TO DO:

1. Review current process for working with PubMed via the Avillach method (https://github.com/OHDSI/KnowledgeBase/tree/master/LAERTES)
2. Read article recommending best practices (https://github.com/OHDSI/KnowledgeBase/issues/66)
3. One page summary of *specific* improvement ideas on current methods.

MEDLINE indexing problems

1. The *chemically induced* qualifier is not always necessary to denote ADEs.
2. The *adverse effects* qualifier is not the only qualifier to denote ADEs.
3. The ADE context is sometimes borne by a broader term rather than the drug of interest.
4. ADEs indexed with MeSH are skewed towards case reports.
5. MEDLINE does not record a relation between a drug and the manifestation of an adverse event.
6. MeSH descriptors sometimes conflate several drugs.
7. MEDLINE indexing rules sometimes aggregate multiple drugs under a broader MeSH descriptor.
8. Changes to MeSH have consequences on the retrieval of ADEs.
9. MEDLINE indexing is not always immediately available at publication time.

*Defining an ADE: What Is Necessary and What Is Sufficient?*

Edwards and Aronson (2000) defined an adverse drug reaction (ADR) as "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrant prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product". Unfortunately, while *ADR* is a relatively well defined term, *adverse drug event* (*ADE*) is not. However, attempts have been made to solidify a definition, most notably the FDA and WHO, who define an ADE as "any untoward medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment".

But this definition causes a few more problems than it offers solutions, namely that an *ADE* might not have a causal relationship whatsoever. The MeSH (Medical Subject Headings, National Library of Medicine) thus includes a few methods for bypassing these issues. In line with the diverse array of possible headings for ADEs, MeSH lists a few "clues" for finding ADEs.

The first is the subheading *chemically induced* which, as the scope note points out, is "used for biological phenomena, diseases, syndromes, congenital abnormalities, or symptoms caused by endogenous or exogenous subsances". Current work suggests that the *chemically induced* qualifier is not always necessary to denote ADEs. This seems somewhat odd at first, namely because our above definition of ADEs mentions "pharmaceutical product[s]" which would imply the usage of a chemical, however, certain natural language descriptors such as *Drug-Induced Liver Injury* or *Drug-Induced Abnormalities*, etc., would be missed because they are not filed as *Liver Diseases/chemically induced* or *Abnormalities/chemically induced*.

Furthermore, the *adverse effects* (*AE*) qualifier (scope note: "used with drugs, chemicals, or biological agents in accepted dosage – or with physical agents or manufactured products in normal usage – when intended for diagnostic, therapeutic, prophylactic, or anesthetic purposes. It is used also for adverse effects or complications of diagnostic, therapeutic, prophylactic, anesthetic, surgical, or other procedures, but excludes contraindications for which "contraindications" is used") is most certainly not the only qualifier for which ADEs can be obtained.

Starting with these two problems, there seem to be two possible solutions. The first and easiest would seem to be to simply dig through the list of MeSH headings and subheadings for qualifiers and descriptors seen as relevant. The second would be a more complicated machine learning algorithm which would be able to determine intuitively which keywords or phrases relate to ADEs. Current work suggests a combination of efforts, with the first method being used in concordance with *pre-coordinated MeSH descriptors* as well as the *adverse effects* qualifier and its pertinent subheadings and the second method being utilized for "indirect links between drug entities and broader drug descriptors bearing the ADE context". This is a marked improvement over the previous Avillach, *et al*. (2012) method because that methodology only considers ADEs that are marked and indexed as such. Avillach, *et al*. even hinted at the burgeoning of a machine learning-type possibility, saying that "because specific subheading and keywords are used in the queries that are automatically built, the automated search may be more specific than a manual query".

*Identification of Other Problems*

In addition to the above issues, it should be noted that many drugs are represented only as SCRs (supplementary concept records) and so cannot be accessed directly via a qualifier. Luckily, all SCRs have a *Heading Mapped* criterion which relates to at least one MeSH descriptor. Best practices currently suggests linking these SCRs (via qualifiers; i.e. SCRs are mapped to descriptors which have qualifiers). Of such descriptors, a pertinent one seems to be *pharmacological action* (*PA*).

Furthermore, with current MeSH headings, there is a bias towards case reports involving specific clinical trials because those trials tend to outline safety profiles and are thus easily annotated and indexed. However, this bias can overshadow more conspicuous results and give the illusion of a full-drug profile even if one does not exist.

As mentioned earlier, for most ADEs a qualifier/descriptor pair (with qualifiers such as *adverse effects* or *chemically induced*) is required in order to locate an ADE. Unfortunately this leads to two problems: (1) no direct record of this relationship actually exists in the database and (2) the presence of both the qualifier and descriptor does not necessarily mean that there is a direct link between the two.

Conflation of several drugs, especially stereoisomers, is also an issue because the two may have completely different effects or potential ADEs. Aggregation of several drugs is also a problem, in that a study involving several related drugs might be indexed under a group name instead of individually, leading to an inaccurate potential ADE.

And, as if these weren't problematic enough, MeSH is not a static organism––its headings and subheadings are re-issued and changed somewhat often, but previous indexes are not always updated. This also leads into the next problem: MEDLINE citations with MeSH indexing are not always available at the time an article is published. They may be incomplete or missing for days to weeks, depending on the journal.

*Independent Generalized Solutions*

MeSH terms are incredibly useful and diverse, but difficult to quantify into a category like *ADE* directly. Two methodologies exist, and the newer recommended method will be discussed shortly, but independent of that, I would like to develop a machine learning methodology.

In order to accurately reflect the existence of an ADE, a computer needs to take into account several possibilities. As discussed earlier, an ADE is "any untoward medical occurrence that may present during treatment with a pharmaceutical product but which does not necessarily have a causal relationship with this treatment".

So our hypothetical program needs to be able to learn what an *untoward medical occurrence* is and determine if it is related in any way to a *pharmaceutical product*.

An *untoward medical occurrence* is either an unexpected and inappropriate or inconvenient event or incident that pertains in some capacity to medicine.

A *pharmaceutical product* is simply any substance or combination of substances presented as having properties for treating or preventing disease in human beings or animals; or that may be used in, or administered to, human beings or animals, either with a view to restoring, correcting or modifying physiological functions by exerting a pharmacological, immunological or metabolic action; or making a medical diagnosis.

An untoward medical occurrence can thus be determined by comparing a pharmaceutical product's intended purpose(s) to the observed result. Thanks to pharmaceutical databases such as RxNorm and (more or less) standard terminology, it can be relatively easy to find the normal effects of a drug, its ingredients and chemical makeup, and its known side-effects.

For medical occurrences, MeSH does not provide an outstanding, all-inclusive effort. However, the ICD-10 (International Classification of Diseases) includes diseases and related health problems (including, but not limited to injuries, symptoms, and signs) in 43 different languages.

Now our hypothetical program has a list of pharmaceutical products and their intended purposes, as well as a list of diseases and related health problems (which will be considered *untoward medical occurrences*).

The program will run through two stages of analysis with the first being a MeSH heading/subheading search for (1) any condition or health problem in ICD-10 and any pharmaceutical (or pharmaceutical ingredient) and then (2) a second pass through the literature's actual text (if available) for the same information. The two lists will be matched and ranked via a corresponding measure of relationship. Finally, all matches including the positive effects of said products (i.e. a cancer drug producing positive cancer treatment) will be eliminated.

This should eliminate most, if not all, of the aforementioned issues, especially if automatic updating is done for the reference databases (most of which have already been loaded in one form or another into LAERTES).

Unfortunately, such a large machine learning system is not entirely feasible within certain constraints and should be further discussed.

*Compromise and Solution*

With a proposal for a better solution already existent, the above more permanent solution can be sidelined for the time being. Thus the following procedure should be attempted:

1. Query for a *Chemical and Drugs Category* drug or chemical term which will be restricted with a known qualifier such as *adverse effects*.
2. Query for an unbound *chemically induced* qualifier (the query should manually include the pre-coordinated MeSH descriptors, which will be discussed further).
3. If an article fulfills at least one "drug facet" (1) and at least one "manifestation facet" (2), then it is captured and kept as a potential ADE.

Current work suggests utilization of the NCBI Entrez Programming Utilities in order to obtain PubMed IDs for MEDLINE records (XML).

1. The results of (1) through (3) should then be able to index the potential ADEs and ADEs amongst the MEDLINE citations gathered.
2. Once a citation is gathered, all indexed descriptors located at *Chemicals and Drugs* are considered "drug candidates" (in addition to SCRs). SCR processing is performed to accommodate further possibility as a more specific potential ADE.
3. ADE pairs are constructed based on filtering from (4) and (5), then classified as either involved or concomitant based on outlined criteria.
4. Clinical significance is another added filter, done by RxNorm mapping.
5. Finally, metadata is extracted which could be relevant to the proposed ADE (MeSH tags such as gender, age groups, publication types, epidemiologic methods, etc.).

It should be noted that because this is not an accomplished machine learning solution, that it presents several limitations that were also present in the previous Avillach method; namely, it does not compare or contrast different ADE vocabularies such as MedDRA or systems which are completely outside of a terminology (a text mining method could crawl the entire internet, finding relationships in social media or, with extended language support offered by ICD-10, results in papers published in different languages or those that are totally unindexed). A UMLS (Unified Medical Language System) Metathesaurus solution has been proposed, but not fleshed out in full-capacity. Furthermore, using RxNorm as the only standard of comparison may eliminate trials or effects of components of drugs that may not otherwise be detected (similarly, unregistered or drugs in testing phases may also be eliminated).

*Implementation over Previous Standard*

The file "queryDrugHOIAssociations.psql" currently implements what has been referred to as the Avillach method. It first copies data between a MEDLINE database image which had been previously loaded into a PostgreSQL DBMS and a newly created file with the recommended name "drug-hoi-associations-from-mesh.tsv".

It does so by selecting data (namely a PubMed ID, descriptor name, and descriptor name UI) from the aforementioned DBMS (specifically the MeSH headings) and joining the relevant qualifiers where the MeSH heading PubMed ID equals the qualifier PubMed ID and where the MeSH heading medical citation MeSH heading list MeSH heading order is also equal to the qualifier's likewise named column.

This is done twice. First into a "drug\_of\_ade" WITH clause and second with an "effect\_of\_ade" clause. The main difference being that the qualifier value in the former must be "adverse effects" while the second should be "chemically induced".

From these newly selected fields, an ADE profile is built (including PubMed ID, drug, drug UI, effect, effect UI, publication type value, and publication type UI). Finally, this profile is added as an "ADE" to a table and sent to the STDOUT.

While this method is quick and useful, it runs into all of the aforementioned problems inherent to MeSH and MEDLINE.

The following is a brief outline of my suggestions for a revamped system, based on a non-machine learning approach.

1. Assuming that the MEDLINE database image is constant, first query for all *Chemical and Drugs Category* drugs and chemical terms which have the qualifier *adverse effects* (this seems like it has already been done, but could possibly be revamped using a different set of standards).
2. Query for all *chemically induced* qualifiers (this has also already been done, however, the following pre-coordinated MeSH descriptors should be added as a "manual" option). I personally consider this to be a "patch" rather than an actual solution to the problem, especially since MeSH is not particularly static and there could be other relevant implicit "chemically induced" qualifiers.
   1. Drug-Induced Liver Injury
   2. Drug Eruptions
   3. Drug-Induced Liver Injury, Chronic
   4. Erythema Nodosum
   5. Serotonin Syndrome
   6. Hand-Foot Syndrome
   7. Stevens-Johnson Syndrome
   8. Neuroleptic Malignant Syndrome
   9. MPTP Poisoning
   10. Dyskinesia, Drug-Induced
   11. Neurotoxicity Syndromes
   12. Psychoses, Substance-Induced
   13. Akathisia, Drug-Induced
   14. Anticholinergic Syndrome
   15. Acute Generalized Exanthematous Pustulosis
   16. Asthma, Aspirin-Induced
   17. Drug Hypersensitivity Syndrome
   18. Chemotherapy-Induced Febrile Neutropenia
   19. Abnormalities, Drug-Induced
3. From all selected articles, compare both tables and keep those PubMed IDs which fulfill at least one "area" of the above qualia.
4. These should then be gathered into an "ADE" or rather a potential ADE-type table with relevant citations (which also appears to have already been completed).
5. From this potential table (or rather a "drug candidates" table), all descriptors which belong to the *Chemicals and Drugs* should be explored further, in addition to all SCRs. These should form a larger table of ADE pairs post-SCR processing.
6. Each created pair on this new list should then be classified as either "involved" (if the SCR inherits *adverse effects* from its descriptor and the PA descriptor) or "concomitant" (if it does not display an ADE context in any form or fashion).
7. The next round of filtering should be done with either a loaded RxNorm table or the appropriate RxNorm API call, in which the MeSH descriptor and SCR of a "potential ADE" can be mapped to an RxCUI. Thus, clinical relevance of drug (which can be normalized via an IN or PIN mapping from RxCUI) can be established if the ingredients are associated with at least one SCD in RxNorm.
8. With this final list, relevant information can be extracted from MeSH/MEDLINE. While the current system does do this to a certain extent, it does not seem to capture pertinent metadata such as *gender*, *age groups*, or *epidemiologic methods*.
9. This new final table can then be sent to STDOUT as it currently done.

The largest disadvantage to this new implementation is certainly time and scalability. With manual "overrides" in check to make sure things don't slip pass, there is a huge potential for the program to execute an order of magnitude or so slower (especially since the amount of information to sift through is projected to grow exponentially.