# Supplemental File 12

## Introduction

This supplemental material describes two sets of methods; first, it briefly describes the process used to create the EPA’s LitDB database, and second, it describes how a subset of records were extracted from LitDB to be included in the reference chemical database RefChemDB.

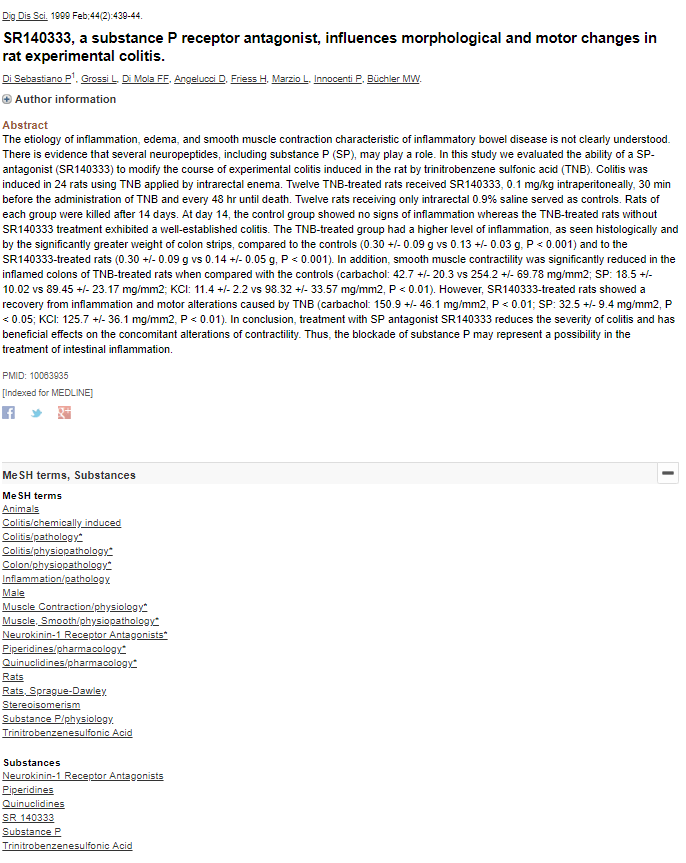
### LitDB construction

LitDB is a database of data elements extracted from the xml download of all MEDLINE PubMed records. Perl scripts are used to extract the identifying information from each citation record, information like title, abstract, authors and PubMed ID. Additionally, the MeSH (Medical subject heading) terms are extracted with subheadings (also known as qualifiers).

The Perl scripts extract to text files which are then loaded into a mysql database.

The MeSH heading and descriptor tree files are also downloaded into mysql tables. They are available at <https://www.nlm.nih.gov/mesh/filelist.html>.

To make the data more useful for research in chemicals, the data is passed stepwise through a series of algorithms. To illustrate the processing of the annotations, we will use PubMed ID 10063935 as an example. This article is entitled “SR140333, a substance P receptor antagonist, influences morphological and motor changes in rat experimental colitis”. The figure below shows the citation and the accompanying annotations of substances and MeSH terms provided by the PubMed website.



The initial database table entries for this article before the set of algorithms is run looks like the table below.

Table 1. Relevant annotations for PMID 10063935 before processing. (Note – this table is a simplified illustration and not the exact database format.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MeSH identifier | MeSH term / substance | Qualifier | Major heading flag |  |
| C085215 | SR 140333 |  |  |  |
| D010880 | Piperidines | pharmacology | Y |  |
| D011812 | Quinuclidines | pharmacology | Y |  |
| D064729 | Neurokinin-1 Receptor Antagonists |  |  |  |

In one set of algorithms, supplemental concept chemicals (any chemical with a MeSH identifier starting with “C”) are mapped to their MeSH headings (identifiers starting with “D”). The supplemental concept chemical then inherits the qualifiers assigned to its mapped-to parent in the article. In the case of our example article, the supplemental concept is *SR 140333*. The MeSH dictionary contains the information about the mapping of the concept to a bona fide MeSH heading. See <https://meshb.nlm.nih.gov/record/ui?ui=C085215> .

The supplemental concepts like SR 140333 will then inherit or be assigned the qualifier of the MeSH headings they are mapped to. After the inheritance steps, SR 140333 be associated with the qualifier *pharmacology* and the Y major heading flag as shown in Table 2.

In another process, coordinated pharmaceutical action MeSH terms will be processed and broken down into its components. Pharmaceutical action terms often describe that activity by specifying a target and the activity against that target. In our example citation, the coordinated MeSH term is *Neurokinin-1 Receptor Antagonists*. Our algorithms take the meaning of that D27 term and insert another row into the database that reflects that meaning in terms of a target MeSH heading and a qualifier that specifies the agonism or antagonism mode. In the following table one can see the newly added record in the last row.

Table 2. Relevant annotations for PMID 10063935 after processing. Italics show the additions to the database through the processing steps. (Note – this table is a simplified illustration and not the exact database format.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MeSH identifier | MeSH term / substance | Qualifier | Major heading flag |  | |
| C085215 | SR 140333 | *pharmacology* | *Y* | |  |
| D010880 | Piperidines | pharmacology | Y | |  |
| D011812 | Quinuclidines | pharmacology | Y | |  |
| D064729 | Neurokinin-1 Receptor Antagonists |  |  | |  |
| *D018040* | *Receptors, Neurokinin-1* | *antagonists & inhibitors* |  | |  |

After this series of algorithms, the resulting database contains more than 300 million records representing annotations in PubMed. A subset of these records is extracted to be pipelined to RefChemDB in steps described below.

### RefChemDB records from LitDB

The purpose of this set of steps is to extract only the records useful to the reference chemical application and, secondarily, to perform further entity mapping to retrieve CAS RN numbers and gene symbols where possible.

The entity mapping steps rely on cross-reference tables. The chemical mapping table MeSH uids and their corresponding CAS RN numbers. Note that there are some challenges in this mapping that must be taken into consideration. The MeSH chemical name is more general than the average CAS number; one MeSH name can encompass many salt forms of that chemical, for instance. In other words, there is not a simple one-to-one relationship between a MeSH chemical name and a CAS RN. Similarly, there is not always a straightforward one-to-one relationship between the MeSH term for a target and a gene symbol. Establishing these relationships in a database, therefore, is an inexact process that includes some computational steps and many hours of curation. Given the number of chemicals and the volume and pace of publications, we cannot claim that this information is complete and accurate.

The general steps in extracting from LitDB for inclusion in RefChemDB are:

1. A query is performed on the MeSH article annotation table to extract any protein/gene targets with the qualifier (subheading) *antagonists & inhibitors* OR *agonists*.
2. For the PubMed IDs identified in step 1, a second query is run to find any annotations of chemicals in those articles that are a) not proteins and b) annotated as a major topic and c) annotated with one of the following qualifiers: *pharmacology, administration & dosage, therapeutic use, toxicity, poisoning, adverse effects*.
3. For each protein/gene target, the gene id is looked up in the MeSH – gene cross reference table.
4. For each chemical, the CAS RN is looked up in the MeSH chemical – CAS table and insert that value if found.
5. Output the chemical – target combination to an Excel file with PubMed ID and mode flags (agonism or antagonism). An example is provided below.

Table 3. Sample output to RefChemDB

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chemuid | ChemName | Target | targetuid | Gene symbol | ag | antag | pmid | casrn |
| C085215 | SR 140333 | Receptors, Neurokinin-1 | D018040 | TACR1 | 0 | 1 | 10063935 | 153050-21-6 |