Alexander Obeid
Analysis on the DisGeNET dataset

Summary of results

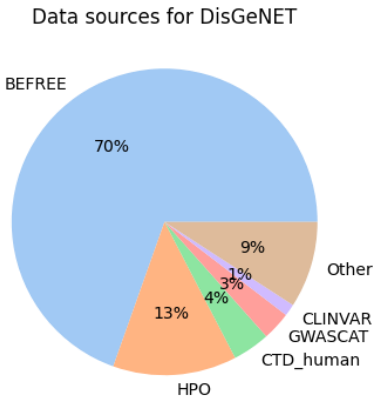
The principal research question on the study was to identify the most reported genes associated with human diseases. The method to do so was mainly creating a bipartite graph between genes and diseases reported on the DisGeNET data set. It was found that the main genes reported are associated with cancer varieties in humans.

Internal validity

I consider that the approach to answer the research question is valid, because if we want to find the most reported genes associated with human diseases it is valid to generate a bipartite network between genes and diseases and then analyze the degree of the genes nodes. This will in fact tell us which genes were the most reported on the data set.

Possible bias

The possible bias in the data concerns the source of the data itself. The DisGeNET data set was made by combining various sources of information. The problem is that the number of entries for each resource is not the same. Actually, 70% of the data comes from a single source as we can see in the graph below.



This comes clearly with a selection bias because most of the data comes from one source called BEFREE. Furthermore, I investigated what was this source and it resulted that BEFREE is a text mining algorithm that extracted gene-disease associations from the abstract of MEDLINE papers. This means, that mostly, the data I used comes from only gene-disease associations reported on the MEDLINE data set. Which is clearly a selection bias problem because MEDLINE might have a bias to the type of papers they publish.

Creating the graph above and investigating the way that BEFREE works was my way of proving that there is a bias on my results. Because I can not infer that the genes I found are the most reported in the literature, as I wanted to prove. In reality, because of the selection bias that the source of the data brings into, I can probably only infer that the genes I found are the most reported genes in the MEDLINE data base (which is not the whole literature).

I will suggest that for preventing this bias in future studies, one should create its own data set, maybe scripting different journal web pages and not only MEDLINE so a more representative data set (of the whole literature) can be generated.