Investigating the Automation of Mapping the British National Formulary Codes into Anatomical Therapeutic Chemical Codes

A CAPSTONE PROJECT

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Joshua Butler, M.D.

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF HEALTH INFORMATICS

Adviser: Terrance Adam, M.D., Ph.D.

October 2019

©2019 Joshua Butler, M.D.

**Table of Contents**

[**List of Figures** i](#_Toc23328587)

[**List of Tables** ii](#_Toc23328588)

[**List of Abbreviations** iii](#_Toc23328589)

[**Introduction** 1](#_Toc23328590)

[**Background: Drug Classification Systems** 2](#_Toc23328591)

[ATC 2](#_Toc23328592)

[BNF 4](#_Toc23328593)

[RxNorm 5](#_Toc23328594)

[SNOMED-CT 7](#_Toc23328595)

[Previous Works in Classification Mapping 8](#_Toc23328596)

[Machine Learning Approaches to Mapping 9](#_Toc23328597)

[**Data and Material** 10](#_Toc23328598)

[**Methods** 11](#_Toc23328599)

[Method A 12](#_Toc23328600)

[Method B 12](#_Toc23328601)

[Method C 12](#_Toc23328602)

[**Results** 13](#_Toc23328603)

[Method A 13](#_Toc23328604)

[Method B 16](#_Toc23328605)

[Method C 17](#_Toc23328606)

[Full BNF File Mapping 17](#_Toc23328607)

[**Discussion** 18](#_Toc23328608)

[Significance 18](#_Toc23328609)

[Limitations 18](#_Toc23328610)

[Future Work 19](#_Toc23328611)

[**Bibliography** 20](#_Toc23328612)

# **List of Figures**

[Figure 1: A classification breakdown of BNF code 5](#_Toc23329595)

[Figure 2: The benefit of using SNOMED-CT 7](#_Toc23329596)

[Figure 3: The design of SNOMED-CT ontological foundation 8](#_Toc23329597)

[Figure 4: Machine Learning Technology Trend from 1990 to 2016 9](#_Toc23329598)

[Figure 5: A partial screenshot showing an output of the fuzzy match (Method A) using ratio scoring algorithm 13](#_Toc23329599)

[Figure 6: A screenshot showing the table of tp, fn, fp, tn tallies to measure the scoring performance 14](#_Toc23329600)

[Figure 7: A screenshot showing an output of Method B Script 16](#_Toc23329601)

[Figure 8: Method B Outcome 16](#_Toc23329602)

# **List of Tables**

[Table 1: Levels of ATC Classification 2](#_Toc23329632)

[Table 2: First level of ATC classification 3](#_Toc23329633)

[Table 3: ATC code breakdown for metformin 3](#_Toc23329634)

[Table 4: Multiple ATC codes for prednisolone 4](#_Toc23329635)

[Table 5: Example of BNF Chapters 5](#_Toc23329636)

[Table 6: A collection of source terminologies for RxNorm 6](#_Toc23329637)

[Table 7: Number of not mapped and mapped entries. 14](#_Toc23329638)

[Table 8: Performance scores for each scoring algorithm, from the best to worse (F1 Score) 15](#_Toc23329639)

[Table 9: Overall number of matches with missed opportunities (could match but did not score well) 15](#_Toc23329640)

[Table 10: Performance of combining Method B and A 17](#_Toc23329641)

[Table 11: Performance result of full BNF file mapping 17](#_Toc23329642)

# **List of Abbreviations**

|  |  |
| --- | --- |
| Abbreviation | Definition |
| ATC | Atomic Therapeutic Chemical |
| BNF | British National Formularies (U.K.) |
| CSV | Comma Separated Values |
| DDD | Defined Daily Dosage |
| NDC | National Drug Code (U.S.) |
| NHS | National Health Services (U.K.) |
| NLM | National Library of Medicine |
| NLP | Natural Language Processing |
| SNOMED | Systematized Nomenclature For Medicine |
| SQL | Structured Query Language |
| UMLS | Unified Medical Language System |
| WHO | World Health Organization |

# **Introduction**

The growth in international prescription usage and associated research creates information comparison problems for prescribing patterns between two countries due to different drug code systems and lack of common drug data terminologies. This is essentially an informatics category mapping problem, where a one-to-one correspondence needs to be established between different drug code systems used to specify medications in administrative data. For example, the UK uses the British National Formulary convention, while the United States uses the RxNorm and SNOMED systems. One approach is to map drug codes to a standardized system (e.g., the Anatomical Therapeutic Chemical Classification) to facilitate research comparing prescription drug use between two countries.

For virtually all medications, such mapping is non-trivial. Internationally, there are a plethora of differences in drug nomenclatures in use within each country (e.g., naming the actual chemical compound), packing sizes, pricing, and dosage that make it harder for researchers to precisely compare drug use between different entities. Further, there are common differences in the medications approved in one country but not in another (i.e., formulary differences). In addition, slight variations in the compound formulation that have minimal therapeutic differences (e.g., metoprolol tartrate versus metoprolol succinate), can make matching problematic.

The ability to compare data from different parts of the world is vital because having these data would be useful for research in prescribing patterns, pharmaceutical economic studies, drug utilization, healthcare decision making, and establishing pricing and reimbursement amount. In this paper, we address this problem by applying machine learning methods to map medications between the US Medicare System prescribing data and the UK British National Formulary system.

A problem became apparent while working on a project studying prescription prescribing patterns in the UK utilizing the data supplied by National Health Services (NHS). The supplied data uses British National Formulary (BNF) codes, to which each unique drug has a unique code assigned based on the drug properties and chemical name. As part of that project, we wanted to compare their prescribing patterns with the US; there were no common identifiers between two systems. At this time, there is no data source that has the BNF codes mapped to Anatomical Therapeutic Chemical (ATC) codes. The purpose of this capstone project is to develop a script to automate the mapping process as a proof of concept and identify a methodology that produces the best yield.

# **Background: Drug Classification Systems**

This project investigates how we can leverage the data available from BNF, SNOMED, RxNorm, and ATC data sets to aid in the mapping from one classification system into another. The following classification systems listed below describe a brief history and description of each drug classification system:

## ATC

The Anatomical Therapeutic Chemical (ATC) classification was developed in the 1970s in Norway and endorsed by the World Health Organization (WHO) as a gold standard. Before the ATC classification came into fruition, the committee at the WHO Symposium in 1969 determined that there was a significant need for the internationally accepted classification system. The development of a classification system began in the early 1970s. In 1981, the classification system was formally recognized for drug utilization studies and recommended for use in Europe. Over a few decades, they have continued to refine the classification system and eventually introduced Defined Daily Doses (DDD) to standardize the drug consumption units in drug usage studies. Finally, in 1996, the WHO has stated that there is a need to strengthen the system as an international standard.

ATC is a classification system that is organized by the active chemical substances according to the organ or system upon which they act according to the therapeutic, pharmacologic, and chemical properties. The sole purpose of this classification system is to simplify the comparison of the drug use statistics across different international, national, and regional levels despite the difference in nomenclature, packing sizes, pricing, and dosages.

*Table 1* below shows different levels of classifications1 used by ATC:

Table : Levels of ATC Classification

|  |  |
| --- | --- |
| ATC Level | Description |
| 1st level | Main anatomical or pharmacological groups |
| 2nd level | Pharmacological or therapeutic subgroup |
| 3rd & 4th level | Chemical, pharmacological or therapeutic subgroup |
| 5th level | Chemical substance (generic name - nonproprietary name) |

The first level of ATC classification denotes under which main anatomical or pharmacological group the drug falls.  *Table 2* shows fourteen different first level main classification groups plus one experimental classification:

Table : First level of ATC classification

|  |  |
| --- | --- |
| 1st Level | Anatomical or Pharmacological Group |
| A | Alimentary tract and metabolism |
| B | Blood and blood-forming organs |
| C | Cardiovascular system |
| D | Dermatologicals |
| G | Genitourinary system and sex hormones |
| H | Systemic hormonal preparations, excluding sex hormones and insulins |
| J | Anti-infective for systemic use |
| L | Anti-neoplastic and immunomodulating agents |
| M | Musculoskeletal system |
| N | Nervous system |
| P | Anti-parasitic products, insecticides and repellents |
| R | Respiratory system |
| S | Sensory organs |
| V | Various |
| X | Experimental |

The nomenclature of chemical substances follows the International nonproprietary name (INN). If a chemical substance is not assigned with an INN, an alternative name will be adopted from either USAN (United States Adopted Name) or BAN (British Approved Name). An example ATC code for metformin (a diabetes medication), is shown in *Table 3*:

Table : ATC code breakdown for metformin1

|  |  |
| --- | --- |
| ATC Code | Description |
| A | Alimentary tract and metabolism |
| A10 | Drugs used in diabetes |
| A10B | Blood glucose lowering drugs, excl. Insulins |
| A10BA | Biguanides |
| A10BA02 | Metformin |

Only one unique ATC code is given for each route of administration, or multiple ATC codes are given for the same chemical substance with different therapeutic uses. For example, prednisolone has several ATC codes (see *Table 4*) due to different formulation or therapeutic uses2.

Table : Multiple ATC codes for prednisolone2

|  |  |
| --- | --- |
| ATC code | Description |
| A07EA01 | Intestinal anti-inflammatory agents (enemas and foams) |
| C05AA04 | Anti-hemorrhoidal for topical use (suppositories) |
| D07AA03 | Dermatological preparations (creams, ointments and lotions) |
| H02AB06 | Corticosteroids for systemic use (tablets, injections) |
| R01AD02 | Nasal decongestants (nasal sprays/drops) |
| S01BA04 | Ophthalmological (eye drops) |
| S02BA03 | Otological (ear drops) |

## BNF

The British National War Formulary was established in 1939 during World War II. There were 380 preparations listed in the formulary under the apothecaries’ system of minims and grains. The birth of the British National Formulary (BNF) occurred after WWII ended. Two non-governmental bodies joined forces to develop a new formulary, which was first published in 1949 and continues to print for years until a significant overhaul began in 19763. During the 1960s, BNF’s old apothecaries system was replaced with the metric system. A new edition was printed every three years until the 1970s when the committee found that the formulary was not being used much by physicians and pharmacists because the information was outdated. During that time, the committee consisted of experts (physicians and pharmacists) who spent most of their time deciding which drugs and preparations to include in the formulary and was not very productive4. From that point, NHS finally recognized the issues they were having and decided to form a new joint formulary committee in 1978. The committee went on to develop a new national formulary to address new changes and began printing a new copy of BNF in February 19814. It is still in use today, and a new edition is published every six months.

The current BNF code used by the National Health Services (NHS) is depicted in *Figure 1,* which illustrates the drug classification hierarchy. The top-level starts from left; each box represents each level and goes down as you move to the right. The top level of the ontology is broken down into chapters, which defines physiological or organ system. A practical example, as excerpted from a website5, shown in *Figure 1*, Tradorec is a pain medication, classified as an opioid. The first nine characters represent the chemical substance of a drug, while the last six characters relate to the brand name and the dosage.

|  |  |
| --- | --- |
|  |  |

Figure : A classification breakdown of BNF code

*Table 5* shows a list of top-level BNF chapters, represented by a different physiological system. The rationale of this system is to provide the researcher with a quick glimpse at the code, and would immediately know under which classification the drug is classified and be able to cluster the drugs together and stratify them for statistical analysis.

Table : Example of BNF Chapters

|  |  |  |  |
| --- | --- | --- | --- |
| Chapter | Description | Chapter | Description |
| 01 | Gastro-Intestinal System | 12 | Ear, Nose And Oropharynx |
| 02 | Cardiovascular System | 13 | Skin |
| 03 | Respiratory System | 14 | Immunological Products & Vaccines |
| 04 | Central Nervous System | 15 | Anaesthesia |
| 05 | Infections | 18 | Preparations used in Diagnosis |
| 06 | Endocrine System | 19 | Other Drugs And Preparations |
| 07 | Obstetrics,Gynae+Urinary Tract Disorders | 20 | Dressings |
| 08 | Malignant Disease & Immunosuppression | 21 | Appliances |
| 09 | Nutrition And Blood | 22 | Incontinence Appliances |
| 10 | Musculoskeletal & Joint Diseases | 23 | Stoma Appliances |
| 11 | Eye |  |  |

## RxNorm

U.S. National Library of Medicine (NLM) created a different model nomenclature describing clinical drugs in the Unified Medical Language System in late 2001 because they suspected that there is a lack of synonymy, which became problematic and traditional methodologies of identifying missing synonymy was not successful.6 This led to the creation of RxNorm, which improved the interoperability between different classification vocabulary and semantics, terminology methodologies gave a clearer representation of different drug concepts, drug ingredients, or dosage forms. The goal of RxNorm is to provide a reliable and consistent way to identify the essential components of a drug – medication name, dosage, route of administration, and ingredients.7

RxNorm is “a normalized naming system for generic and branded drugs, and a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems.”8 This tool provides key stakeholders with the ability to process prescription information, and because there are so many systems with different drug names, it would become difficult to communicate from one system to another. With the normalization of a system, it would make it easier to transit from one system to another, and it would behoove users to adopt the standard terminology.

RxNorm is built upon what is already publicly available using various drug terminologies from several pharmacy management software and drug interaction software packages. RxNorm contains terminologies derived from different terminology sources8, as depicted in *Table 6*. RxNorm attempts to maintain its data by preserving the drug names, attributes, and relationships from various sources. If a drug name does not appear in the database, then it did not receive the information from any of the data sources.

Table : A collection of source terminologies for RxNorm

|  |  |
| --- | --- |
| Source | Source Abbreviation (SAB) |
| Anatomical Therapeutic Chemical Classification System | ATC |
| Vaccines Administered | CVX |
| DrugBank | DRUGBANK |
| Gold Standard Drug Database | GS |
| Medi-Span Master Drug Data Base | MDDB |
| Multum MediSource Lexicon | MMSL |
| Micromedex RED BOOK | MMX |
| Medical Subject Headings (MeSH) | MSH |
| CMS Formulary Reference File | MTHCMSFRF |
| FDA Structured Product Labels | MTHSPL |
| FDB MedKnowledge (formerly NDDF Plus) | NDDF |
| RxNorm Normalized Names and Codes | RXNORM |
| US Edition of SNOMED CT (drug information) | SNOMEDCT\_US |
| USP Compendial Nomenclature | USP |
| Veterans Health Administration National Drug File | VANDF |

## SNOMED-CT

SNOMED began in 1965, initially called Systematized Nomenclature of Pathology (SNOP), was developed by the College of American Pathologists.9 SNOMED-CT is a restructured merger between SNOMED RT (Reference Terminology) and United Kingdom’s National Health Service Clinical Terms (CT). SNOMED-CT is a “comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting.”10 It has been designated as a US standard by the US federal government for interoperability between systems to facilitate health information exchange. It covers several different domains, including but not limited to: clinical findings, disorders, procedures, health-related activities, and observable entities. The goal of SNOMED-CT is to provide comprehensive, high-quality health information in health records that benefit different stakeholders *(Figure 2)* by tracking clinical assessment and treatment, population monitoring, and research to increase clinical knowledge.

The National Library of Medicine (NLM) entered an agreement with the College of Pathologists to make SNOMED-CT available for free through UMLS in July 2003. The SNOMED-CT concept id does not have a specific coding pattern like BNF, and ATC does but provides relationship links between different attributes. It is broken down into concepts map and hierarchies, as depicted in *Figure 3*.

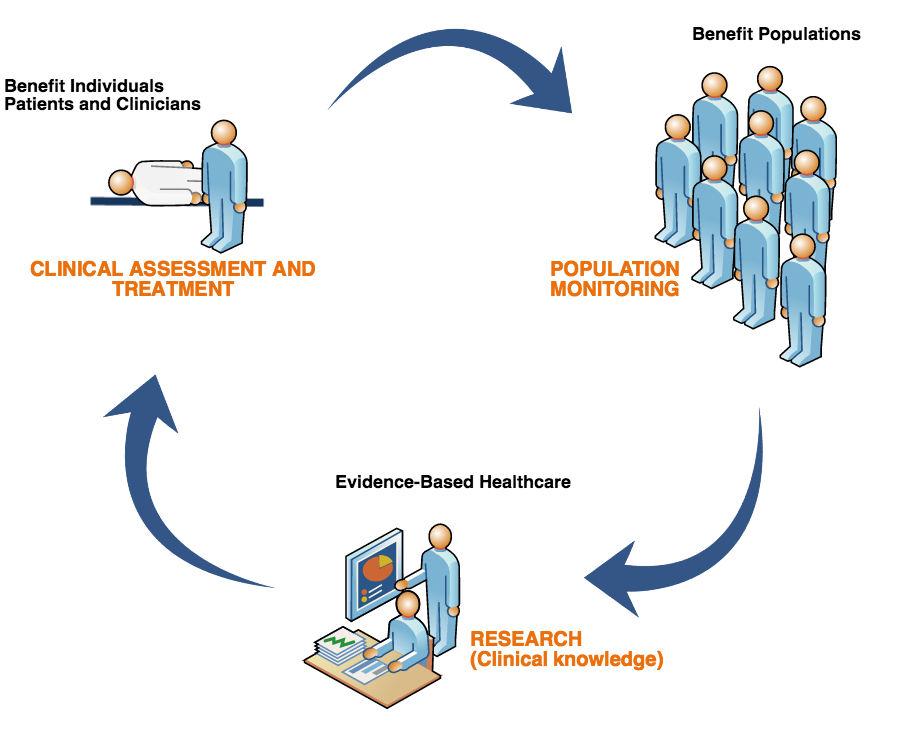


Figure : The benefit of using SNOMED-CT11

Previously, general practices in the UK used two different versions of clinical codes (Read v2 and CTV3). Not all use the same system. NHS had decided that they needed a single clinical terminology so that the clinical data exchange is consistent and accurate across all settings – allowing better patient care and improved reporting of clinical data analysis12. NHS had mandated that all general practices across the nation to begin implementing SNOMED-CT in their practices beginning April 2018 and through the next 18 months; which were deployed in two phases: (1) behind the scene changes and (2) the changes that impact the end-users (User Interface)13.

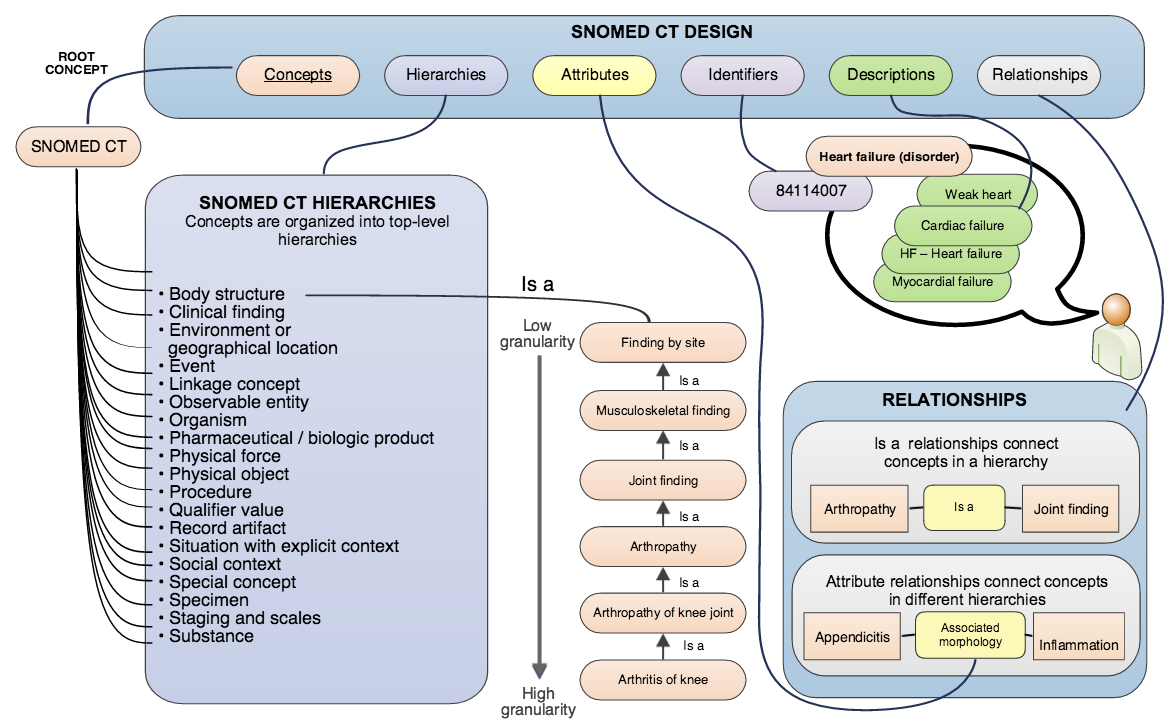


Figure : The design of SNOMED-CT ontological foundation14

While SNOMED-CT provides codes to many different concepts, it also provides codes for medications along with multiple relationships between different concepts. SNOMED-CT will be used in this project as one of the methodologies to map the BNF codes to ATC codes directly by cross-referencing with the RxNorm database.

## Previous Works in Classification Mapping

There are not many peer-reviewed publications that discuss mapping from one drug code system to another system. Below are papers from the AMIA symposium that discussed mapping and issues with the automation of the mapping.

An AMIA symposium paper by Dr. Oliver Bodenreider and Dr. Laritza Taft previously evaluated the suitability of the ATC/DDD index for prescription list analysis in the US15. They mapped the clinical drugs to ATC by using the RxNorm database through ingredients and route of administration. They found that only 68% of their data was mapped to ATC, and out of that subset, they only mapped 44% of entries to the ATC 5th-level codes for single-ingredient drugs. They concluded that the mapping of RxNorm ingredients to ATC was incomplete due to missing data and ambiguities.

An AMIA symposium poster by Dr. Fabricio Kury in 2017 demonstrated that the mapping of NDC codes to ATC codes using an R script to make API queries to RxNorm to obtain ATC 4th-level code16. They found that out of the “Medicare dataset, only 77.9% of the NDCs were mapped to at least one ATC class.” They concluded that their approach to utilizing RxNorm is a scalable and straightforward solution and was concerned about the ambiguity of some mappings that ATC has multiple classes for the same ingredient. This particular work does not address the drug mapping to ATC 5th-level codes.

## Machine Learning Approaches to Mapping

In 1959, Arthur Samuel, a pioneer in the field of computer gaming and artificial intelligence, coined the term “machine learning “ and the term was expanded by Tom Mitchell, “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.*”17 Machine learning is one of the fastest-growing fields in data science (*Figure 4*).18 It leverages the high-speed computer to analyze big data and classify the dataset through unsupervised, supervised, and reinforcement learning algorithms.

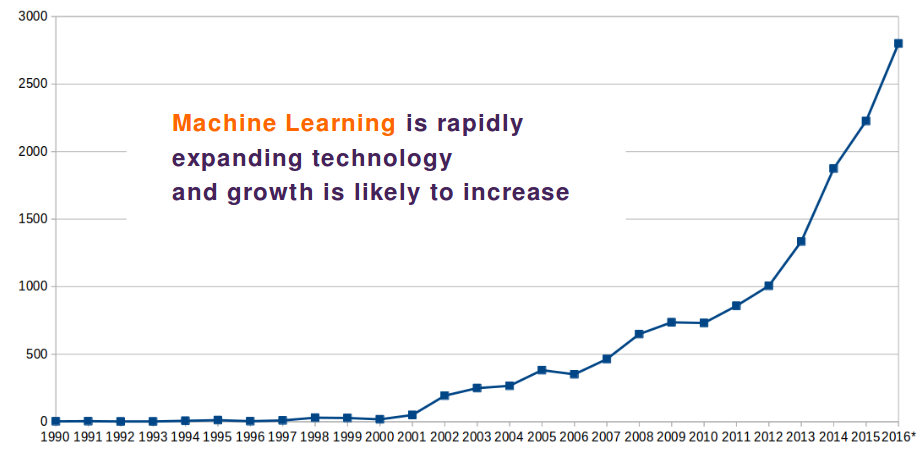


Figure : Machine Learning Technology Trend from 1990 to 2016 (Source: USPTO, Teqmine Analytics Ltd.)

This capstone project focuses on leveraging the natural language processing (NLP) models, a subset of the machine learning field. NLP is a field of study on how to program computers to process and analyze a significant amount of natural language data.19 A fuzzy string match is one of the NLP models we will use for the mapping project. The model uses Levenshtein distance,20,21 an algorithm named after a Soviet mathematician, Vladimir Levenshtein, in 1965. It is a mathematical metric to measure the difference between two strings of interest and returns a score of how well the strings match.

Below is an example from a Wikipedia article how the algorithm works:20

*The Levenshtein distance between "kitten" and "sitting" is three, since the following three edits change one into the other, and there is no way to do it with fewer than three edits:*

***k****itten →* ***s****itten (substitution of "s" for "k")*

*sitt****e****n → sitt****i****n (substitution of "i" for "e")*

*sitti****n*** *→ sittin****g*** *(insertion of "g" at the end)*.

To calculate the similarity score, the formula below gives you the percentage of how similar two strings are to each other:

**similarity = (1 – (Levenshtein\_distance / length\_of\_the\_longest\_string)) \* 100**

As a result of the equation above, the similarity percentage for the example strings above is calculated with a Levenshtein distance of three and the length of the longest string is seven:

(1 – (3/7))\*100 = 57.2%

The above example is just one of many ways of how it would calculate the similarity score. The method section of the paper will describe in detail how it will use the Levenshtein distance and different scoring models it has. Providing technical details of different scoring models is beyond the scope of this paper.

# **Data and Material**

This project utilizes several different data sets that are publicly available for download. Most sites are publicly accessible except for the RxNorm database, which requires a UMLS account to download and install on a local MySQL database. Although there is a public website for RxNorm, this project requires automation of data queries, and their public API endpoint does not provide ATC codes. Anyone can look up the SNOMED-CT code on RxNorm to get the chemical substance and then look up all associated codes like NDC, SNOMED-CT, ATC, and other standardized codes. It is a very intense and tedious process to do so manually.

The following data sources were used in this project:

1. BNF Code Information CSV File22 from the NHS website
2. RxNorm Files23 from UMLS download page (UMLS account required)
3. BNF SNOMED Mapping24

Instructions to get these files and replicating the project results are found on the project GitHub site (<https://github.com/jbutlermd/bnf-atc-mapping>) containing the source code of the scripts and some data files.

# **Methods**

Considering the data size of BNF codes, over ~83,000 entries are used in the UK to track prescriptions written by physicians and prescriptions dispensed by pharmacies. It is tedious to map those codes to ATC codes manually and would require a specialist with a doctorate in pharmacy to correctly classify some drugs because there is variation in spellings or chemical names but are identical in properties. One potential approach is to automate the mapping process if a script is well designed and would not take too much time to run to completion. For this project, Python programming language was chosen, because it is the fastest-growing in popularity, and its versatile libraries are available to enrich the programming code and speed implementation25. It is also one of the most commonly used programming languages in the data analysis and data science, as well as machine learning over R.

The purpose of this paper is to investigate the best way to automate mapping; thus, three different methods are proposed to evaluate which method would produce the best yield with the most accurate matches. Each method will have a statistical analysis to determine the performance of the mapping. In order to analyze the performance of a fuzzy match, a confusion matrix will be built. The data set used for this work contains over ~55,000 entries to manually determine whether the match is true positive, true negative, false positive, or false negative. Due to the size of the full BNF file containing >55,000 entries, it is impractical to create a confusion matrix manually with this many entries. Therefore, for practical consideration, it would make the most sense to simplify the process by using only 0.25% of all data (n=138). The drug entries were randomly selected for a test file for the performance analysis to determine the best scorer and method that has the highest yield.

Each method has a python script written to automate all of the steps outlined in each method, and the study can be reproduced easily. The script for Method A (***method-a.py***) generates an Excel spreadsheet file with each tab created for each scorer system and a summary tab containing statistical calculations, including F1 scores. And the rest of the resulting output, a summary tab is created manually.

## Method A

A python script (***method-a.py***) was written using one of python’s libraries called **fuzzywuzzy**, which handles the fuzzy match using the Levenshtein algorithm, described above, for scoring. The library has six different scoring systems: Ratio, Partial Ratio, Token Set Ratio, Partial Token Set Ratio, Token Sort Ratio, and Partial Token Sort Ratio. The details of each scoring method are explained in an article entitled, “Natural Language Processing for Fuzzy String Matching with Python” by Susan Li26 on the *Towards Data Science* website.

The script will peruse all six different scoring systems; each output is stored in an individual tab along with matches and their scores. After executing the script, the confusion matrix is entered manually for each scorer tab. As you update it, the statistics in the summary tab are updated automatically.

## Method B

This method is very straightforward; it cross-references the SNOMED-CT code to ATC code through the RxNorm database. This method uses two data sources, and they are required in order to execute this method. A full RxNorm database would need to be installed on a MySQL server.23 The script (***method-b.py***) will make four separate SQL queries to map the code for each SNOMED-CT code correctly. NHS supplies a data source containing BNF code mapped to SNOMED-CT codes, which is publicly available for download.24

## Method C

A combined method of both methods A and B. The first step, Method B will be applied, and the second step will map the remaining unmatched entries using Method A. To perform a fuzzy match in this method, we select the best fuzzy match scoring algorithm and use that algorithm in the second step. To determine the best scoring algorithm, we choose the algorithm with the highest F1 score from Method A test cases. The goal of this method is to achieve a higher yield than method A or B.

# **Results**

Each method is run individually, and the time to execute on a smaller sample (n=138) only took less than one minute for each scorer (Method A), less than one minute for Method B, less than two minutes for Method C, and approximately six minutes for the full file mapping (n=55,706) using Method C with the best scorer from Method A. For the final script, full file mapping, it utilizes caching to store previously matched drug name to speed up the mapping process of a large file. If caching was not used, it would take longer than one hour to finish.

## Method A

The script ***method-a.py*** was executed and produced the results, as seen in *Figure 5*. Within the Excel file, it contained seven different worksheets – six different scoring algorithms and a summary worksheet with statistical calculations to measure the performance of the fuzzy match. The ‘*matches’* column shows the scoring result (drug name, match score, and internal id number for cross-reference). The ‘*atc\_best\_match’* column lists the best-matched ATC code with a score > 90.

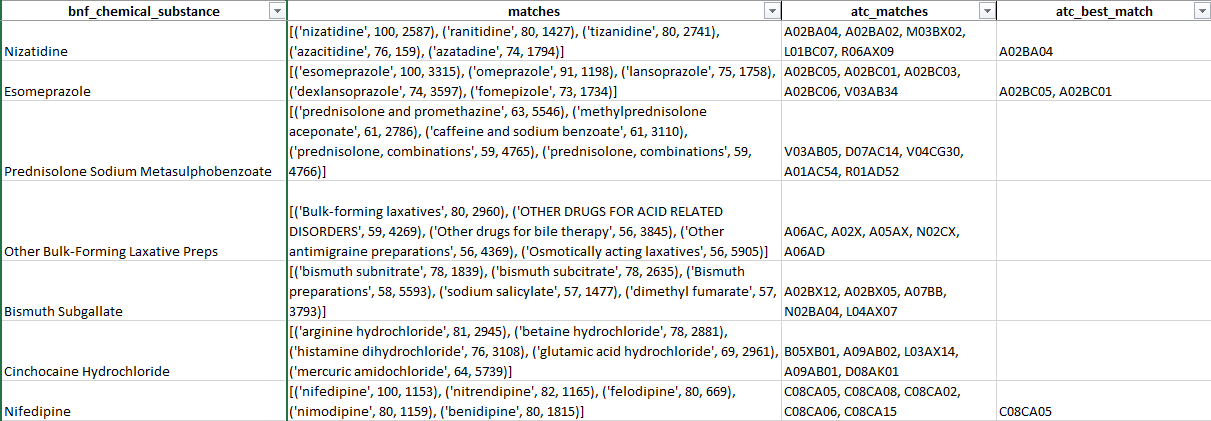


Figure : A partial screenshot showing an output of the fuzzy match (Method A) using ratio scoring algorithm

The next step was to evaluate the matches manually to measure how accurate the fuzzy match algorithm matched the correct ATC codes by tallying the number of true positives, false positives, true negatives, and false negatives *(Figure 6)*.

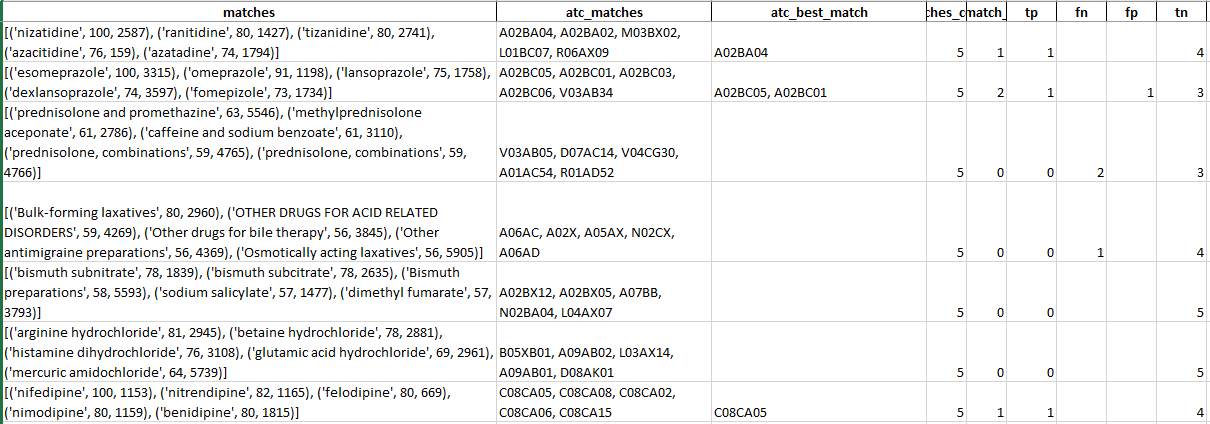


Figure : A screenshot showing the table of tp, fn, fp, tn tallies to measure the scoring performance

The tables below show the fuzzy match performance of each scorer. In *Table 7*, the “Not Mapped” and “Mapped” column indicates the number of drug entries that have no matches which were not mapped and have suitable matches, respectively. The “number of matches” column shows the actual number of suitable matches with a score >90, so one drug entry may have more than one match due to having multiple ATC codes for one drug (note multiple matches in *Figure 5*). The confusion matrix is a sum of respective columns from the results tab.

Table : Number of not mapped and mapped entries (Nbr of Match indicates a number of ATC codes Mapped).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Scorer | Not Mapped | Mapped | Nbr of Matches | TP | FN | FP | TN |
| Partial Ratio | 21 | 117 | 234 | 137 | 2 | 97 | 454 |
| Partial Token Set Ratio | 5 | 133 | 433 | 112 | 1 | 321 | 256 |
| Partial Token Sort Ratio | 21 | 117 | 230 | 129 | 5 | 101 | 455 |
| Ratio | 72 | 66 | 80 | 79 | 36 | 1 | 574 |
| Token Set Ratio | 28 | 110 | 195 | 135 | 2 | 60 | 493 |
| Token Sort Ratio | 71 | 67 | 82 | 80 | 23 | 2 | 585 |

*Note: TP=True Positive, FN=False Negative, FP=False Positive, TN=True Negative*

*Table 8* demonstrates scorer performance based on the confusion matrix in *Table 7*. Three scorers with a high F1 Score (>0.800) in the decreasing order: Token Sort Ratio, Token Set Ratio, and Ratio. F1 score is a harmonic mean between precision and recall (also known as sensitivity), whereas the perfect precision and recall approach to 1. Because Token Sort Ratio has the highest F1 score, this scorer will be used in Method C.

Table : Performance scores for each scoring algorithm, from the best to worse (F1 Score)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Scorer | Sensitivity | Specificity | Precision | Accuracy | F1 Score |
| Token Sort Ratio | 0.777 | 0.997 | 0.976 | **0.964** | **0.865** |
| Token Set Ratio | 0.985 | 0.892 | 0.692 | 0.910 | 0.813 |
| Ratio | 0.687 | **0.998** | **0.988** | 0.946 | 0.810 |
| Partial Ratio | 0.986 | 0.824 | 0.585 | 0.857 | 0.735 |
| Partial Token Sort Ratio | 0.963 | 0.818 | 0.561 | 0.846 | 0.709 |
| Partial Token Set Ratio | **0.991** | 0.444 | 0.259 | 0.533 | 0.410 |

*Table 9* shows the number of the overall successful match (by entries), whereas “Missed Opportunity” indicates the percentage of data that could have been mapped but was marked mismatch (scored < 90). To illustrate a missed opportunity (*Figure 5*), *Prednisolone Sodium Metasulphobenzoate* had no matches, but on a closer look, there are two matches for “prednisolone, combinations.” That particular entry was marked with two false negatives because they were correct matches, but it scored too low to meet the threshold to accept it as a good match.

Table : Overall number of matches with missed opportunities (could match but did not score well)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scorer | Overall Match | Missed Opportunity | Gain if not missed | Net Gain |
| Token Set Ratio | 71.7% | 7.1% | 73.2% | 1.4% |
| Partial Ratio | 71.0% | 9.5% | 72.5% | 1.4% |
| Partial Token Sort Ratio | 69.6% | 23.8% | 73.2% | 3.6% |
| Partial Token Set Ratio | 61.6% | 20.0% | 62.3% | 0.7% |
| Token Sort Ratio | 48.6% | 29.6% | 63.8% | 15.2% |
| Ratio | 47.8% | 38.9% | 68.1% | 20.3% |

Another issue to note from fuzzy matching, the algorithm showed the drug esomeprazole two matches meeting the threshold: esomeprazole (100) and omeprazole (91). Omeprazole was marked as a good match, but in reality, it is a false positive because of similar spellings and partial matches. Additionally, the algorithm does not recognize drug combinations and does not correctly match the correct ATC code for combination drugs.

Token Sort Ratio scorer yielded 48.6% of the match with fewest false positives. Although the Partial Ratio and Token Set Ratio scorer yielded higher but generated many false positives, and it is not practical. It would not be ideal to use those two because if the script is run on a full BNF file containing >50,000 entries, it would be very tedious to go through the entire file and delete poorly matched ATC codes.

## Method B

This method is straightforward, and it directly maps the SNOMED-CT code to the ATC code (*Figure 7*).

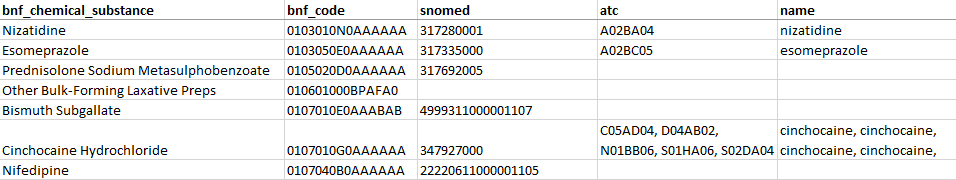


Figure : A screenshot showing an output of Method B Script

Out of 138 entries in the test file, only 83 have SNOMED-CT code assigned, and only 24 ATC codes were mapped. The yield was only 17.4%, as opposed to the best match (token sort ratio scorer) in Method A has yielded 48.6% with reasonable accuracy (*Figure 8*).

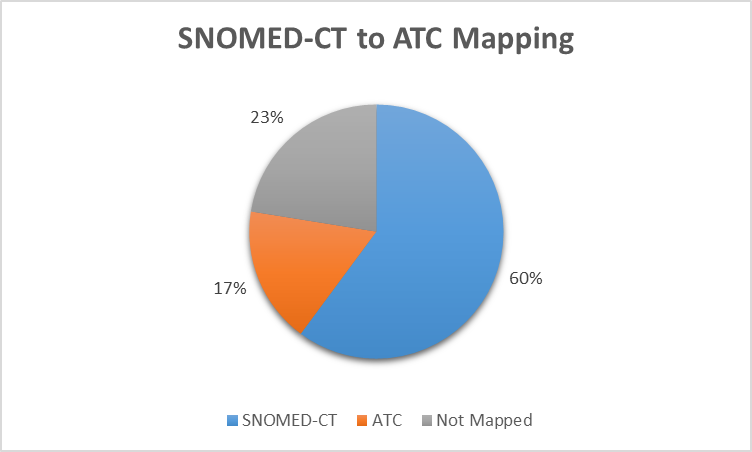


Figure : Method B Outcome

One issue to note is that one of the entries, “Timolol with diuretic” (see the result file), was mapped to several separate codes rather than a singleton code, C07DA06 - timolol, thiazides, and other diuretics. Subsequently, combination medicines were not appropriately mapped due to SNOMED-CT having a one-to-many relationship with different ATC codes. The individual code listed represent all active ingredients in that combination drug. This poses an issue when performing data analysis of prescription patterns because it could count as three or four different drugs prescribed when it was actually a single drug.

## Method C

The script, **method-c.py**, starts off with mapping the SNOMED-CT codes, from a test file, only 60.1% of total entries have SNOMED-CT code assigned. Of those, 17.4% were mapped to ATC codes (Method B), and the remaining unmapped entries (82.6% of total entries) are mapped through the fuzzy match (Method A) using the scorer with the highest F1 score from Method A which is Token Sort Ratio.

Table : Performance of combining Method B and A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *(n=138)* | Not Mapped | % | Mapped | % | Total |
| Method B - SNOMED-CT to ATC | 59 | 42.8% | 24 | 17.4% | 83 |
| Method A - Fuzzy Match | 64 | 46.4% | 50 | 36.2% | 114 |
| *Skipped (Mapped in B)* | *24* | *17.4%* |  |  |  |
| Combined | 64 | 46.4% | 74 | 53.6% | 138 |

Method C yielded about 53.6% of test data that are mapped to ATC codes.

## Full BNF File Mapping

Finally, the script, **bnf-atc-full-map.py**, performs all of the same tasks as **method-c.py** but with a cache enabled to buffer all matches returned by the fuzzy match calls to increase the execution efficiency. Without it, it would take over an hour to run instead of approximately six minutes to complete!

Table : Performance result of full BNF file mapping

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *(n=55,706)* | Not Mapped | % | Mapped | % | Total |
| Method B - SNOMED-CT to ATC | 10,538 | 18.9% | 2,008 | 3.6% | 12,546 |
| *Cached matches mapped* |  |  | *17,358* | *31.2%* |  |
| Method A - Fuzzy Match | 26,730 | 48.0% | 9,610 | 17.3% | 36,340 |
| *Skipped (Mapped in B)* | *19,366* | *34.8%* |  |  |  |
| Combined | 26,730 | 48.0% | 28,976 | 52.0% | 55,706 |

In the full BNF file, only 12,546 entries (22.5%) were mapped to SNOMED-CT. Out of that subset, only 2,008 (16% of subset, 3.6% overall) were mapped to ATC codes. With the cache enabled, it yielded 17,358 matches, whether SNOMED-CT was mapped or not. That leaves the remaining 36,340 entries (65.2%) to be mapped by the fuzzy match algorithm. Through the fuzzy match, 9,610 of the entries were mapped to ATC code with good scores. Finally, the complete file mapping using both methods yielded 52.0% of the entries that are mapped to ATC codes. However, there are false positives that need to be reviewed manually before finalizing the mapping dataset for use in research.

# **Discussion**

## Significance

Although the full BNF file only yielded 52.0% of the ATC match, it would save many person-hours from mapping each BNF code to an ATC code manually. It is most likely that almost all of the most commonly prescribed drugs from BNF formulary are mapped to ATC. Further work is needed to verify the fact that the less commonly prescribed drugs are lacking ATC mapping. One way of doing this is to gather prescriptions frequency and generate a list of the most commonly prescribed drugs along with the ATC mapping to determine how much of the most common drug prescribed is mapped to ATC.

## Limitations

The most significant limitation in this project is a lack of thoroughness of both BNF to SNOMED-CT mapping and SNOMED-CT to ATC mapping. It was evident from the full BNF file mapping, only 12,546 entries (22.5%) have a SNOMED-CT code mapped, but out of that subset, only 16% are mapped to ATC. Secondly, another limitation of this study is the size of the test file to accurately measure the performance of the match because the full-file map has produced a different yield rate for the fuzzy match component. Thus, limiting the statistical power and using a larger sample size should be considered; however, a larger sample size would require more time to build the confusion matrix manually.

Finally, the mapping algorithm does not distinguish between different organ systems according to the coding system – BNF chapter and ATC first-level. For example, as illustrated above, there are 118 BNF codes for prednisolone under seven different BNF chapters: 01, 06, 10, 11, 12, 13, and 19 (see *Table 5*). So for the ATC code, A07EA01 (Intestinal anti-inflammatory agents) would map under BNF chapter 01 because it relates to the gastrointestinal system. As a result, the script does not discriminate and lists all possible matches by drug name, ignoring the proper organ system classification.

With the above limitations and the conclusions from previous works15,16 by Dr. Bodenreider and Dr. Kury is validated through this project with similar findings: low yield mapping and ambiguity concerns. These can be addressed further in future work with improvement in algorithms and other approaches using machine learning.

## Future Work

In addition to the 52.0% yield for the full BNF file, the focus of the future work is to refine the method A algorithm to improve the yield rate by adding multi-step matching. Multi-step matching would be a series of fuzzy matches with different scoring algorithms in a specific order so that there will be fewer false positives and fill in the remaining unmatched entries. It is also worthwhile to check into the SOUNDEX match algorithm and see if the performance would be comparable or improved. The application of machine learning in this mapping project should be taken into consideration, more specifically, an exploration into supervised or reinforcement learning may give us a better result. However, the implementation design of machine learning is beyond the scope of this paper.

# **Bibliography**

1. *Guidelines for ATC Classification and DDD Assignment*. 22nd ed. Oslo, Norway: WHO Collaborating Centre for Drug Statistics Methodology; 2018. https://www.whocc.no/filearchive/publications/2019\_guidelines\_web.pdf. Accessed October 18, 2019.

2. WHOCC - Structure and principles. https://www.whocc.no/atc/structure\_and\_principles/. Accessed August 2, 2019.

3. About BNF Publications. https://www.bnf.org/about/. Accessed October 18, 2019.

4. Wade OL. British National Formulary: its birth, death, and rebirth. *Br Med J*. 1993;306(6884):1051-1054. doi:10.1136/bmj.306.6884.1051

5. Prescribing Data: BNF Codes. https://ebmdatalab.net/prescribing-data-bnf-codes/. Accessed October 21, 2019.

6. RxNorm Project History. https://www.nlm.nih.gov/research/umls/rxnorm/history.html. Accessed October 21, 2019.

7. Bennett CC. Utilizing RxNorm to support practical computing applications: Capturing medication history in live electronic health records. *J Biomed Inform*. 2012;45(4):634-641. doi:10.1016/j.jbi.2012.02.011

8. RxNorm Overview. https://www.nlm.nih.gov/research/umls/rxnorm/overview.html. Accessed October 21, 2019.

9. SNOMED CT. In: *Wikipedia*. ; 2019. https://en.wikipedia.org/w/index.php?title=SNOMED\_CT&oldid=919385008. Accessed October 25, 2019.

10. SNOMED CT FAQs. https://www.nlm.nih.gov/healthit/snomedct/faq.html. Accessed October 24, 2019.

11. 2. SNOMED CT Benefits - SNOMED CT Starter Guide - SNOMED Confluence. https://confluence.ihtsdotools.org/display/DOCSTART/2.+SNOMED+CT+Benefits. Accessed October 25, 2019.

12. SNOMED CT implementation in primary care. NHS Digital. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care. Accessed October 28, 2019.

13. Digital NHS. Delen: Home: Supplier Deployment Schedule & System Functionality. https://hscic.kahootz.com/connect.ti/t\_c\_home/viewcontent?contentId=353907. Accessed October 28, 2019.

14. 4. SNOMED CT Basics - SNOMED CT Starter Guide - SNOMED Confluence. https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics. Accessed October 25, 2019.

15. Bodenreider O, Rodriguez LM. Analyzing U.S. prescription lists with RxNorm and the ATC/DDD Index. *AMIA Annu Symp Proc*. 2014;2014:297-306.

16. Kury F, Bodenreider, Olivier. Mapping U.S. FDA National Drug Codes (NDC) to Drug Classes and Codes. *AMIA Annu Symp Proc*. 2017;2017:2070.

17. Mitchell TM. *Machine Learning*. New York: McGraw-Hill; 1997.

18. Toivanen H. Why You Must Understand Machine Learning - And The Sooner The Better. Teqmine. https://teqmine.com/must-understand-machine-learning/. Published June 1, 2016. Accessed October 30, 2019.

19. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc JAMIA*. 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464

20. Levenshtein distance. In: *Wikipedia*. ; 2019. https://en.wikipedia.org/w/index.php?title=Levenshtein\_distance&oldid=915176394. Accessed September 21, 2019.

21. Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov Phys Dokl*. 1966;10:707-710.

22. BNF Code Information Download. https://apps.nhsbsa.nhs.uk/infosystems/data/showDataSelector.do?reportId=126. Accessed October 14, 2019.

23. RxNorm Files. https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html. Accessed October 14, 2019.

24. BNF SNOMED mapping. https://www.nhsbsa.nhs.uk/prescription-data/understanding-our-data/bnf-snomed-mapping. Accessed October 14, 2019.

25. Hayes B. Programming Languages Most Used and Recommended by Data Scientists |. https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/. Accessed October 14, 2019.

26. Li S. Natural Language Processing for Fuzzy String Matching with Python. Medium. https://towardsdatascience.com/natural-language-processing-for-fuzzy-string-matching-with-python-6632b7824c49. Published December 6, 2018. Accessed October 21, 2019.