

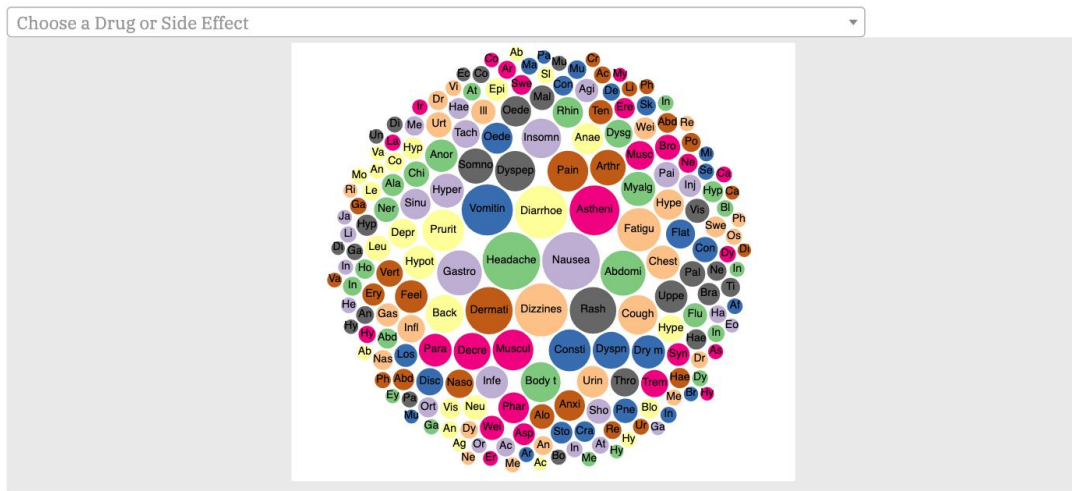
CS 480x Process Book

Suryansh Goyal, Jean-Philippe Pierre, Remy
Allegro, Joshua McKeen



Drug Side Effects Visualization

CS480x Final by Joshua McKeen, Suryansh Goyal, Jean-Philippe Pierre, and Remy Allegro



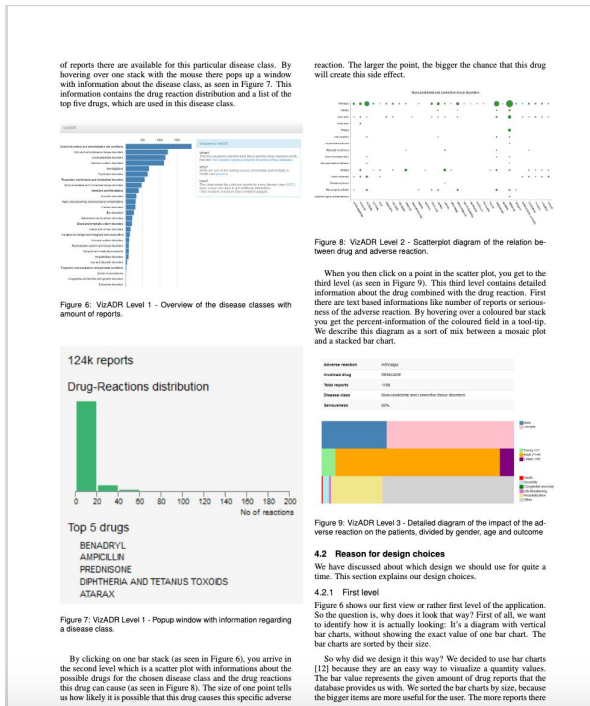
Our project uses the SIDER 4.0 dataset to visualize how many drugs cause a certain side effect, which side effects are most prevalent among drugs, and the frequency of a certain drug's side effects.



Motivation

Upon finding the SIDER 4.0 dataset, we thought it would be interesting to visualize the side effects of drugs. We thought that visualizing side effects would be worthwhile because many people take medicines, and as a result, side effects affect them personally. Thus, people would be interested in knowing more about the side effects in medicine. We also were curious to know what the most common side effects were, and how frequencies of these side effects varied among drugs.

Related Work



“VizADR - Visualizing Adverse Drug Reactions” by Pfunder and Kuric

This work provided us with insight on visualizing different qualities of a dataset within a single visualization tool. It visualizes adverse drug reactions by transitioning between 3 levels of visualizations. The first level is a bar chart showing the amount of adverse reaction reports for each drug disease class. Hovering over its bars shows a sorted bar chart visualizing the amount of reports for each drug, in descending order. The second level shows a scatter plot that displays the likelihoods of certain side effects occurring in the presence of drugs in a certain disease class. The third level uses a stacked bar chart to visualize the distributions of gender, age, and side effects in adverse reaction reports for a given drug. The user can go from one level to another by interacting with the visualization.



Questions

We wanted our visualization to attempt to answer the following questions about the dataset:

- Which side effects are the most prevalent across the entire dataset?
- Given a particular side effect, what are the drugs with the highest frequencies of that side effect?
- For a particular drug, what are the side effects that occur most frequently?
- How wide is the range of frequencies for a given drug - side effect pair?
 - How far apart are the upper and lower bounds of the frequency for a given drug - side effect pair?



Data

SIDER 4.1


Home

Drug list

Side Effects

Download

About



SIDER 4.1 : Side Effect Resource

SIDER contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. The available information include side effect frequency, drug and side effect classifications as well as links to further information, for example drug-target relations.

Search for drugs or side effects :

type 3 or more characters...

Database statistics

Number of drugs and side effects

# of SE	# of drugs	# of drug-SE pairs	Pairs with frequency information
5868	1430	139756	39.9%

Number of drug-side effect pairs in different frequency ranges

frequency	infrequent	common	post-marketing	total

- The dataset used for this visualization was the SIDER 4.0 dataset. It contains side effects for numerous drugs and their likelihoods.



Data

Our visualization used two files from the dataset:

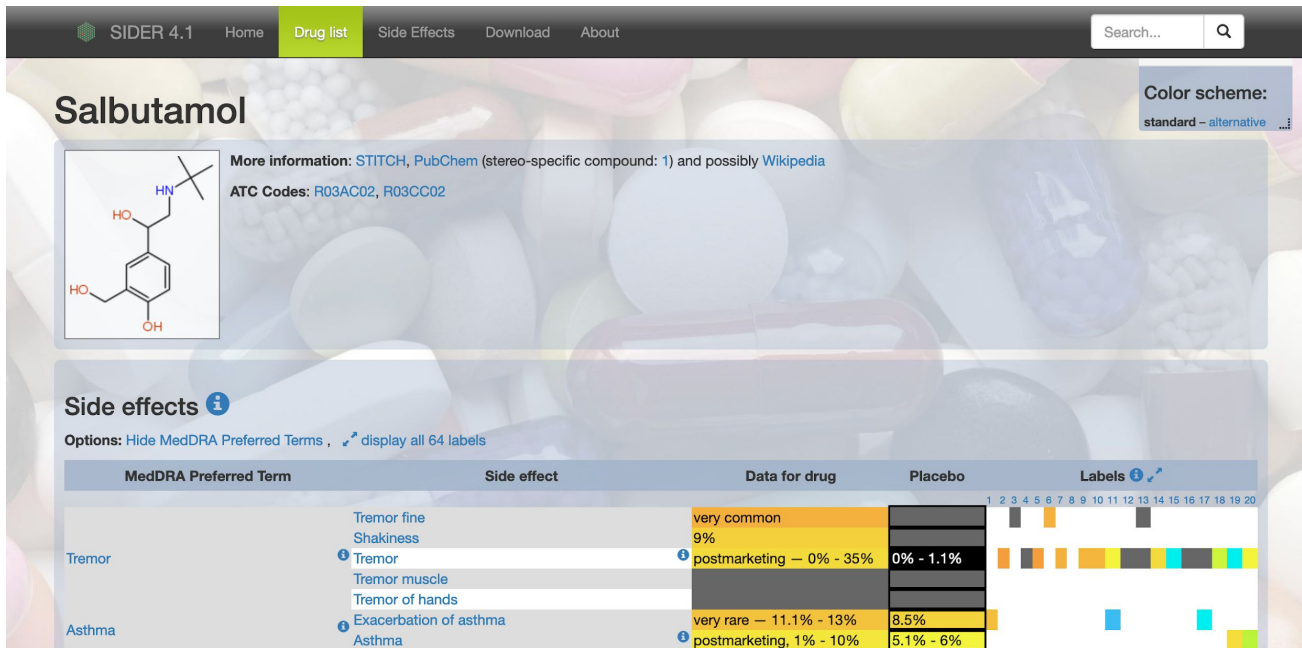
Stitch CID1	Stitch CID2	Umls IDLabel	Placebo	Frequency De	Frequency Lo	Frequency Up	Meddra Type	Umls IDMedd	Side Effect
CID100000085	CID000010917	C0000737		21%	0.21	0.21	LLT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		21%	0.21	0.21	PT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		21%	0.21	0.21	PT	C0687713	Gastrointesti
CID100000085	CID000010917	C0000737		5%	0.05	0.05	LLT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		5%	0.05	0.05	PT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		5%	0.05	0.05	PT	C0687713	Gastrointesti
CID100000085	CID000010917	C0000737		6%	0.06	0.06	LLT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		6%	0.06	0.06	PT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		6%	0.06	0.06	PT	C0687713	Gastrointesti
CID100000085	CID000010917	C0000737		9%	0.09	0.09	LLT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		9%	0.09	0.09	PT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737		9%	0.09	0.09	PT	C0687713	Gastrointesti
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	LLT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	PT	C0000737	Abdominal p.
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	PT	C0687713	Gastrointesti
CID100000085	CID000010917	C0002418		3%	0.03	0.03	LLT	C0002418	Amblyopia
CID100000085	CID000010917	C0002418		3%	0.03	0.03	PT	C0002418	Amblyopia
CID100000085	CID000010917	C0002418		6%	0.06	0.06	LLT	C0002418	Amblyopia
CID100000085	CID000010917	C0002418		6%	0.06	0.06	PT	C0002418	Amblyopia
CID100000085	CID000010917	C0002418	placebo	2%	0.02	0.02	LLT	C0002418	Amblyopia

meddra_freq.tsv: Each row in this table contained a side effect and a drug that caused it. Each row contains the name of the side effect, a description of its frequency, upper and lower bounds of its frequency, and whether the side effect name was a preferred medical term. The drug ID's are STITCH ID's, which are ID's from a database of drugs. Each row contains two ID's: one of the combined compound (STITCH CID1), and another of an individual chemical (STITCH CID2).

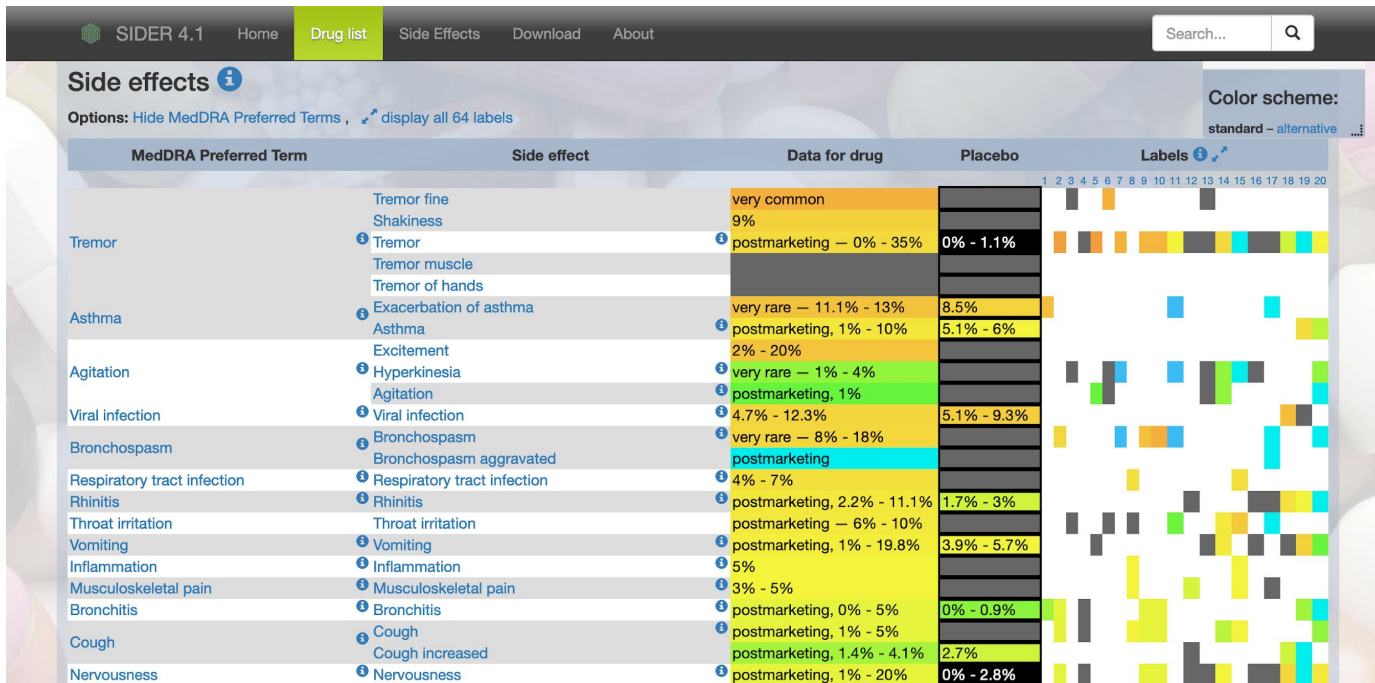
Stitch CID1	Name
CID100000085	carnitine
CID100000119	gamma-amino
CID100000137	5-aminolevulin
CID100000143	leucovorin
CID100000146	5-methyltetrah
CID100000158	PGE2
CID100000159	prostacyclin
CID100000160	prostaglandin
CID100000175	acetate
CID100000187	acetylcholine
CID100000191	adenosine
CID100000206	glucose

drug_names.tsv: Each row maps a drug ID to the name of the drug. The ID's in this table are used in meddra_freq.tsv.

Data - Exploratory Data Analysis



Data - Exploratory Data Analysis





Data - Exploratory Data Analysis

- The website for the SIDER 4.0 dataset allows users to view the side effects of specific drugs. We decided to use it for exploratory data analysis.
- As a part of exploratory data analysis, we also read the dataset's README and a few rows in its tables to understand the structure of its data.



Data - Exploratory Data Analysis

The following findings influenced our design choices:

- The frequencies of the side effects had upper and lower bounds. We found that some drugs lacked frequency information for their side effects. In addition, some drugs represented their frequencies in verbal rather than numerical terms. (For example, labelling their frequencies as “common”, rather than as a percentage.)
- The entries in the dataset for side effects were based on individual labels of a drug. Because a single drug could have multiple labels, a pair of side effect and drug could appear more than once in the dataset.
- The dataset could refer to a side effect using its preferred term (the term that is generally agreed upon) or its lowest-level terms (terms other than the preferred term that refer to the same side effect). As a result, there were duplicate rows for the same side effect which used lowest-level terms instead of the preferred term.
- Some side effect frequencies occurred during placebo administration rather than actual drug dosages.

We will cover how these findings influenced our design in the Data Preprocessing section.



Data Preprocessing

Since the data we wanted to use was in the two files “meddra_freq.tsv” and “drug_names.tsv”, we planned to merge them so that we can have the respective names of the drugs associated with the side effects. Upon merging, we realised that the size of the data set was around 300,000 records, so we planned to reduce it to a more scalable size.

The processing included the following:

1. Since there were only STITCH CID1 drugs (which represent merged compounds) in the drug_names.tsv file, we planned to drop the STITCH CID0 (which represent individual compounds) field from the meddra_freq.tsv file. This made it convenient to join the two data files on the field STITCH CID1.
2. There were duplicate records for all drug-side effect pairs with same frequency labelled according to the “MeddraType” as either LLT (lowest level term) or PT (Preferred Term). We dropped all the duplicate records (rows) so that there were only PT side effects.
3. We dropped all the “Placebo” trials from the dataset.
4. We dropped all rows with frequencies described as “common”, “postmarketing”, “rare”, “infrequent” and “frequent”.
5. There were multiple frequencies for some drug-side effect pairs. We computed the mean, maximum and minimum for frequencies for a given drug-side effect pair, so that there was one row for each pair.



Data Preprocessing

We utilized the pandas library in Python to accomplish data preprocessing.

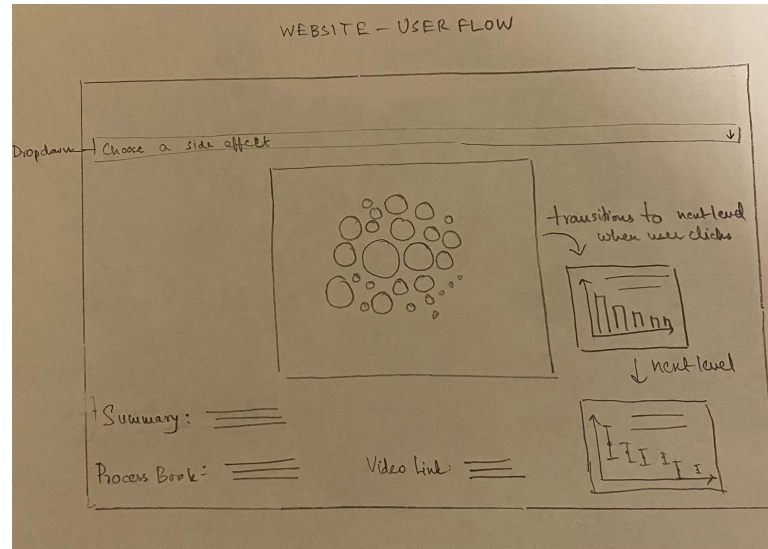
After applying all the preprocessing steps in the previous slide, we get the master.tsv file with the size of the dataset around 30,000 records:

master.tsv

	Name	SideEffectName	MinFrequency	MaxFrequency	MeanFrequency
1	1,25(OH)2D3	Discomfort	0.02	0.03	0.025
2	1,25(OH)2D3	Hypercalciuria	0.03	0.03	0.03
3	1,25(OH)2D3	Laboratory test abnormal	0.08	0.08	0.08
4	1,25(OH)2D3	Pruritus	0.01	0.03	0.02
5	1,25(OH)2D3	Psoriasis	0.04	0.04	0.04
6	1,25(OH)2D3	Urine abnormality	0.04	0.04	0.04
7	1,25(OH)2D3	Urine analysis abnormal	0.04	0.04	0.04
8	17-hydroxyprogesterone	Diarrhoea	0.007	0.023	0.015
9	17-hydroxyprogesterone	Gestational diabetes	0.046	0.056	0.051000000000000004
10	17-hydroxyprogesterone	Gestational hypertension	0.046	0.088	0.067
11	17-hydroxyprogesterone	Injection site nodule	0.02	0.045	0.0325
12	17-hydroxyprogesterone	Injection site pain	0.327	0.348	0.3375
13	17-hydroxyprogesterone	Injection site pruritus	0.033	0.058	0.0455
14	17-hydroxyprogesterone	Injection site swelling	0.078	0.171	0.1245

Design Evolution

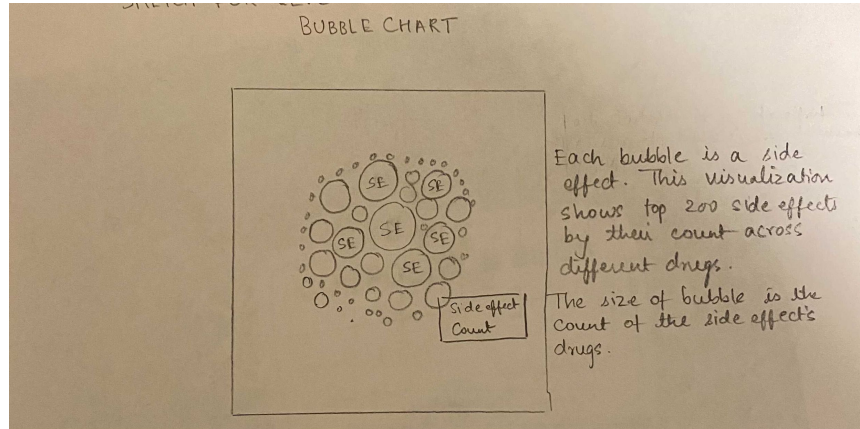
We planned to make a website which features visualizations at three levels. Our idea was to start off with a general view about the side effects related to drugs and then become more specific to a particular drug or a side-effect through user interaction.



Design Evolution

Level 1: This level will exhibit a general view by displaying the most common side effects across all drugs. We planned to accomplish this using a **bubble chart**. Each side effect can be represented by a bubble. Since, the size of the dataset is 30,000, clearly we cannot achieve a neat visualization with that many bubbles. So we decided to show the top 200 side-effects using this visual idiom. The size of the bubble can be determined by either of the following two:

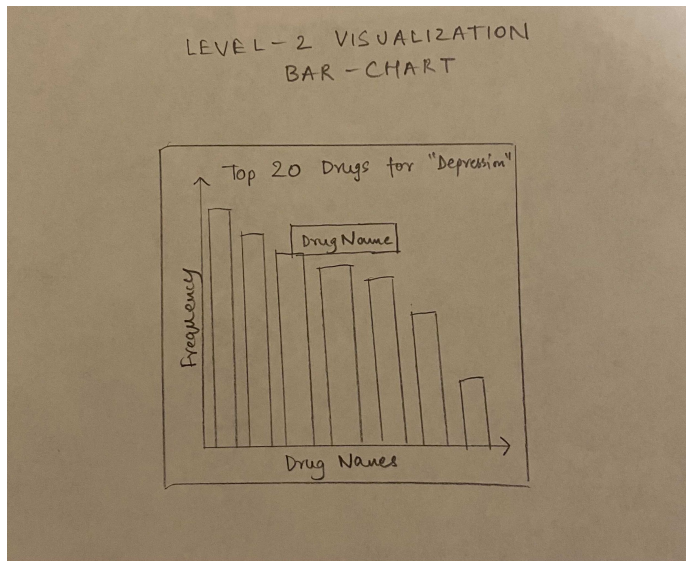
- Could be based on number of drugs related with that side effect
- Could be weighted based on frequency of those side effects across drugs





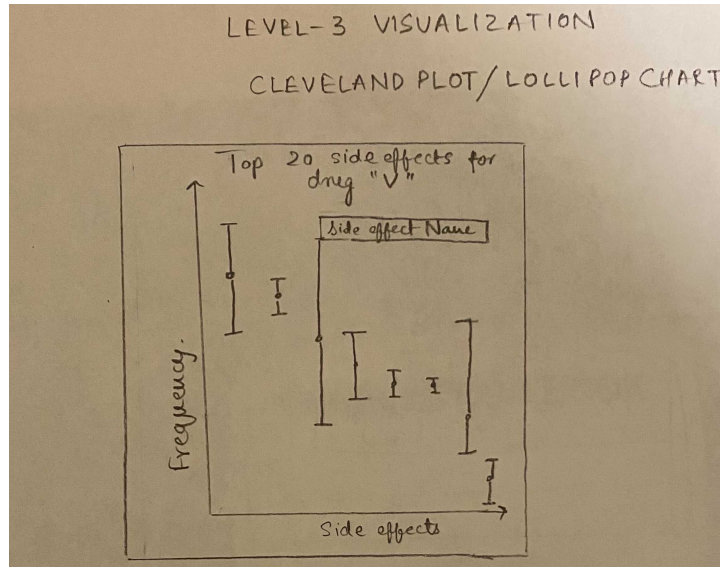
Design Evolution

Level 2: This level will drill down into a side effect chosen by a user and show the top 20 drugs with that side effect, ranked by their frequencies. We planned to use a **sorted bar chart** to accomplish this.



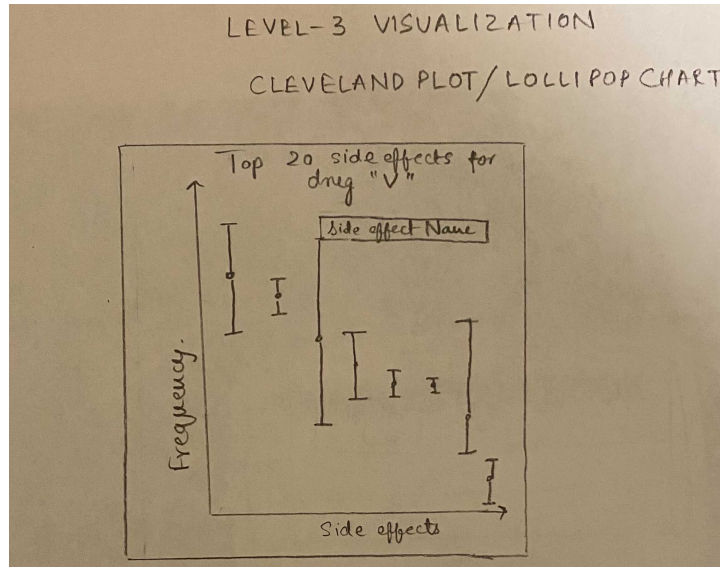
Design Evolution

Level Three: This level will feature a particular drug and will show the frequencies of the side effects for that drug. To achieve this, we planned to utilize a **modified cleveland plot** (also known as the **interval plot**). We displayed the range of frequencies for a particular side-effect for that drug along with the mean frequency as well.



Design Evolution

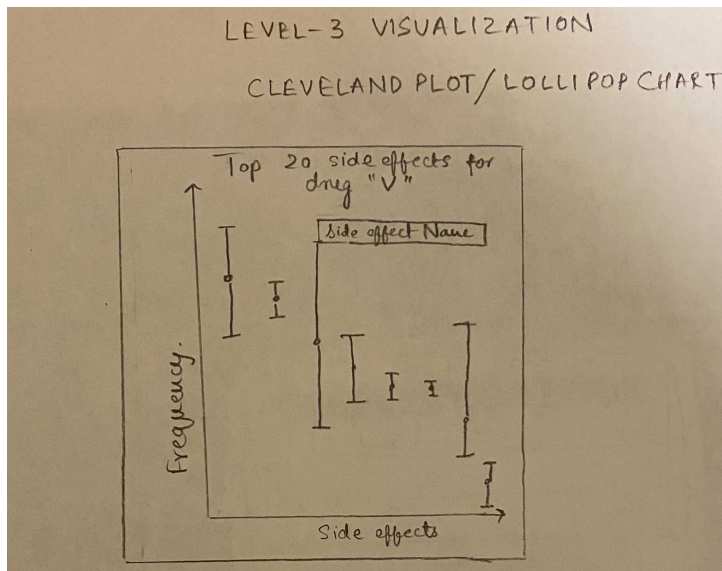
We initially thought of using a sorted bar chart to display the frequencies of the side effects. We decided not to use one because we believed it would be too similar to the level 2 visualization. We also wanted the level 3 visualization to show side effect frequency ranges, which a bar chart would not be able to communicate.

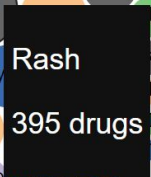




Design Evolution

The original level 3 visualization used circles to show the minimum and maximum frequency, without showing the mean. We then decided to display the mean in the form of a red circle, because we thought it would be helpful to display the mean. Finally, we represented the minimum and maximum frequency using lines instead of circles to decrease visual clutter.





- This takes the form of a clustered bubble chart showing the top 200 side effects ranked by the number of drugs the side effect occurs with.
- Users may hover their mouse cursor over a side effect to see its full name and number of drugs the side effect occurs with.
- Users may click on a side effect to move to a specific view of drugs which cause that side effect.



Implementation

Choose a Drug or Side Effect

Side Effects

Discomfort

Hypercalciuria

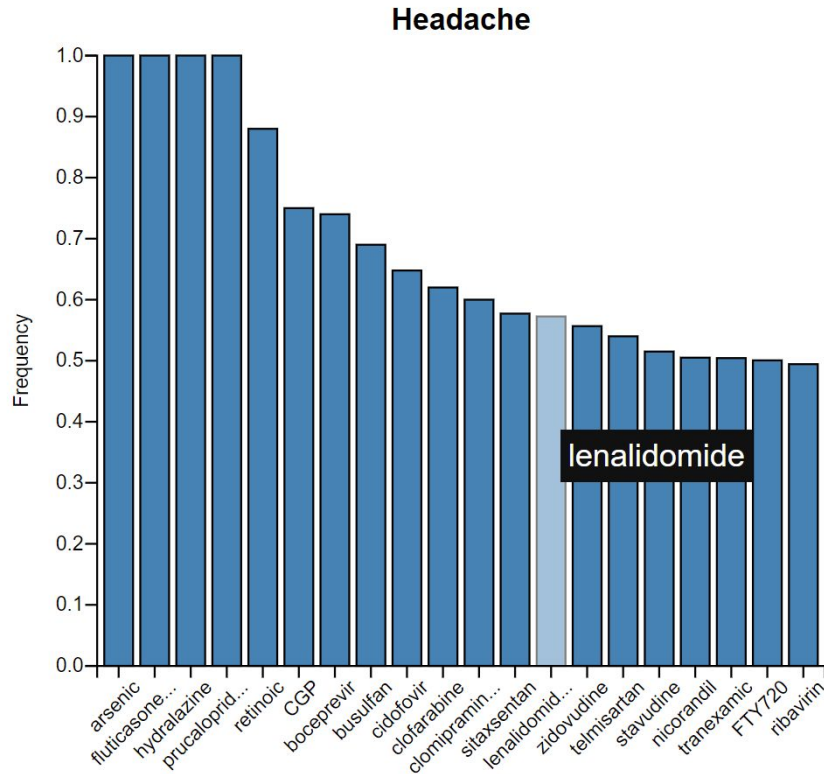
Laboratory test abnormal

Pruritus

Psoriasis

- Users may also choose a particular side effect or drug to learn more about using the bar above the bubble chart.
- Users may select a drug or side effect from the drop-down list or begin typing into the search box to see results.

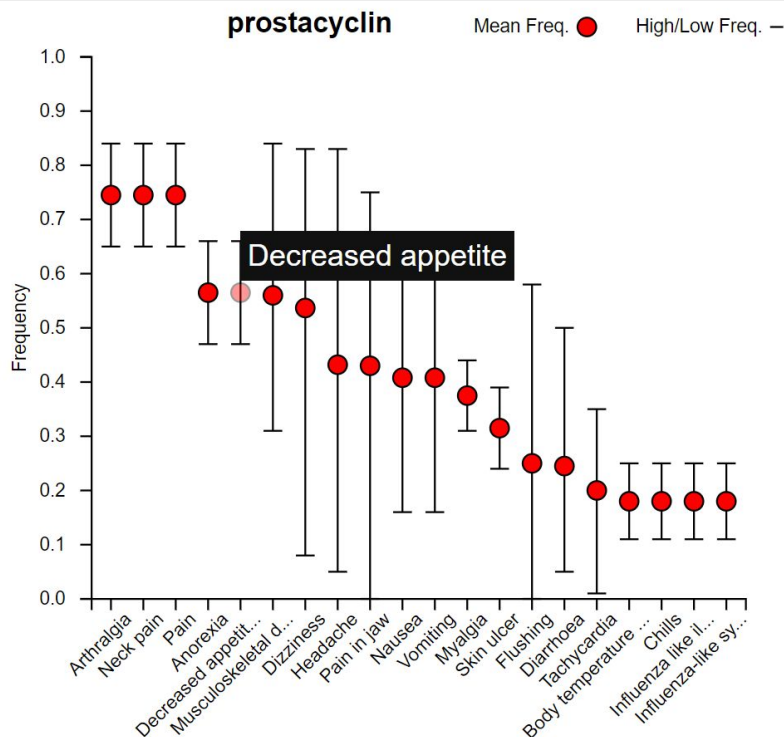
Implementation



If a user chooses a particular side effect, either with the bubble visualization or with the search box, they are presented with a sorted, ranked bar chart showing the top 20 drugs which have that side effect ranked by mean side effect frequency.

Users may roll their mouse cursor over a bar to see the drug's full name, and may click on the bar to see a detailed visualization for that drug in the final layer.

Implementation



The final “layer” of the visualization shows the frequencies of side effects for a particular drug.

Because frequency data was available across multiple trials for the same side effect/drug pairings, we decided to implement a view that shows both the mean frequency of the drug as well as the range (high/low). While the idea started off based off of the Cleveland Dot Plot, over design iterations it morphed to use lines instead of dots at the two ends to represent the range.

Just like the ranked bar chart showing drugs for a given side effect, a user can roll their mouse over a side effect to see the full name, as well as click on a side effect to move back “up” a layer to the ranked bar chart view for that side effect.

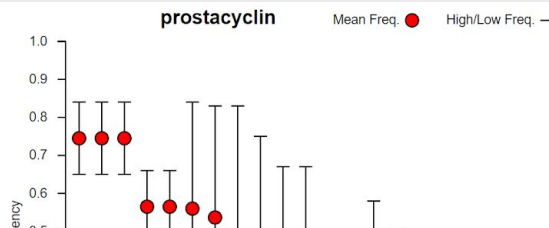


Implementation

prostacyclin



Remove all items

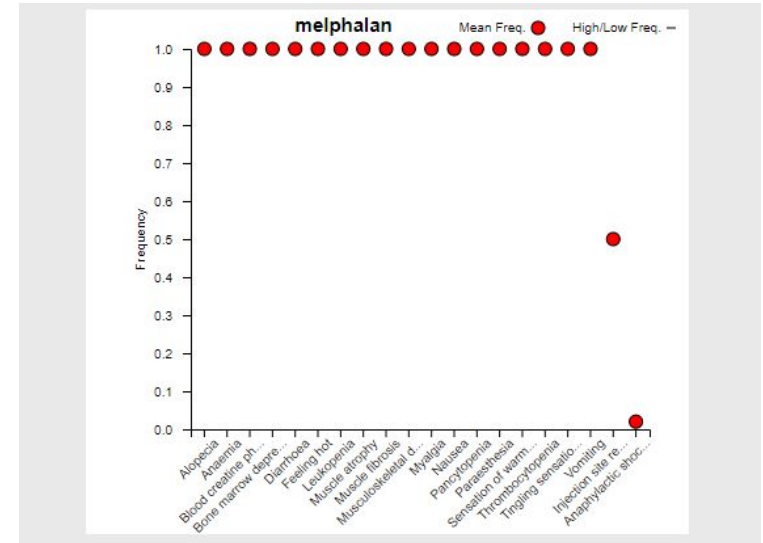


Additionally, the user may directly search for any drug or side effect using the search bar from any view in the visualization and always be brought to the specific drug or side effect view for their selection.

The “X” button on the right side of the search bar clears the selection and allows the user to go back to the top-level bubble chart and browse all side effects.

Evaluation

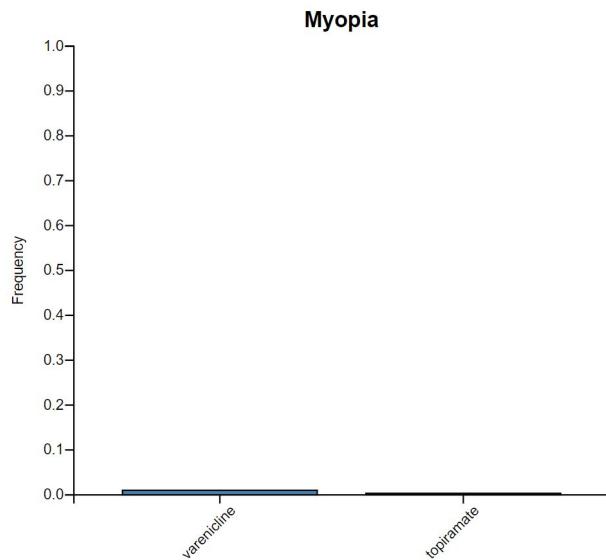
- Our visualization shows that the most common side effects are headache, nausea, diarrhea, and dizziness.
- There are some drugs which, according to the dataset, are guaranteed to have one or many side effects. That is, the upper and lower bounds of frequency for those side effects was equal to 100%. For example, melphalan causes alopecia, anaemia, diarrhoea, nausea, and other side effects with a frequency of 100%.
- Overall, there is a strong tendency for the same common side effects to appear across many different drug-side effect frequencies.





Evaluation

- Because the dataset is very large, we had to limit the number of datapoints shown in our visualizations. Because of this, the visualizations are very good at showing the most common side effects and drugs, but more obscure side effects require a specific search to find information.
- Future improvements could include better accessibility for side effects with small frequencies, possibly through a “show more” button that would allow users to scroll through the side effects and drugs to see smaller frequencies.



Example of a side effect with only two drugs and low frequencies. This side effect required searching to find.