data

MDPI

Data Descriptor

Drugs, Active Ingredients and Diseases Database in Spanish. Augmenting the Resources for Analyses on Drug–Illness Interactions

Irene López-Rodríguez [1], César F. Reyes-Manzano [2], Israel Reyes-Ramírez [1], Tania J. Contreras-Uribe [2] and Lev Guzmán-Vargas [1,*]

[1] Unidad Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Av. IPN No. 2580, L. Ticomán, Ciudad de México 07340, Mexico; ilopezr0600@alumno.ipn.mx (I.L.-R.); ireyesr@ipn.mx (I.R.-R.)

[2] Tecnológico Nacional de México, Tecnológico de Estudios Superiores de Ixtapaluca, Km. 7 Carretera Ixtapaluca-Coatepec S/N San Juan, Ixtapaluca, Estado de México 56580, Mexico; cesarrm5@hotmail.com (C.F.R.-M.); jetzabel.tania@hotmail.com (T.J.C.-U.)

* Correspondence: lguzmanv@ipn.mx; Tel.: +52-55-57296000 (ext. 56873)

Abstract: Quantitative and qualitative data on active-ingredient drug composition are essential information for characterizing near-field exposure of consumers to product-related chemicals, among other things. Equally as important is the characterization of the relationship between one or many active ingredients in terms of the diseases they are prescribed for. Such evaluations, however, require quantitative information at different anatomical levels. To complement the available sources of information on active substances and diseases, we have designed a database with enough versatility to potentially be used in a variety of analyzes. By using information provided by a well-established online pharmacological dictionary, we present a database with 11 tables which are easy to access and manipulate. Specifically, we present datasets containing the details of 12,827 marketed drug products, 40,164 diseases, 6231 active pharmaceutical ingredients and 4093 side effects. We exemplify the usefulness of our database with three simple visualizations, which confirm the importance of the data for quantifying the complexity in the associations among active substances, diseases and side effects. Although there are databases with detailed information on active substances and diseases, none of them can be found in Spanish. Our work presents an option that contributes substantially to obtaining well classified information in order to evaluate the roles of active pharmaceutical ingredients, diseases and side effects. These datasets also provide information about clinical and pharmacological groupings which may be useful for clinical and academic researchers. The database will be regularly updated and extended with the newly available Virtual Medicinal Products.

Dataset: https://dx.doi.org/10.6084/m9.figshare.7722062.

Keywords: drug–disease classification; drug–disease interaction; pharmacological database

Citation: López-Rodríguez, I.; Reyes-Manzano, C.F.; Reyes-Ramírez, I.; Contreras-Uribe, T.J.; Guzmán-Vargas, L. Drugs, Active Ingredients and Diseases Database in Spanish. Augmenting the Resources for Analyses on Drug–Illness Interactions. *Data* **2021**, *6*, 3. https://doi.org/10.3390/data6010003

Received: 8 November 2020
Accepted: 4 January 2021
Published: 9 January 2021

1. Background

Since 1998, Allen Kratz classified the process of diversification or repurposing for the discovery of new drugs as high risk [1]; many other authors have executed studies of drug–disease interactions from different perspectives [2–4]. It is recognized that drug discovery is a complicated process (see [5] for a historical perspective), where pharmaceutical companies could spend between 7 and 12 years to bring a new drug to the market, with recent estimates on cost ranging from $314 million to $2.8 billion [6]. An important impediment for the commercialization of new drugs is the low odds of survival in the first phase of clinical

trials [7]. Despite these low odds and long times of discovery and development, billions of dollars in revenue are generated per year for at least a decade. Therefore, to increase the odds of creating successful drugs, pharmaceutical companies began to diversify the process of discovery of new drugs using new technologies to increase the range of possible candidates that benefit from a new drug [1].

As a result, the rapid growth in the field of drug discovery is associated with advances in technology, faster innovation and the minimization of unnecessary spending. This economic risk is particularly high in developed and some developing countries, where the idea of paying for medicines on the basis of their current performances is seen as a factor dependent on patient choice [8]. The idea of reducing risks through diversification is very practical, especially after a group of drugs continues to fail for many reasons, such as drug resistance, allergic and immune reactions, adverse drug interactions and problems related to bio-availability. For instance, a drug can cause abnormalities in processes such as absorption, distribution, metabolism and excretion, which can occur more frequently among those patients with different comorbidities. In the medical area, doctors usually have a table of medications for the treatment of a specific disease, which enables doctors to decide on the best treatment option, while giving the patient details of the possible side effects or risks involved. In this sense, a comprehensive data source that links the anatomical drug classification of an active ingredient to a disease classification system is needed to reduce risk in drug repurposing or drug prescription scenarios. We remark that these datasets already exist (for instance, DrugBank [9], DrugCentral [10] and SIDER [11]), but are difficult to access and not readily available in Spanish. This article presents a new set of data that will be useful to researchers in order to evaluate the roles of active pharmaceutical ingredients and diseases at different levels. Specifically, our repository provides the possibility of analysis from at least two perspectives: (i) clinical researchers will be able to analyze the interdependence of active ingredients and diseases, and will be able to classify the presence of comorbidities and provide information about possible side effects; and (ii) network science investigators will be able to analyze the active substances space to characterize the relationship between every active ingredient.

2. Methods

2.1. Data Request to Vademecum

In order to comprise our database, the authors examined different national and international databases for identifying a complete dataset which contained the majority of the target–event relationships between diseases, active substances and side effects for each drug in the Spanish language. Website managers of the most important online pharmacological dictionaries were contacted in order to inquire about their willingness to contribute data. We received one answer from one major international source: Vademecum, a recognized pharmacological handbook which includes advertising of medicines and prescriptions of pharmaceutical specialties addressed to health professionals [12]. They shared with us 56 tables; the archived copies were possible to access via web services.

2.2. Data Collection Process

Our procedure to gather data, consisted of converting the information obtained from XML requests into flat files for better data manipulation. Through XML web services, we were able to access the tables that interested us. We then proceeded to filter the fields and collected the data in flat files. Nevertheless, these web services are not publicly available to everybody. These web services do not provide a publicly downloadable file for all chemical–disease pairs which would help describe drugs, or compare drug makers and indications from health professionals. To address this problem, we generated an adapted algorithm as a toolkit to extract information from the available data source, mainly using the flat files that Vademecum provided us with.

2.3. Load Data from Flat Files to PostgreSQL Database

To fully understand how defined the relationships between each table are and identify which data were useful for our purpose, it was necessary to load the 56 flat files (tables) to a PostgreSQL instance which was installed in our local server. After checking the attributes query results, and relationships, it was discovered that we only required 11 tables to get the needed interactions for future experiments and also for our technical validation. A graphical representation of entities and their relationships is depicted in Figure 1.

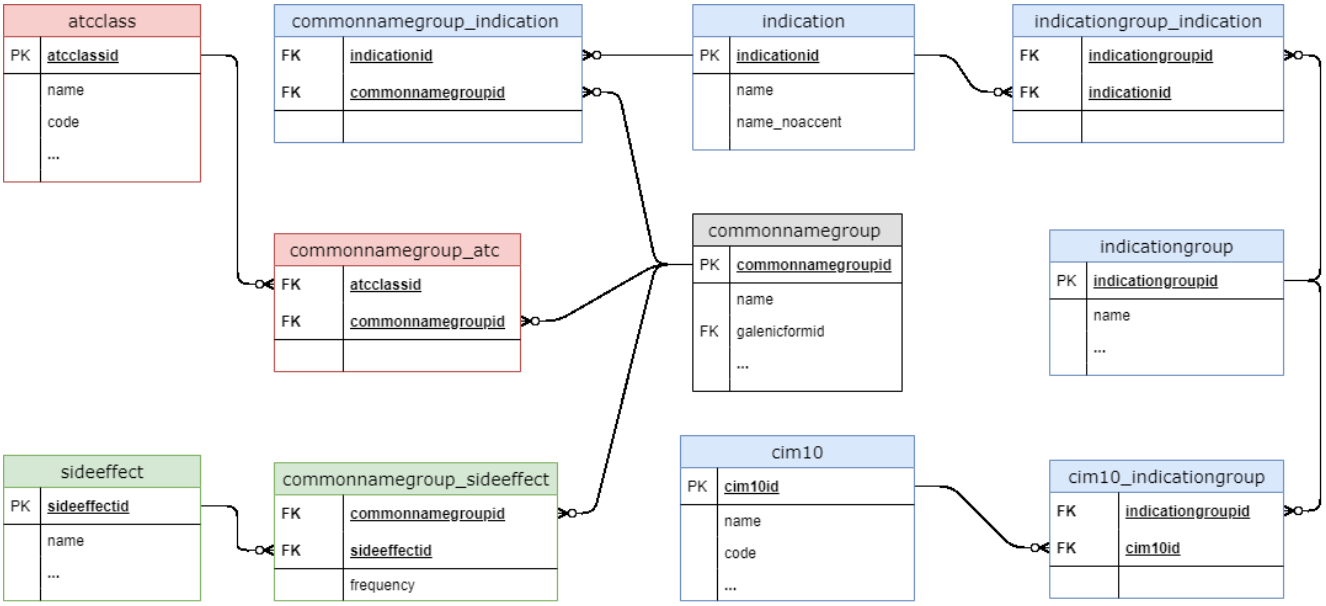

Figure 1. Schematic representation of our database's structure. The database contains 11 tables which comprise details of drugs, active ingredients, diseases and side effects.

2.4. Standarization

Our datasets were constructed from a variety of available data sources on drugs, active pharmaceutical ingredients, side effects and diseases, in addition to associated categorical groupings such as the Anatomical Therapeutic Chemical Classification System (ATC) and the International Classification of Diseases (ICD). The ICD and ATC standards for healthcare data are better because they support semantic interoperability with other systems, and the benefits of standardization and reproducibility using those standards allow for combination with other home data sources to obtain different insights. To calculate the amount of processed records out, a schematic representation of the manipulation of the information is depicted in Figure 2.

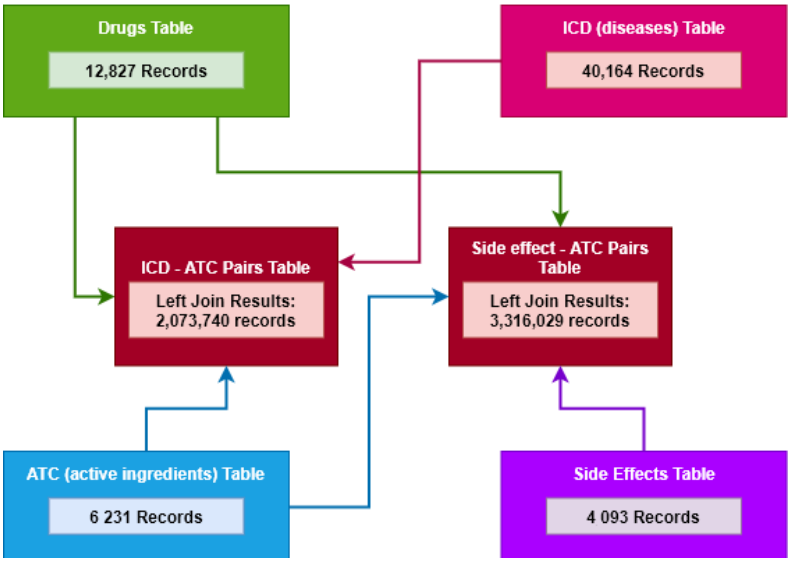

Figure 2. Schematic overview of the operational workflow to construct the *ATC-ICD-DATA* and *ATC-SIDEEFFECT-DATA*. The integrated datasets contains 2,073,740 records for ICD–ATC pairs and 3,316,029 records for side effects–ATC pairs. The data was extracted from relational database using web services XML (see main text for details); the information such as drug name, disease name, active ingredient name, side effects and other attributes were obtained from Vademecum (VDM) relational database by using SQL's SELECT statement.

2.5. Data Content of Drugs, ATC and ICD Tables

As we mentioned before, the tables were queried from our database using SQL syntaxis. This SQL SELECT statement retrieves records from a table according to clauses that specify criteria. The existing datasets of medicine compositions are described briefly below together with details of the data included in our database.

2.5.1. Drugs Table

The table only presents the prescriptions of all specialties on the market for common denomination (active ingredient + doses + pharmaceutical form + administration route). For practical recognition, this group is called Virtual Medicinal Product (VMP), except in the following cases: (a) multivitamins or drugs with more than three active ingredients, (b) insulin, (c) drugs with sequential doses, (d) cosmetics, disinfectants and pesticides, (e) the drugs whose technical datasheets do not recommend the use of common denomination VMP, (f) radio-pharmaceuticals, (g) homeopathic medicinal products and (h) medicinal plants. For this table, it was not possible to find a convenient, downloadable and up-to-date list in Spanish language with all the ATC–ICD code pairs among the principal health websites such as DrugBank [9], the European Commission [13] and the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology [14].

2.5.2. ATC Table

ATC classifies the active ingredients of drugs according to the organs on which they act and their therapeutic, pharmacological and chemical properties [14]. We resorted to a complete catalog with 6231 records in Spanish language, and also from the provided flat files it was possible to get a convenient dataset.

2.5.3. ICD Table

The ICD is a classification system which provides a system of diagnostic codes for classifying diseases [15]. This table integrates information from the 10th revision of the International Classification of Diseases and Related Health Problems (ICD), a medical classification list by the WHO with more than 40,164 different codes in Spanish language.

2.6. Integrated Datasets

Our integrated datasets are the results of the joining of the nine tables and their attributes listed in Table 1. Two tables, *indication* and *indicationgroup*, were omitted from the entity relationship diagram to generate our SQL queries. As the foreign keys were used to retrieve information, they act as a cross-reference between tables and also reference the primary keys of the omitted tables. Thereby, it is possible to establish a link while dispensing with both tables.

Each of these tables has specific information about ATCs (*atcclass*), classification of diseases (*cim10*), relationships between ICD10 and indications (*cim10_indicationgroup*), VMP group (*commonnamegroup*), VMP group and ATC class (*commonnamegroup_atc*), VMP group and therapeutic indication (*commonnamegroup_indication*), VMP group and adverse effects (*commonnamegroup_sideeffect*), specific therapeutic indications and general therapeutic indications (*indicationgroup_indication*) and adverse effects (*sideeffect*). The extraction of drug and medication records from each table was achieved via SQL statement. The ICD and ATC catalogs were automatically downloaded from their respective websites [14] as CSV files using Python. CSVs were stored in the same PostgreSQL database. Some missing data fields and relationships in the drug table that were generated after data collection (assignment of ICD and ATC code) are detailed below; all other data fields are replicated exactly as they were in the original data source [14]. Regarding the integration process, for each drug-specific record, the disease was assigned to an ICD subcategory, and the active pharmaceutical ingredient was assigned to an ATC level code joining the disease and chemical name columns with their respective codes stored in ICD and ATC catalog tables. We notice that the table *cim10* contains 275 records that do not show a specific ICD code but only a full chapter (see Table 1). For these 275 records, an additional inspection was necessary; we use the Jaro–Winkler distance to evaluate the similarity between two ICD names (phrase object) with a constant scaling factor $p = 0.1$ [16]. If the score difference is sufficiently low or equal to zero, the relationship within the candidate phrase object is accepted and then added to the database.

Our algorithm provides two options for acquiring data: (i) SQL statements and (ii) Pandas DataFrames using Python language. For option (i), the drugs and medications corpora of 2,073,740 records were retrieved using SQL SELECT statements (see Figure 3 for a schematic description).

For option (ii), the process consists of combining DataFrames on Python using columns in each dataset that contain common values (a common unique ID), and this process is also called joining (see Figure 4 for a schematic description). This process of joining tables is similar to what we do with tables in an SQL database. We use the common type of join, called an inner join. An inner join combines two DataFrames based on a join key and returns a new DataFrame that contains only those rows that have matching values in both of the original DataFrames. For this process it is not necessary to install a database management tool, such as PostgreSQL, for doing the operations with the data.

Regarding the extraction and combination of rows from two or more tables, a join clause was used, which is based on a related column between them. In this case our related columns are the disease name from both the drugs table and the ICD catalog, and the chemical description from the drugs table and ATC catalog.

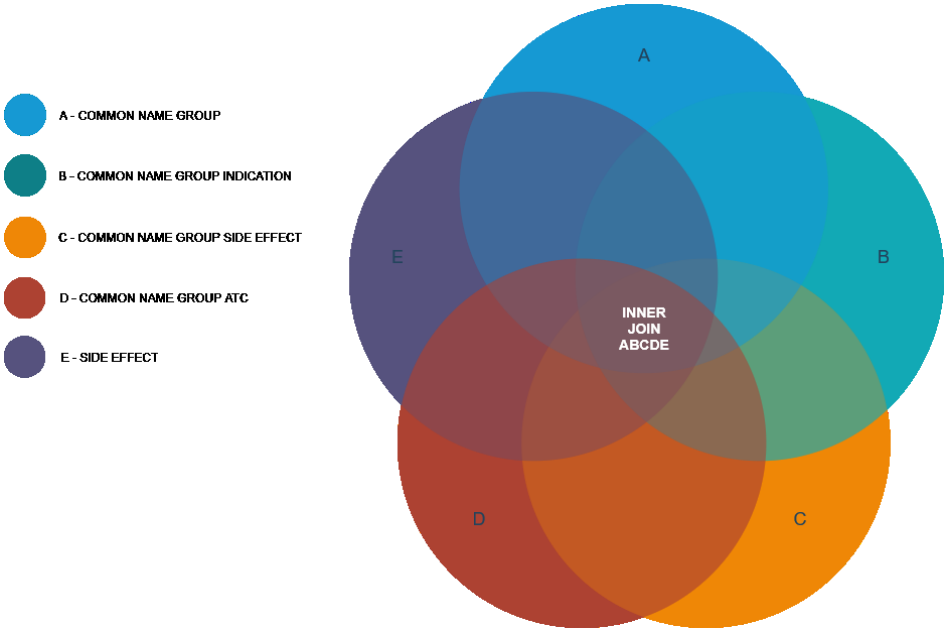

Figure 3. Schematic representation of the process (option (i)) for acquiring data based on SQL staments. This process utilizes an inner join keyword for selecting records that have matching values in the 5 tables shown in the Venn diagram. With the SQL statement, it is possible to select all the attributes defined in the tables, and there is the ability to match other data sources that include ATC and ICD codes with descriptions in different languages.

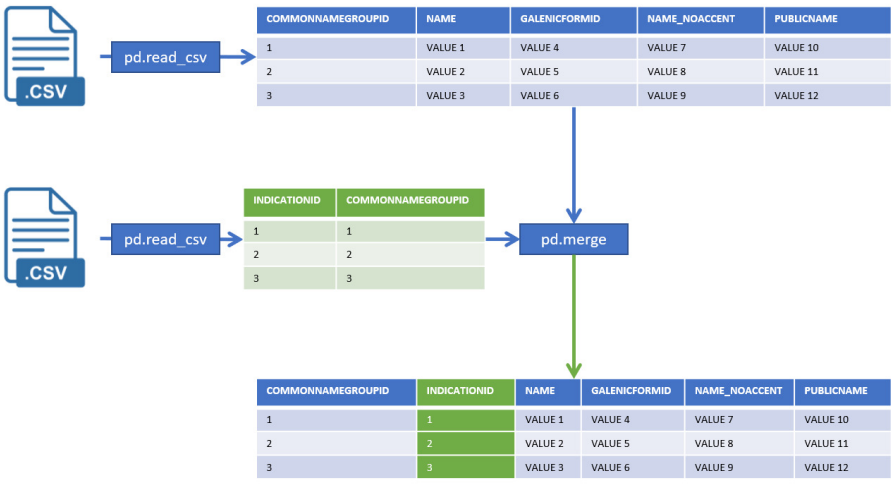

Figure 4. Schematic representation of the process (option (ii)). This process consists of combining Pandas DataFrames as s data manipulation tool for joining tables.

Table 1. List of source tables in our database. We present the table and attributes names, descriptions and number of records for each table.

Name	Description	Rows Number	Attribute Name	Attribute Description
atcclass	Anatomical Therapeutic Chemical Classification System (ATC International)	6231	atcclassid	ATC class identifier
			parentid	Top level ATC code identifier
			code	ATC code
			name	ATC title description
cim10	The 10th revision of the International Classification of Diseases and Related Health Problems	40,164	cim10id	ICD10 identifier
			parentid	ICD10 group identifier
			code	ICD10 code
			name	ICD10 description
cim10_indicationgroup	Relationship between ICD10 and general indications	10,432	indicationgroupid	Indication group identifier
			cim10id	ICD10 identifier
commonnamegroup	VMP Group defines the drugs from active ingredient, doses, doses unit, route and pharmaceutical form	12,827	commonnamegroupid	Virtual Medicinal Product(VMP) group identifier
			name	VMP group description
			galenicformid	Galenic form identifier of the VMP group
			name_noaccent	VMP group description without accent
commonnamegroup_atc	Relationship between VMP Group and ATC class	12,320	atcclassid	ATC class identifier
			commonnamegroupid	VMP group identifier
commonnamegroup_indication	Relationship between VMP Group and therapeutical indication	50,662	indicationid	Therapeutic indication identifier
			commonnamegroupid	VMP group identifier
commonnamegroup_sideeffect	Relationship between VMP Group and adverse effects	631,243	commonnamegroupid	VMP group identifier
			sideeffectid	Side effect identifier
			frequency	Frequency of occurrence of adverse effects
indicationgroup_indication	Relationship between specific therapeutic indications and general therapeutic indications	5292	indicationgroupid	General therapeutic indication identifier
			indicationid	Particular therapeutic indication identifier
sideeffect	Adverse effects (these side effects imply the noxious and non-intentional responses from the use of the drugs)	4093	sideeffectid	Side effect identifier
			name	Side effect description

3. Results

3.1. Data Description

The lack of publicly available datasets with easy access and manipulation of active pharmaceutical ingredients and drugs is a problem for researchers of the field, especially, when working with data whose origin is not in English, and a direct one-to-one correspondence is desirable between elements. Here, we report the construction of the datasets ATC-ICD-DATA and ATC-SIDEEFFECTS-DATA. Both static and dynamic forms of the database can be freely downloaded in CSV formats through Figshare https://figshare.com/s/5b3128788640d7aa7d4f. The data consist of an available web-based application and hub integrating data containing 45,584 records (including drugs, ICD codes and ATC codes). The specific description of the provided files is given in Table 2. Various types of data are associated with each result table, which contain descriptions of the active ingredients, names of each level, disease names, diagnoses, gender and morbidity rates. In the case of drugs, the following information is included: medicine name, laboratory name and active ingredients.

Table 2. Script and dataset file names and descriptions in FigShare and GitHub repositories (in GitHub to find the files listed, is needed to navigate between folders).

File Name	Description
get_full_data.py	Script which load data from CSV files to PostgreSQL database in a methodical and automated manner and extract joined datasets using SQL statements (insert number 1 via command line for this option). For joining dataset using Pandas DataFrames, insert number 2 via command line (The minimum requirement for running option 2 is 8GB RAM)
Technical Validation.ipynb	Jupyter notebook where data technical validation is shown
requirements.txt	The list of Python module names required for Nbviewer
covid_subtances.graphml	Network ready for opening on Gephi o Cytoscape software
atcclass.csv	Anatomical Therapeutic Chemical Classification System (ATC International)
cim10.csv	The 10th revision of the International Classification of Diseases and Related Health Problems
cim10_indicationgroup.csv	Relationship between ICD10 and general indications
commonnamegroup.csv	VMP Group defines the drugs from active ingredient, doses, doses unit, route and pharmaceutical form
commonnamegroup_atc.csv	Relationship between VMP Group and ATC class
commonnamegroup_indication.csv	Relationship between VMP Group and therapeutical indication
commonnamegroup_sideeffect.csv	Relationship between VMP Group and adverse effects
indicationgroup_indication.csv	Relationship between specific therapeutic indications and general therapeutic indications
sideeffect.csv	Adverse effects (these side effects imply the noxious and non-intentional responses from the use of the drugs)
atc_icd_data.csv	Integrated dataset result of our algorithm in format CSV, it is available only in FigShare inside a ZIP file with the same file name
atc_sideeffects_data.csv	Integrated dataset result of our algorithm in format CSV, it is available only in FigShare inside a ZIP file with the same file name

For convenience, the dataset described in this report is stored in a PostgreSQL multi-dimensional database. A copy of the PostgreSQL database as of February 2020 is archived in Figshare [17]. Users may install PostgreSQL and download the dataset for manipula-

tion and data extraction. Within the PostgreSQL database, two types of data are stored: metadata and record-specific data. Metadata refers to descriptions and information to all or most of the records. Record-specific data refers to specific single data records. Besides, metadata also describe where the data records are stored and how the right record can be located. As it was described in the previous section, the data were extracted from different tables, and we focused on the following characteristics:

Medicine compositions. Reported data on the composition of a large number of consumer medicines, including active pharmaceutical ingredients, medicine name, presentation, shape, administration and laboratory.

Active ingredients' categorizations by anatomically therapeutic chemicals. Reported data from ATC that active substances are classified in a hierarchy with five different levels. The classification system has fourteen main anatomical/pharmacological groups or first levels. Each ATC main group is divided into second levels, which could be either pharmacological or therapeutic groups. The third and fourth levels are chemical, pharmacological or therapeutic subgroups, and the fifth level is the chemical substance. The second, third and fourth levels are often used to identify pharmacological subgroups when they are considered more appropriate than therapeutic or chemical subgroups [14]. Table 3 is an example of the complete classification of metformin. As it illustrates the structure of the classification code:

Table 3. Example of the complete classification of metformin. Active ingredients' categorization by the Anatomical Therapeutic Chemical Classification System.

Code	Description
A	Alimentary tract and metabolism (1st level, anatomical main group)
A10	Drugs used in diabetes (2nd level, therapeutic subgroup)
A10B	Blood glucose lowering drugs, excl. insulins (3rd level, pharmacological subgroup)
A10BA	Biguanides (4th level, chemical subgroup)
A10BA02	Metformin (5th level, chemical substance)

International classification of diseases. ICD classifies the universe of diseases, disorders, injuries and other related health conditions in a comprehensive and hierarchical manner. For instance, the ICD-10-WHO, which is the 10th revision of the ICD [15], contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases. Additionally, the information is presented with three-character categories or three-character codes. Here, a category is defined by its class title and an alphanumeric code. The first character denotes the letter of the chapter. In particular cases, the three-character categories, such as "I01 Rheumatic fever with heart involvement," represent individual pathological conditions; however, in general, one three-character category comprises several diseases with similar characteristics.

Almost all categories have four-character subcategories. The four digits are separated from the three-digit code by a decimal point. Their values can range between 0 and 9, but not all values have to be used. Subcategories ending in 8 or 9 are mostly used for other specified and unspecified pathological states within this category; for example, take the code I09.9 with description "Acute rheumatic heart disease, unspecified." Codes without subcategories are referred to as terminal codes. For instance, Table 4 shows an example of the classification of rheumatic fever with heart involvement disease.

Table 4. Example of the classification of rheumatic fever with heart involvement disease–International Classification of Diseases, 10th revision (ICD-10).

Chapter	Group	Category	Codes
IX Diseases of the circulatory system (I00–I99)	Acute rheumatic fever (I00–I02)	I01 Rheumatic fever with heart involvement	I01.0 Acute rheumatic pericarditis
			I01.1 Acute rheumatic endocarditis
			I01.2 Acute rheumatic myocarditis
			I01.8 Other acute rheumatic heart disease
			I09.9 Acute rheumatic heart disease, unspecified

3.2. Technical Validation

The main goal of our database is ensuring that data included in the set accurately reflect the quality of the raw data sources. In some cases, errors were identified in the process of migrating data from CSV files into the PostgreSQL database. In these cases, custom Python scripts were written to check all data and correct these errors. It is likely that relationships between one ATC code and one ICD code can be very diverse—i.e., one ATC code points to various ICD codes. After applying the extractor algorithm to join the data frames, we checked data integrity, accuracy, completeness, consistency, timeliness and uniqueness. Quality checks were implemented to ensure that: the data are logically organized; the data stored in the tables are correct and as per requirements; and unnecessary data present in the dataset are eliminated. Typical actions to ensure data quality that we applied were to: eliminate redundant data, validate input data, handle missing data, ensure data values fall within defined domains and resolve data conflicts.

By relating the anatomical levels from ATC and ICD codes, it is possible to identify that the highest values per anatomical name group (ATC) really fit with organs or systems where the active ingredient should have an effect. For example, the relationship N—G, the letter N from ATC defines those substances which were created to help the nervous system, and the letter G from ICD refers to the diseases related to the nervous system (see Table 5).

Table 5. Examples of one interpretation of the associations between the anatomical group names and diseases. See heatmap in Figure 5b.

Relationship	Anatomical Name Group (ATC)	Disease Names per Blocks (ICD)
A_ATC-K_ICD	A-Alimentary tract and metabolism	K-Diseases of the digestive system
B_ATC-K_ICD	B-Blood and blood forming organs	K-Diseases of the digestive system
C_ATC-I_ICD	C-Cardiovascular system	I-Diseases of the circulatory system
D_ATC-L_ICD	D-Dermatologicals	L-Diseases of the skin and subcutaneous tissue
G_ATC-N_ICD	G-Genito-urinary system and sex hormones	N-Diseases of the genitourinary system
H_ATC-L_ICD	H-Systemic hormonal preparations, excluding sex hormones and insulins	L-Diseases of the skin and subcutaneous tissue
J_ATC-A_ICD	J-Antiinfectives for systemic use	A-Certain infectious and parasitic diseases
L_ATC-C_ICD	L-Antineoplastic and immunomodulating agents	C-Neoplasms
M_ATC-M_ICD	M-Musculo-skeletal system	M-Diseases of the musculoskeletal system and connective tissue
N_ATC-G_ICD	N-Nervous system	G-Diseases of the nervous system
P_ATC-B_ICD	Antiparasitic products, insecticides and repellents	B-Certain infectious and parasitic diseases
R_ATC-J_ICD	R-Respiratory system	J-Diseases of the respiratory system
S_ATC-L_ICD	S-Sensory organs	L-Diseases of the skin and subcutaneous tissue
V_ATC-L_ICD	V-Various	L-Diseases of the skin and subcutaneous tissue

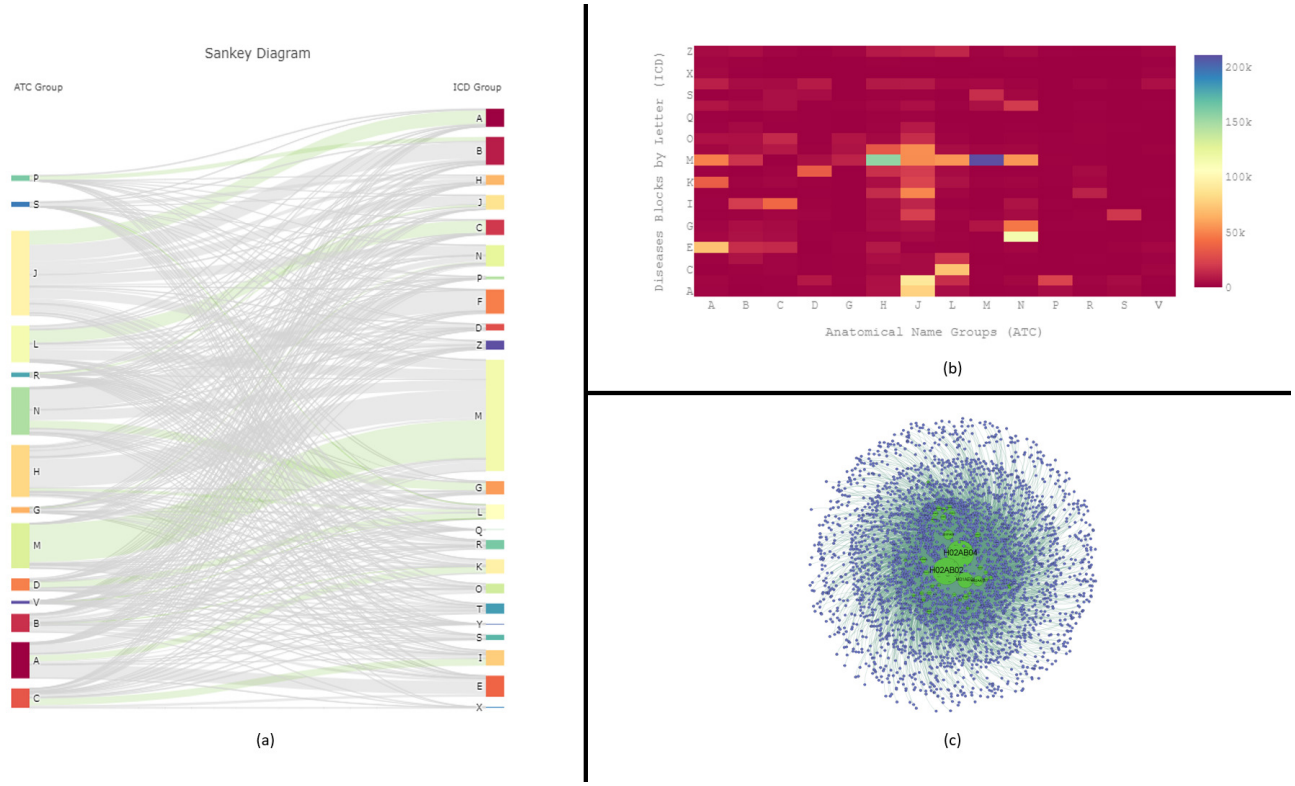

Figure 5. Examples of the utility of our dataset. (**a**) Sankey diagram to represent the association intensities between active ingredients and diseases. Here, the width is proportional to the number of links (associations) between ATC's and ICD's groups. (**b**) Heatmap of the associations between anatomical group names and blocks of diseases. We observe the presence of groups of active ingredients (ATC) and diseases (ICD), which concentrate high numbers of elements. (**c**) Bipartite network of the most promising COVID-19 treatments. In this case, nodes in green represent active ingredients, nodes in blue are diseases and a link exists if the ATC is prescribed for a disease.

To further illustrate the utility of our dataset, we generated a Sankey diagram (see Figure 5a) to represent the association intensities between active ingredients and diseases, where the width of arrows is proportional to the flow quantity. Particularly, the flows in color green color represent the strengths of the associations presented in Table 5. Moreover, with the help of this diagram, it is possible to identify the diversification of the active ingredients on the different anatomical levels that they could act on. We observe that, at the anatomical group level, there are some ATC groups that participate in many ICD groups, and others only point to a few ICD groups. For instance the ATC group J (Antiinfectives for systemic use) is the biggest and most diverse one, while for ICD group, M (Diseases of the musculoskeletal system and connective tissue) is the most diverse. Additionally, in order to visually evaluate the grouping behavior of the anatomical names and the diseases, we generated a heat map where the anatomical group names are plotted against blocks of diseases (see Figure 5b). In this plot, the values in the color bar represent the total number of active substances grouped by both axes. The plot exhibits the presence of blocks or groups which notably contain high numbers of both active ingredient and diseases. The profile of this heatmap highlights the coincidence of ATCs (or ICDs) that belong to the same anatomical system. For instance, the pairs of groups M–M (musculo-skeletal system–muscle-skeletal system) and H–M (systemic-hormonal–musculo-skeletal system) are the ones with the highest incidences of association.

Finally, by making use of our dataset, it was possible to construct a bipartite network where two types of vertices can be identified according with the ATC and ICD sets; i.e., nodes can be active ingredients or diseases, whereas a link exists between one ATC ver-

tex that points to one ICD. For instance, Figure 5c shows a subgraph of the network derived from active ingredients suggested for treating the novel coronavirus SARS-CoV-2 [18]. All nodes in ATC are green and in ICD are purple. The ATC codes presented in the network are the active substances considered for treatment. From a complex network's perspective, it is important to evaluate interactions from the bipartite network, where nodes (either ATC or ICD) can be linked if they share a common node. These projections would be indicative of the roles of active ingredients and diseases in a general map, and eventually for drug repurposing. Besides, by using the information of codes or groups, it would be possible to identify which pairs of ATCs really fit with organs or systems where the active ingredient should have an effect. In addition, by using the information provided in our dataset and the integration of others, an exploration potential drug combinations would be possible.

3.3. Comparison with Other Databases

We evaluated the drug disease information in DrugBank [9], DrugCentral [10] and SIDER [11], in order to contrast the interactions of them and our dataset. We queried the information from the datasets listed above which provide easy access to the main medical terminologies in the Unified medical Language System (UMLS), and we only consider records containing active ingredient–disease pairs. With that premise, we queried all the information and obtained the number of records shown in Table 6.

Our dataset contained more records and active ingredients than other datasets; however, the number of active ingredients decreased when we matched them with their respective ICD10 codes. In Table 7, we used the related data to compare the number of records by anatomical name group.

Table 6. Counts in the four datasets by total, total number of active ingredients and number of active ingredients related to an indication (disease).

Records Analyzed	DrugBank	DrugCentral	Vademecum	SIDER
Records number of unique ICD-ATC codes pairs	13,580	106,916	260,995	30,835
Unique Active Ingredients in the catalog table	4484	5067	6231	1560
Active Ingredients related to an indication	4484	3295	2729	1560

Table 7. Anatomical name group counts in DrugBank, DrugCentral, Vademecum and SIDER.

ATC Group	Anatomical Name Group	DrugBank	DrugCentral	Vademecum	SIDER
A	Alimentary tract and metabolism	554	408	354	169
B	Blood and blood forming organs	190	129	142	64
C	Cardiovascular system	571	409	308	159
D	Dermatologicals	345	254	208	127
G	Genito-urinary system and sex hormones	264	192	160	85
H	Systemic hormonal preparations, excluding sex hormones and insulins	69	51	51	39
J	Antiinfectives for systemic use	402	316	261	151
L	Antineoplastic and immunomodulating agents	344	295	251	123
M	Musculo-skeletal system	217	150	124	75
N	Nervous system	588	415	339	226
P	Antiparasitic products, insecticides and repellents	118	77	56	25
R	Respiratory system	364	251	207	103
S	Sensory organs	264	239	157	148
V	Various	194	109	111	66
Total		**4484**	**3295**	**2729**	**1560**

A considerable proportion of the active ingredients have a greater correspondence for all with DrugBank, followed by DrugCentral and thirdly by our dataset (see Figure 6). There is a clear division with SIDER due to limited amount of active ingredients found within this data set.

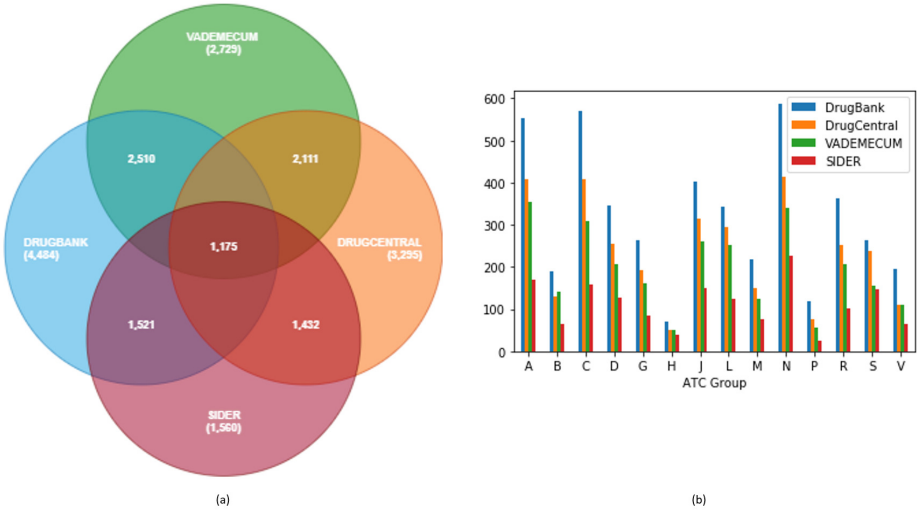

Figure 6. Comparison. (**a**) Venn diagram and (**b**) bar chart which show the frequencies of active ingredients in active ingredient–disease pairs.

Clearly, more could be done to compare the number of diseases using the ICD10 code. It is worth mentioning that the other datasets use different medical terminologies than UMLS, which include SNOMED CT and MedDRA. For future reference, we will map directly the terminologies to unify and also compare by diseases.

4. Code Availability

The Drugs, Active Ingredients and Disease Database is available on Figshare and GitHub. The code used to join and visualize the data can be obtained from both repositories. The joining of all contributed data and algorithms used to do so was done in the version of Python > 2.7. Other necessary packages are required to be added, not to mention how important it is to install PostgreSQL for running get_data_full.py script. Regarding Jupyter notebook, it can be downloaded and run in personal computers or open on Nbviewer. For details, a README file is available on GitHub and FigShare. Originally contributed data or updated data are only available by contacting Vademecum.

5. Conclusions

Obtaining information on the interactions between active substances and diseases is a priority from clinical and academic points of view. However, accessing this type of information is always a challenge due to the differences in the methodologies to classify them, and the nomenclature, which changes according to the country and also due to the continuous development of new drugs. In fact, most of current datasets are difficult to access and not readily available. Our work presents an option that contributes substantially to obtaining well classified information in order to evaluate the roles of active pharmaceutical ingredients, diseases and side effects. These datasets also provides information about clinical and pharmacological grouping which may be useful for clinical researchers and physicians. Though a wide range of consumer drugs are available, the drugs included in our dataset are limited to those for which data were publicly available (Vademecum). Although data included in our dataset can be used in many ways for future analysis, users should be aware of the limitations of the information, and the intended usage of the data—more specifically, when querying data by some specific field or attribute of the resultant table. Currently, it is possible to update the ICD table with all the available versions of the ICD system (the latest versions are ICD-9, ICD-10 and ICD-11). However, using the attributes specified in the catalogs, it can be easily queried by any field to form subsets of active ingredients with desired disease properties or inclusion criteria. Finally, we encourage pharmaceutical corporations to adopt a unified classification system of diseases and active

ingredients in order to get more self-consistent datasets. Full documentation, demos and examples are also available from https://github.com/sydmizar/drugs-datasets.

Author Contributions: Conceptualization and methodology, I.L.-R. and L.G.-V.; software and data curation, I.L.-R., C.F.R.-M., T.J.C.-U. and I.R.-R.; writing—reviewing and editing, I.L.-R., T.J.C.-U. and L.G.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Consejo Nacional de Ciencia y Tecnología (abbreviated CONACyT) and Instituto Politécnico Nacional.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare at https://dx.doi.org/10.6084/m9.figshare.7722062.

Acknowledgments: We thank all of Complex Systems Lab's members for fruitful comments and suggestions.

Conflicts of Interest: There were no competing interests.

References

1. Krantz, A. Diversification of the drug discovery process. *Nat. Biotechnol.* **1998**, *16*, 1294. [CrossRef] [PubMed]
2. Guney, E.; Menche, J.; Vidal, M.; Barábasi, A.L. Network-based in silico drug efficacy screening. *Nat. Commun.* **2016**, *7*, 10331. [CrossRef] [PubMed]
3. Vinayagam, A.; Gibson, T.E.; Lee, H.J.; Yilmazel, B.; Roesel, C.; Hu, Y.; Kwon, Y.; Sharma, A.; Liu, Y.Y.; Perrimon, N.; et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 4976–4981. [CrossRef] [PubMed]
4. Ye, H.; Tang, K.; Yang, L.; Cao, Z.; Li, Y. Study of drug function based on similarity of pathway fingerprint. *Protein Cell* **2012**, *3*, 132–139. [CrossRef] [PubMed]
5. Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960–1964. Available online: https://science.sciencemag.org/content/287/5460/1960.full.pdf (accessed on 2 February 2020). [CrossRef] [PubMed]
6. Wouters, O.J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018. *JAMA* **2020**, *323*, 844–853. Available online: https://jamanetwork.com/journals/jama/articlepdf/2762311/jama_wouters_2020_oi_200015.pdf (accessed on 19 December 2020). [CrossRef] [PubMed]
7. Mohs, R.C.; Greig, N.H. Drug discovery and development: Role of basic biological research. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2017**, *3*, 651–657. [CrossRef] [PubMed]
8. Leaders. Getting Medicines to Market Faster. 2018. Available online: https://www.economist.com/leaders/2018/03/24/getting-medicines-to-market-faster (accessed on 6 March 2020).
9. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2013**, *42*, D1091–D1097. Available online: https://academic.oup.com/nar/article-pdf/42/D1/D1091/3559045/gkt1068.pdf (accessed on 17 June 2020). [CrossRef] [PubMed]
10. Ursu, O.; Holmes, J.; Bologa, C.G.; Yang, J.J.; Mathias, S.L.; Stathias, V.; Nguyen, D.T.; Schürer, S.; Oprea, T. DrugCentral 2018: An update. *Nucleic Acids Res.* **2018**, *47*, D963–D970. Available online: https://academic.oup.com/nar/article-pdf/47/D1/D963/27436360/gky963.pdf (accessed on 17 June 2020). [CrossRef] [PubMed]
11. Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2015**, *44*, D1075–D1079. [CrossRef] [PubMed]
12. Spain, V.V. Vidal Vademecum Spain, Su Fuente de Conocimiento Farmacológico, 2010. Available online: https://www.vademecum.es/ (accessed on 30 November 2019).
13. European Commission. European Commission—Centralised Medicinal Products for Human Use by ATC Code. 2020. Available online: http://ec.europa.eu/health/documents/community-register/html/reg_hum_atc.htm (accessed on 2 February 2020).
14. World Health Organization. World Health Organization, Anatomical Therapeutic Chemical Classification System. 2018. Available online: https://www.whocc.no/ (accessed on 2 February 2020).
15. World Health Organization. World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 2010. Available online: http://www.who.int/classifications/ (accessed on 2 February 2020).
16. Winkler, W.E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In Proceedings of the Section on Survey Research Methods, Anaheim, CA, USA, 6–9 August 1990; pp. 354–359.
17. FigShare. 2020. Available online: https://figshare.com/s/5b3128788640d7aa7d4f (accessed on 30 October 2020).
18. Smith, D.G. The 3 Most Promising Coronavirus Treatments, Explained. 2020. Available online: https://elemental.medium.com/the-3-most-promising-coronavirus-treatments-explained-752e2c6d54d7 (accessed on 30 September 2020).