names of drugs that have matched by firstname, and those matched by full names

**final = firstnamematched + allonly**

1. In the process of extracting unique list of first names from the name list, the salts of Calcium, Sodium, Zinc, Magnesium etc. Would have got removed. Hence to obtain names of these salts the following list can be used:

**elementlist =['calcium','magnesium','manganese','potassium','ammonium','sodium','zinc','barium','lanthanum','titanium','silver','lithium','gallium','arsenic','bismuth']**

This is the list of elements whose various salts are present in openfda namelist. To extract all the salts of these elements the following code needs to be executed in Python

**complist=list()**

**for comp in namelist:**

**for ele in elementlist:**

**if comp.startswith(ele):**

**complist.append(comp)**

1. The list of these salts which is present in Chemical Items substance list as well can be obtained using:

**compmatched = [comp for comp in complist if comp in substances]**

**final = final + compmatched**

1. The list of drugs can then be printed in excel file using:

**finallist = pd.unique(final).tolist()**

**finalframe = pd.DataFrame(finallist)**

**finalframe.to\_csv(r'filepath\matchedDrugsandComp.csv')**

1. The file obtained through the computational processing as mentioned above was manually processed as well, the following drugs were removed.

|  |
| --- |
| **Removed Drug names** |
| amino acids |
| ammonia |
| arsenic |
| barium |
| bismuth |
| carbon |
| copper |
| gallium |
| helium |
| hydrogen |
| indium |
| iron |
| lithium |
| magnesium |
| manganese |
| nitric oxide |
| nitrogen |
| nitrous oxide |
| oxygen |
| potassium |
| radium |
| rubidium |
| samarium |
| selenium |
| sodium |
| strontium |
| sulfur |
| talc |
| technetium |
| titanium |
| xenon |
| zinc |

**Section 2: Reference Drug cluster preparation based on ATC drug classification**

The following steps were performed for preparation of ATC drug clusters for Validation:

1. We identified the classes of drugs which commonly occurred in our clustering results and located them in ATC classification. 26 classes of drugs containing 1681 drugs were identified.

2. 736 drugs among them whose names matched with our drug list, were used further. Only full-length matching was used for this purpose.

3. Using the ATC codes, the drugs belonging to same groups/subgroups were considered as one cluster for validation of the clustered generated in in silico experiments

4. A description of each of these validation clusters, their drugs and ATC codes is provided in the **Supplementary Material.xlsx**

**Section 3: Misclassification Counting**

The grouping/clustering of the drugs achieved through the procedure described in Section 2 has been considered as ideal clustering (‘True’ class) and is used for detecting misclustering/misclassification in our in silico experiments.

The following cases were observed in our clustering results of the in silico experiments:

I) When all the drugs belonging to a ‘True’ class are clustered together in a cluster obtained from the in silico experiments, all of them are considered to be clustered correctly even if they are present together in a noise cluster with drugs of other classes. Such a pattern was observed for antifungal and contrast agents

II) When majority of drugs belonging to a ‘True’ class are present in a cluster and few drugs are present in other cluster(s), then the number of drugs which are not present with the majority are counted as misclassified. Such pattern was observed for antidiabetic and antiviral drugs

III) When drugs belonging to a ‘True’ class are almost equally distributed over more than one classes, for example, two clusters contain 40 percent each of those drugs, then the 80 percent which appears in equally distributed clusters are considered correctly classified and the remaining 20 percent which are present in other clusters are considered misclassified. Such pattern was observed for antibacterial and anticancer drugs

IV) When drugs belonging to a ‘True’ class are spread out sporadically over various classes, the cluster with maximum number of drugs is considered correctly clustered, and all other drugs from the ‘True’ class are counted as misclustered. In the case of K-means Clustering of in silico Experiment 1, Anti-parkinson drugs have demonstrated this pattern, where out of the 19 diuretic drugs present in our ATC-based gold standard, 7 are present together in Cluster#8 while the remaining 12 drugs are distributed across clusters 22,19 and 12, hence the 7 drugs which are present together are considered to be correctly clustered while the rest are considered as misclustered.

Misclassifications/misclustering counted from all the four types were then summed up and reported.

**Section 4: Cluster-wise summary of results obtained from various K-means Clustering Experiments**

The cluster-wise summary of the in silico experiments performed using k-means clustering algorithms is mentioned below. Experiment 1 refers to clustering performed with entire set of MeSH terms, while Experiment 2 refers to clustering performed using a refined set of MeSH terms.

**1) Experiment 1**

*Cluster 1:* Antibacterials/Antimicrobials

*Cluster 2:* Natural and synthetic steroids, retinoids used for treatment of skin infections and inflammatory conditions

*Cluster 3:* Vasoconstrictors used for hypertension, blood vessel relaxants of other nature (asthama, diuretic) were also present.

*Cluster 4:* anti-cancer drugs, and few other drugs used in conjunction with them

*Cluster 5:* Noise cluster

*Cluster 6:* Anti-cancer drugs, with few other drugs used for cancer as well, such as Altretamine

*Cluster 7:* Anti-cancer drugs

*Cluster 8:* medicines related to brain related to anxiety, seizures, migraine, parkinson’s, alzheimer’s etc.

*Cluster 9:* Antivirals and antiretrovirals

*Cluster 10:* antifungal

*Cluster 11:* antihistamines, antiallergens and asthma medications

*Cluster 12:* antidepressent, antipsychotic, antihistamines and antiemetic

*Cluster 13:* Overactive bladder, prostatic hyperplasia and antihypertensives

*Cluster 14:* Steroidal and non-steroidal analgesics, anesthetics

*Cluster 15:* antidiabetic and antihyperlipidemic

*Cluster 16:* Drugs affecting hypertension, diuretics etc.

*Cluster 17:* hormonal therapeutics

*Cluster 18:* antibiotics

*Cluster 19:* neuronal agonists/antagonists

*Cluster 20:* ocular medicines

*Cluster 21:* psychiatric medicines

*Cluster 22:* Noise cluster

*Cluster 23:* Anesthetics

*Cluster 24:* Anticoagulants

*Cluster 25:* Noise cluster

*Summary:* Three noise clusters contain 746 drugs. Apart from that most of the smaller clusters also contain drugs belonging to more than one class and the boundaries of clusters are diffuse.

**Experiment 2:**

*Cluster 1:* Noise cluster

*Cluster 2:* Steroids, used for skin, arthiritis etc.

*Cluster 3:* anti-hypertensives/anti-hypotensives, other drugs such as bronchodilators, ergotamine derivatives used for treatment of migraine etc.

*Cluster 4:* anti-cancer drugs

*Cluster 5:* antibiotic/antimicrobial

*Cluster 6:* anti-cancer

*Cluster 7:* Monoclonal antibody used for psoriarisis, immunosuppression, cancer etc. Also contains antivirals, immunosuppresants and few steroidal drugs

*Cluster 8:* steroidal and non-steroidal analgesics

*Cluster 9:* antivirals and antiretrovirals

*Cluster 10:* antifungal

*Cluster 11:* respiratory medicines

*Cluster 12:* Psychiatric medicines for depression, schizophrenia etc. Also contains antiemetics

*Cluster 13:* medicines for overactive bladder, and induce vasoconstriction. Nasal decongestants and medicines for prostatic hyperplasia also present

*Cluster 14:* non-steroidal analgesics and antiemetics

*Cluster 15:* antidiabetic

*Cluster 16:* Antihypertensives, diuretics and cardiovascular medicines

*Cluster 17:* hormonal therapeutics

*Cluster 18:* antibiotics

*Cluster 19:* psychiatric medicines used for depression, schizophrenia etc.

*Cluster 20:* ocular medicines for intraocular pressure etc., few anesthetics as well

*Cluster 21:* psychiatric medicines for depression, anxiety, seizures etc.

*Cluster 22:* Noise cluster

*Cluster 23:* Anesthetic

*Cluster 24:* anti-coagulants

*Summary:* Noise clusters contain 688 drugs. Smaller clusters contain drugs of less diverse classes, however in few of the small clusters contain multiple clusters of drugs as well.

**Section 5: Cluster-wise summary of results obtained from various Hierarchical Clustering Experiments**

The cluster-wise summary of the in silico experiments performed using hierarchical clustering algorithms is mentioned below. Experiment 1 refers to clustering performed with entire set of MeSH terms, while Experiment 2 refers to clustering performed using a refined set of MeSH terms.

**Experiment 1**

*Cluster 1:* Antimalarial

*Cluster 2:* Neuronal agonists of different types such as anesthetics, antipsychotics etc.

*Cluster 3:* Noise cluster

*Cluster 4:* Neuronal agonists of different types, mostly antihypertensives, but also contains gastrointestinal medicines and anesthetics as well

*Cluster 5:* Contains medicines which act near the brain. Contains psychiatric medicines for depression and schizophrenia. Also contains medicines used for Alzheimer’s and Parkinson’s. It contains few other medicines related to these diseases as well which are not delivered to the brain.

*Cluster 6:* Steroidal and non-steroidal analgesics, few anti-emetics are also present

*Cluster 7:* biomolecules and simple inorganic salts. Chemical dyes and antimicrobials also present

*Cluster 8:* hypertension and arrhythmia

*Cluster 9:* anticancer drugs

*Cluster 10:* anticancer drugs

*Cluster 11:* hormonal therapeutics

*Cluster 12:* psychiatric medicines

*Cluster 13:* respiratory medicines and antihistamines

*Cluster 14:* Anesthetics

*Cluster 15:* hypertension and arrhythmia

*Cluster 16:* psychiatric medicines

*Cluster 17:* arrhythmias

*Cluster 18:* antidepressants

*Cluster 19:* antibiotic

*Cluster 20:* anticancer drugs

*Cluster 21:* psychiat~~e~~ric drugs

*Cluster 22:* antihyperlipidemic

*Cluster 23:* gastrointestinal disease

*Cluster 24:* Antibiotic

*Cluster 25:* Antihistamine

*Cluster 26:* non-steroidal analgesics

*Cluster 27:* anti coagulants

*Cluster 28:* antibiotic

*Cluster 29:* dermatological

*Cluster 30:* Ocular medicines

*Cluster 31:* Immunosupressant

*Cluster 32:* Osteoporosis

*Cluster 33:* diuretic

*Cluster 34:* antivirals

*Cluster 35:* antifungal

*Cluster 36:* diabetes

*Cluster 37:* radiocontrast agent

*Cluster 38:* overactive bladder and benign prostatic hyperplasia

*Cluster 39:* cancer

*Cluster 40:* gastroesophageal reflux disease

*Cluster 41:* antibiotic

*Cluster 42:* antivirals

*Cluster 43:* antifungal

*Cluster 44:* migraine

*Cluster 45:* small cluster (8 drugs) of unrelated drugs

*Summary:* Contains drug clusters of limited sizes only, one very large cluster present, however. The small clusters contain drugs which are closely related to each other. The large cluster contains unrelated drugs, one small cluster (number 45 also contains unrelated drugs). Otherwise, the clusters show high resolution in terms of being present with similar drugs.

**Experiment 2**

*Cluster 1:* Antimalarials

*Cluster 2:* Anesthetics and Analgesics

*Cluster 3:* arrhythmia, angina and hypertension (cardiovascular drugs)

*Cluster 4:* psychiatric medicines

*Cluster 5:* noise cluster

*Cluster 6:* Nervous system drugs

*Cluster 7:* biomolecules and simple inorganic salts. Chemical dyes and antimicrobials also present

*Cluster 8:* anticancer

*Cluster 9:* immunosuppressant

*Cluster 10:* psychiatric medicines

*Cluster 11:* respiratory medicines and antihistamines

*Cluster 12:* anesthetics

*Cluster 13:* hypertension

*Cluster 14:* antidiabetic

*Cluster 15:* Psychiatric medicines

*Cluster 16:* arrhythmia, angina and hypertension (cardiovascular drugs)

*Cluster 17:* hormonal therapeutics

*Cluster 18:* Psychiatric medicines

*Cluster 19:* antibiotics

*Cluster 20:* anticancer

*Cluster 21:* Psychiatric medicines

*Cluster 22:* antiemetic

*Cluster 23:* antihyperlipidemic

*Cluster 24:* antiallergic medicines

*Cluster 25:* antibiotic

*Cluster 26:* analgesics

*Cluster 27:* anticoagulants

*Cluster 28:* antibiotic

*Cluster 29:* Opthalmologic medicines

*Cluster 30:* Osteoporosis

*Cluster 31:* diuretic

*Cluster 32:* antifungal

*Cluster 33:* radiocontrast agents

*Cluster 34:* overactive bladder and benign prostatic hyperplasia

*Cluster 35:* anticancer drug

*Cluster 36:* antivirals

*Cluster 37:* gastrointestinal disorders

*Cluster 38:* Dermatological agents

*Cluster 39:* antibiotic

*Cluster 40:* antivirals

*Cluster 41:* antifungal

*Cluster 42:* migraine

*Cluster 43:* small cluster (8 drugs) of unrelated drugs

*Summary:* Most of the clusters are small and contain uniform type of drugs. Only one very large cluster of 384 drugs is present, rest of the clusters have limited sizes only. Cluster number 43 is a small cluster of unrelated drugs.