Yael Ozair
Mentor: Dr. Olivier Bodenreider
Computational Analysis
CgSB LHNCBC NLM NIH
2017 Internship Report
30 July 2017

My project this summer entailed working on research that would assess the risk of prescription drug use during pregnancy. Previously, the old FDA standards for drug risk recommendation categories made it difficult for the assessment of risk-benefit ratios for pregnant women and their fetus, and has also been a challenge to healthcare providers in decision making, due to lack of uncertainty from the categories. As an attempt to improve this, the new FDA recommendation requires risk severity in one sentence with level of human evidence. To facilitate research, prior work by members from the Cognitive Science Branch at the National Library of Medicine, Lister Hill National Center for Biomedical Communications, used ingredient-level RxNorm concept unique identifiers (RxCUIs) and mapped the ingredients to their respective level of risk and evidence. This information was provided by Briggs' reference textbook on evidence and risk of drugs in pregnancy and lactation.

To conduct the research on real data, the Innovation in Medical Evidence Development and Surveillance (IMEDS) data cloud had provided authorized users to use its collection of public and private insurance claims data of patients in a controlled and standard format using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). To extract the data without exposing patient-level information, the researchers produced counts of patients by categories of risk and level of evidence and other demographic relevant data from the database. Pregnancy cases were identified through a delivery code, and drugs dispensed 270 days prior delivery were analyzed, as 270 days is the average length in pregnancy in humans. The rationale for this work is that decision-making in care can improve by the available evidence, provide potentially new knowledge about risk of drugs in pregnancy, as well as provide potential findings that suggest differences between populations covered by public and private insurance data. This could ultimately help improve the quality of care

Last year Dr. Olivier Bodenreider and others at the NLM produced a preliminary investigation of this assessment as a podium abstract presented at the American Medical Informatics Association (AMIA) Annual Symposium. Subsequently, an upcoming publication is underway, and I have contributed to this research in the form of computational analysis.

The objective for the investigation includes characterizing prescription drugs dispensed during pregnancy according to the new FDA recommendations, in terms of level of risk and type of evidence (human evidence for example). For my contribution, I investigated some differences between the two types of insurances, stratified by pregnancy period and age. During my time in Bethesda, I was able to filter the data in such a way that it would provide ways for other researchers to quickly compute the statistics in determining differences among age groups and different pregnancy periods. Specifically, I was able to perform computational analysis upon the data that was during pregnancy, across both insurance claims datasets (public vs. private).
To be able to conduct this research, I inherited aggregated data produced by these researchers. The data was prepared as counts of beneficiaries exposed to prescription drugs in different categories of age groups, pregnancy periods, and drugs with level of risk and evidence.

During my time here I learned R, as the data aggregation I obtained from my supervisor was available for computation in the form of R data files. Aggregations were formulated in such a way that allowed computation without exposing patient-level information. I was able to learn the R language pretty quickly and produce descriptive and inferential statistics from the datasets. I found that R Studio is a great environment and is pretty intuitive. The packages available allow for a speedy data exporting and importing process, as well as convenient ways to tidy up and transform the data, which also reduced the time to compute such statistics.

The methods to conduct the investigation included identifying the most frequently prescribed drugs during pregnancy between both insurance datasets. To do this, I used descriptive statistics to compute the relative frequency of each prescription drug in both datasets using the number of beneficiaries exposed to that drug at least once during pregnancy. I then compared the frequencies of each prescription drug across both datasets that made it in the top 20 ranked drugs for each set, and did this using the inferential statistics test, known as, 'Comparison of Two Proportions from Independent Populations'. Lastly, I compared the ranks of the top 20 drugs. To compare the rankings between each dataset, I used the Wilcoxon Rank Sum Test to determine how similarly ranked the drugs are across both sets.

I first produced two tables that were considered the top 20 drugs based on the proportion of beneficiaries that were exposed to that drug at least once during pregnancy, and this was for each set. I then consolidated the tables to determine which drugs were common between both top 20 drugs, and found 14 are common between both (Table 1).

I then compared the frequencies using the proportions test and found that *Terbutaline* was the only drug frequency not significantly different between both sets (Table 2). The test uses the absolute frequency of beneficiaries exposed to a drug and those not, by subtracting exposed from the total and did this is for both datasets (Table 3). The R statistics package provides this test, and implemented as, *prop.test()*, which facilitated the computation by taking the values described above as the parameters (Table 3). To add, correcting for multiple tests was not necessary as each test is independent from one another. The next task was to rank the drugs based on the proportions of beneficiaries exposed to each drug. It was found that *Azithromycin* is ranked quite similarly and the one highlighted in orange is ranked a bit differently (Table 4). I used the Wilcoxon rank test to determine how similar the two datasets are ordered and this is done globally between both sets. This test in nonparametric, meaning that it uses the medians rather than mean and standard deviation to provide insight on the distribution of the sets of ranked drugs (Figure 1). The boxplot shows that the medians are similar and this further solidified my results that suggested the hypothesis was correct—that there is no significant difference in the order of the drugs across the datasets (Figure 2). This was determined, as the p-value computed from the test is greater than the significant level of 0.05.

This is only a small part of what we need to do, as there are many other tasks like assessing risk and evidence. I found it to be a steep learning curve for using R on such complex datasets. The datasets were complex to learn, as those who aggregated them were either out of the country and difficult to reach or working on other projects. Therefore, there is a lot of work left to do and I definitely plan to continue this work. This is my first experience conducting research and computation for a prospective publication and it's such a privilege, as this experience has only furthered my ability to conduct a scientific investigation. I have definitely found this helpful to my career, as well as myself in general, as it is a goal of mine to become a biomedical researcher and contribute to medical knowledge.

Table 1. Top 20 drugs for each dataset, Private vs. Public Insurance

Commercial Insurance

	RXN_ID	RLABEL	PRCNT	INSR_PRCNT
1	18631	Azithromycin	7.929367	CCAE_PRCNT
2	26225	Ondansetron	7.182582	CCAE_PRCNT
3	723	Amoxicillin	5.632606	CCAE_PRCNT
4	8745	Promethazine	4.990360	CCAE_PRCNT
5	214182	Acetaminophen / Hydrocodone	3.046541	CCAE_PRCNT
6	2231	Cephalexin	3.026764	CCAE_PRCNT
7	10582	Thyroxine	2.584257	CCAE_PRCNT
8	6922	Metronidazole	2.347959	CCAE_PRCNT
9	8727	Progesterone	2.257703	CCAE_PRCNT
10	4450	Fluconazole	2.241382	CCAE_PRCNT
11	73645	valacyclovir	2.046693	CCAE_PRCNT
12	10368	Terbutaline	1.961860	CCAE_PRCNT
13	19711	Amoxicillin / Clavulanate	1.938946	CCAE_PRCNT
14	817579	Acetaminophen / Codeine	1.811121	CCAE_PRCNT
15	6915	Metoclopramide	1.421372	CCAE_PRCNT
16	36437	Sertraline	1.259751	CCAE_PRCNT
17	39993	zolpidem	1.225248	CCAE_PRCNT
18	7417	Nifedipine	1.225106	CCAE_PRCNT
19	6902	Methylprednisolone	1.031569	CCAE_PRCNT
20	214183	Acetaminophen / Oxycodone	1.008567	CCAE_PRCNT

Public Insurance

	RXN_ID	RLABEL	PRCNT	INSR_PRCNT
1	8745	Promethazine	7.302233	MDCD_PRCNT
2	18631	Azithromycin	6.794953	MDCD_PRCNT
3	6922	Metronidazole	6.129653	MDCD_PRCNT
4	26225	Ondansetron	5.117727	MDCD_PRCNT
5	723	Amoxicillin	4.897023	MDCD_PRCNT
6	214182	Acetaminophen / Hydrocodone	4.070242	MDCD_PRCNT
7	2231	Cephalexin	3.745473	MDCD_PRCNT
8	817579	Acetaminophen / Codeine	3.507754	MDCD_PRCNT
9	4450	Fluconazole	3.030999	MDCD_PRCNT
10	9143	Ranitidine	2.368291	MDCD_PRCNT
11	10368	Terbutaline	1.954115	MDCD_PRCNT
12	6915	Metoclopramide	1.697193	MDCD_PRCNT
13	5553	Hydroxyzine	1.497858	MDCD_PRCNT
14	19711	Amoxicillin / Clavulanate	1.385148	MDCD_PRCNT
15	10831	Sulfamethoxazole / Trimethoprim	1.256124	MDCD_PRCNT
16	73645	valacyclovir	1.237070	MDCD_PRCNT
17	214183	Acetaminophen / Oxycodone	1.216784	MDCD_PRCNT
18	71722	Docusate Sodium	1.174916	MDCD_PRCNT
19	82001	Docusate Calcium	1.174916	MDCD_PRCNT
20	82002	docusate potassium	1.174916	MDCD_PRCNT

Table 2. Comparison of top 20 drugs between both data sets, using the comparison proportion test

	RLABEL	Commercial	Public	pval	is_sig
1	Azithromycin	7.92936701	6.7949527	0.000000e+00	TRUE
2	Promethazine	4.99036038	7.3022328	0.000000e+00	TRUE
3	Ondansetron	7.18258245	5.1177274	0.000000e+00	TRUE
4	Metronidazole	2.34795872	6.1296535	0.000000e+00	TRUE
5	Amoxicillin	5.63260574	4.8970231	0.000000e+00	TRUE
6	Acetaminophen / Hydrocodone	3.04654077	4.0702422	0.000000e+00	TRUE
7	Cephalexin	3.02676356	3.7454733	0.000000e+00	TRUE
8	Acetaminophen / Codeine	1.81112092	3.5077542	0.000000e+00	TRUE
9	Fluconazole	2.24138154	3.0309989	0.000000e+00	TRUE
10	Thyroxine	2.58425738	0.4652846	0.000000e+00	TRUE
11	Ranitidine	0.70439470	2.3682911	0.000000e+00	TRUE
12	Progesterone	2.25770305	0.3781925	0.000000e+00	TRUE
13	valacyclovir	2.04669300	1.2370700	0.000000e+00	TRUE
14	Terbutaline	1.96186011	1.9541147	3.717263e-01	FALSE
15	Amoxicillin / Clavulanate	1.93894620	1.3851479	0.000000e+00	TRUE
16	Metoclopramide	1.42137168	1.6971928	9.531952e-281	TRUE
17	Hydroxyzine	0.79378202	1.4978580	0.000000e+00	TRUE
18	Sertraline	1.25975148	0.8405245	0.000000e+00	TRUE
19	Sulfamethoxazole / Trimethoprim	0.84912630	1.2561241	0.000000e+00	TRUE
20	zolpidem	1.22524770	0.8222139	0.000000e+00	TRUE
21	Nifedipine	1.22510593	1.0969578	1.911289e-81	TRUE
22	Acetaminophen / Oxycodone	1.00856676	1.2167839	9.855943e-224	TRUE
23	Docusate Sodium	0.01081012	1.1749159	0.000000e+00	TRUE
24	Docusate Calcium	0.01081012	1.1749159	0.000000e+00	TRUE
25	docusate potassium	0.01081012	1.1749159	0.000000e+00	TRUE
26	Methylprednisolone	1.03156928	0.4475901	0.000000e+00	TRUE

Table 3. The parameters used in the proportion test method in the R statistics package

Azithromycin	**Exposed**	**Not Exposed**	**Total**
Commercial Insurance	P_A	T_A-P_A	T_A
Public Insurance	P_B	T_B-P_B	T_B

Table 4. The top 20 drugs ranked in both datasets, blue border represents similar ranking, and orange border represent different ranking

Commercial Insurance			Public Insurance		
	RLABEL	RANK		RLABEL	RANK
1	Azithromycin	1	1	Promethazine	1
2	Ondansetron	2	2	Azithromycin	2
3	Amoxicillin	3	3	Metronidazole	3
4	Promethazine	4	4	Ondansetron	4
5	Acetaminophen / Hydrocodone	5	5	Amoxicillin	5
6	Cephalexin	6	6	Acetaminophen / Hydrocodone	6
7	Thyroxine	7	7	Cephalexin	7
8	Metronidazole	8	8	Acetaminophen / Codeine	8
9	Progesterone	9	9	Fluconazole	9
10	Fluconazole	10	10	Ranitidine	10
11	valacyclovir	11	11	Terbutaline	11
12	Terbutaline	12	12	Metoclopramide	12
13	Amoxicillin / Clavulanate	13	13	Hydroxyzine	13
14	Acetaminophen / Codeine	14	14	Amoxicillin / Clavulanate	14
15	Metoclopramide	15	15	Sulfamethoxazole / Trimethoprim	15
16	Sertraline	16	16	valacyclovir	16
17	zolpidem	17	17	Acetaminophen / Oxycodone	17
18	Nifedipine	18	18	Docusate Sodium	18
19	Methylprednisolone	19	19	Docusate Calcium	18
20	Acetaminophen / Oxycodone	20	20	docusate potassium	18

Figure 1. This box plot represents the actual distribution of the ranks of drugs globally for both datasets (the medians are represented as the bold lines that divide each box). Box ‘1’ represents the private insurance data set, while ‘2’ represents the public.

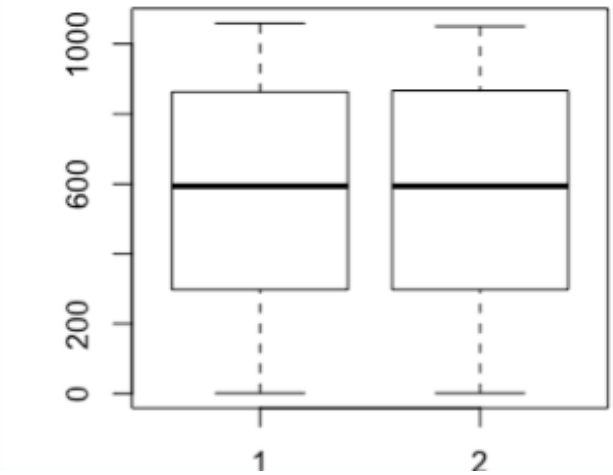

Figure 2. Each bullet point is a result of the Wilcoxon Rank Sum Test, the p-value is greater than the significance level of 0.05.

- W = 716000
- **P-value = 0.4904**
- 95% confidence interval:
  - -20.99996 23.00002
- Sample estimates in location
  - 7.999986
- **P-value > alpha : H_0 = True & H_A = False**