W266 Final Project Milestone:
Evaluating Food-Drug Interactions with ACE Inhibitors and Common Food Items

Adam Lenart, Hyera Moon, Lisa Barceló

Abstract: Using the PubMed API, we looked at 50,000 abstracts containing our drug keyword "ACE" or "ACE inhibitor", and filtered out relevant sentences containing common food components. We used a random predictor baseline to label each of the resulting 300,000 sentences based on the interaction between the food item and the drug, assigning a label of *positive*, *negative,* or *neutral* with equal probability. A variety of models will be used to compare outcomes on the corpus. One thousand manually labeled sentences will be used to evaluate to train and test these models. A sentiment lexicon will first be used to test a model that labels sentences based on the presence and frequency of words categorized as *positive*, *negative,* or *neutral*. Next, we will use an LSTM that has been trained on the manually labeled sentences.

Keywords: *Angiotensin II receptor blockers; ACE inhibitors; food-drug interactions; NLP; Natural Language Processing; Information Extraction*

Introduction

An estimated 40 million Americans[1] use prescription medication to control hypertension, and the majority are also juggling prescriptions for other medications to manage additional diseases. Medication management is a healthcare best practice, and patients and providers alike are educated on the risks of drug-drug interactions that can occur when managing medications for a variety of comorbidities. Less emphasized, however, are the risks for drug-nutrient interactions. Our team aims to use select neurolinguistic processing tools to elucidate relationships between specific drug classes (i.e., ACE inhibitors) used to treat hypertension, and nutrients found in common food items, whether they are naturally occurring (such as isoflavonoids and plant estrogens) or synthetic additions (i.e., monosodium glutamate and bisphenol-A).

[1] https://www.cdc.gov/bloodpressure/facts.htm

Background

Scientific literature contains much work on the elucidation of drug-drug interactions, which has been critical in pharmaceutical research. As the number of people on multiple types of medication continues to increase, it will continue to be pertinent. Almost one in four Americans[2] used 3 or more prescription drugs in the last 30 days. Less emphasized are the risks of certain food compounds with prescription drugs. Some well-known examples exist, such as the relationship between grapefruit juice and certain cholesterol-lowering drugs like lovostatin[3,4]. But of the hundreds of food particles humans consume per day, are there other compounds that are interacting with your blood pressure medication, for example? Our preliminary results suggests there are.

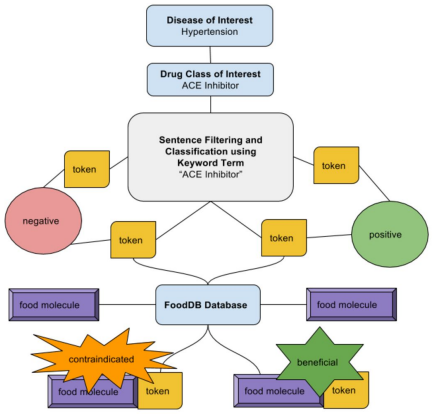

Figure 1. Data workflow.

[2] https://www.cdc.gov/nchs/fastats/drug-use-therapeutic.htm

[3] https://www.ncbi.nlm.nih.gov/pubmed/9585793

[4] http://www.health.harvard.edu/heart-health/grapefruit-juice-and-statins

Methods

We did some initial data exploration of the pubmed corpus and found over 10 million tokens. The most frequent words can be shown in Figure 2.

	Most frequent words			
rank	Without stopwords		Including stopwords	
1	patients	82302	the	391549
2	ACEI	72976	of	360643
3	treatment	37852	and	304381
4	blood	35821	in	267195
5	pressure	34732	to	153961
6	p	33444	with	139923
7	inhibitors	28649	a	115994
8	renal	27998	was	89689
9	heart	26985	patients	82302
10	captopril	23806	were	75261
11	group	23798	ACEI	72976
12	study	23671	for	67000
13	ii	23643	by	57982
14	angiotensin	22946	is	54093
15	hypertension	22857	or	53074

Figure 2. Most frequent words in Pubmed corpus.

For our baseline, we chose to study only one particular drug, an ACE inhibitor. Using the PubMed API, we downloaded 50,000 abstracts containing the keyword “ACE” or “ACE inhibitor”. We parsed the abstracts with the NLTK Python library, reducing the abstracts into sentences containing our key word only, and ignored the rest--this left us with about 300,000 sentences. Our database of common foods and the components came from FooDB.ca. We created a dictionary of foods with their components (i.e., {milk: ‘whey’,’casein’}), and exported as pickle files. We have also received access to the MIMIC II data set from MIT, which we will use to test our model in the future.

Because of the large volume of data processing, we utilized PySpark, and created an RDD of the filtered sentences. Within each sentence, we identified the number of words comprising the food name (i.e., if "goat milk" was listed as a food item in the database, we noted that it contained n = 2 words, and we would search for the given n-gram in the sentence). This allowed for less vagueness when comparing items like "goat milk" with "milk". In order to look at similarities between words, we used the Jaro-Winkler distance to compare.

Regarding sentence classification, our baseline model will simply use random prediction, assigning labels of *positive*, *negative*, and *neutral* with equal probability. We have also collected 1,000 random sentences from our original corpus and are manually labeling the sentences. These will then be used to test our model and improve accuracy.

Results and discussion

One of the challenges with our dataset is that we saw some overlap between some common food names and non-food-related words in the scientific corpus! For example, filtering on the food "pie" initially included sentences containing the word "therapies". The Jaro-Winkler string matches were helpful in resolving this issue.

Another problem we are currently trying to evaluate is incongruencies with word contexts. The food item "date" has yet to be distinguished from the "date" of an experiment, for example. Time will tell whether this is a significant problem, or one that can be addressed with a simple method such as entity recognition.

Our random predictor baseline has been recommended as a good comparison against which we can measure our resulting model. We will need to evaluate whether or not the model is better than chance at predicting the classification of the sentence.

Next Steps

Now that we have figured out the baseline method, we will be testing a simple sentiment analysis using a sentiment lexicon. This looks at the presence and frequency of positive or negative words in the sentence to determine the sentiment. This is not as sophisticated as the most modern sentiment analysis algorithms, as it does not always take context into account. A sentence like "Food X increases ACE inhibition activity" could be interpreted as a positive interaction, although the food is really exerting a negative effect on ACE, and should be labeled as such.

As one of our final models, we will use an LSTM to classify the filtered sentences as positive, negative or neutral. The LSTM will be trained with about 1,000 labeled sentences.

Appendix

We have uploaded some relevant code into our shared github folder[5], including:

1) **downloadPubmed.ipynb**, to download abstracts from PubMed's API
2) **compound_food_id.ipynb**, which is where we build the food dictionaries
3) **baseline_v2.ipynb**, the notebook that filters our sentences and applies the random model
4) **Filtered_Sentences.txt** which is the results of our filtered sentences

5 https://github.com/adamlenart/food_drug_interaction/tree/master/progress_submission