# Process Report and Evaluation

## Preface:

**Data** *(I changed my datasets!)*:

I started out with the search for a dataset not with the aim to find a clean and ready to go dataset but with the aim to find data for a specific topic that I considered important and promising. Thus, I first ended up with datasets I could not get much out of. Thereafter, I ended up with datasets that were in a format nobody can work with and were there was not much to analyse (but still much to show and visualise!).

### Tools/language and resulting files:

*Excel*:

To bring the data into an accessible format [Heatmap: comorb_both.csv, comorb_fem.csv, comorb_mal.csv ; Sankey: sankey1_cleaned.csv, sankey2_cleaned.csv,]

*Python [in Google Collaboratory]*:

To clean the data and transform it to the format needed to load and visualise it in d3.js. [Heatmap: heatmap.ipynb; Sankey: Sankey_MDPD.ipynb, Sankey_PDMD.ipynb]. Please be aware that the code needs to be executed in Google Colab, as opening and storing files is different to a local Jupyter Notebook.

*HTML, CSS & D3.js [in Visual Studio Code]*:

To visualise the data [Heatmap & Sankey: index.html, heatmap.html, sankey.html]

To view the coursework, you need to start a local server with path to the DAV_1961593 folder, which will automatically open the index.html file that allows you to access both visualisations. Alternatively, you may want to open heatmap.html and sankey.html individually.

## **Vis 1** Comorbidity in mental disorders [heatmap]

### Dataset source:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4392551/ [article]

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4392551/bin/jmir_v17i3e55_app2.xlsx [download]

### Process:

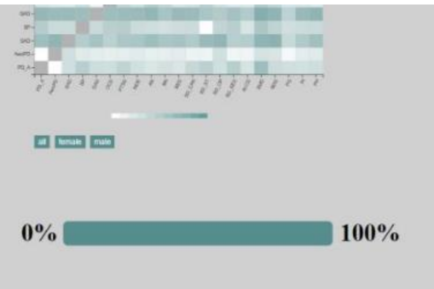
After having converted the matrix like data format into a column format, I prepared the dataset in python and then loaded it with d3.js. The challenges for both visualisations were mainly in getting started with d3.js in its not very supported 5th version (starting with loading the dataset) and getting the datasets into the right format.

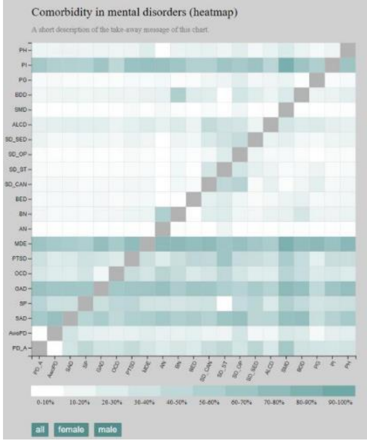
## Prototypes:

While the heatmap was straightforward to build, it was hard to customize as you will see in the prototypes.

*Prototype of colour scale*

The colour scale was a huge issue as I did not find any code examples for a customised colour scale:
Here: First version of my own colour scale and the colour scale I tried to adapt (bottom).



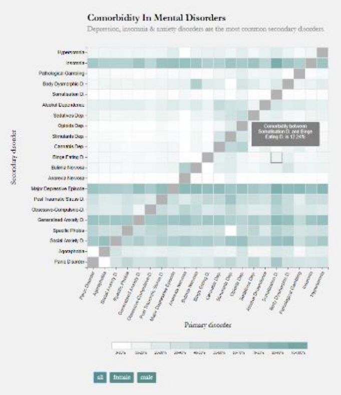


*First prototype of the heatmap*

**1.** Lacking axis labels: it was not clear what was the primary and what was the secondary disorder

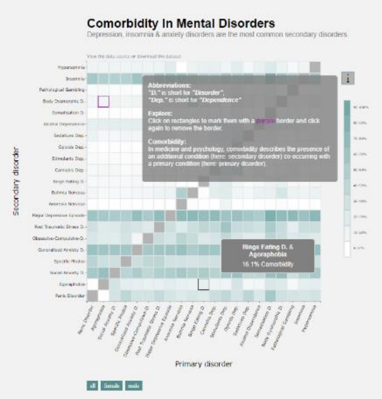
**2.** I realised I needed a table of abbreviations for the tick labels which would make it very hard to have a quick look at the data.

**3.** Background too dark.



*"First final" version of the heatmap*

**1.** Axis labels present
**2.** Tick labels written in full
**3.** Lighter background (forced me to add a stroke everywhere)
**4.** A bit overwhelming for viewer



Abbreviations:
*"D."* is short for *"Disorder"*,
*"Dep."* is short for *"Dependence"*

Explore:
Click on rectangles to mark them with a purple border and click again to remove the border.

Comorbidity:
In medicine and psychology, comorbidity describes the presence of an additional condition (here: secondary disorder) co-occurring with a primary condition (here: primary disorder).

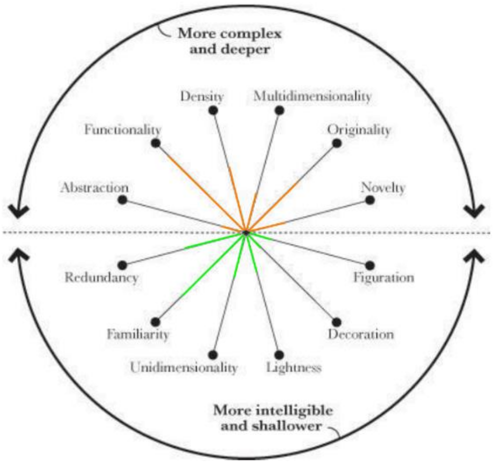
*Final version of the heatmap*

**1.** Rectangles selectable on-click
**2.** Tooltip more readable
**3.** Everything sans-serif
**4.** Information box
**5.** Rather common & intuitive colour scale (high values top, low values bottom)

## Reflective Evaluation:

There are a few limitations to this visualisation. The most obvious one is the long tick labels that simply make it a bit ugly but were a compromise between design and simplicity as well as ease of understanding. Secondly, the placement of the primary disorder on the x-axis and secondary on the y-axis might be a bit unintuitive. It was still the better choice to leave it like that, because the trends in secondary disorders become easier to grasp as they lie in the F-shape of eye movement ( (Nielsen, F Shaped Pattern For Reading Web Content, 2006)). Still, it is unclear if cost and benefit balance each other out here. In depression and insomnia, for example, they blue rectangles almost form a line from left to right which shows that they are quite common 'side effects' of other disorders. This is the main message of the data and it is also conveyed in the visualisation.

My main design goal in this visualisation was to keep it as simple as possible, while still allowing for insight. In human centred design, this is a common approach to interactivity (see 'Aesthetic and minimalist design' (Nielsen, Ten usability heuristics, 2005))). To achieve the maximum of simplicity and reduce cognitive load () I created an information button. However, without any user testing I cannot be sure whether this is very intuitive and if it is not it might hinder the viewer from understanding the abbreviated labels. Another measure to enhance simplicity was the decision to mark rectangles grey where primary and secondary disorder are the same (as it is not meaningful). Simplicity could still be improved for example by reducing the number of disorders to the most relevant (e.g. most frequent ones).



The strengths and weaknesses of the visualisation become clear, when you draw in how it would score on Cairo's visualisation wheel (Cairo, 2012), although this is very subjective without user (viewer) testing:

As you can see, some of the complex dimensions are balanced out by the more intelligible ones. The visualisation is very functional and dense and might not attract enough attention for the average user to view it. However, the new and unusual elements in the visualisation are balanced out quite well with redundancy and use of common features.

While the visualisation wheel might be insightful, Tufte's visualisation principles (Tufte, 2006) allow a more objective and measurable approach:

**Is the lie factor is one?** *Yes. the data is represented accurately, tick marks are correct and labelled, as is the colour scale. However, the ends of the axes look like tick marks, which might be confusing.*

**Do dimensions in data equal dimensions in design?** *Yes. The data and the visualisation have two dimensions each (nominal categories and covariation).*

**Is Necessary context provided?** *Yes, but it remains unclear whether it is easily accessible.*

**High Data:Ink Ratio?** *Not much unnecessary information is displayed, the amount of ink used to show data relative to the total amount of ink used in the graphic is high.*

**Chart Junk?** No, the visualisation is very minimalistic.

## Vis 2 Comorbidity of Mental and Physical Diseases [Sankey diagram]

### Dataset source:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074457/ [article]

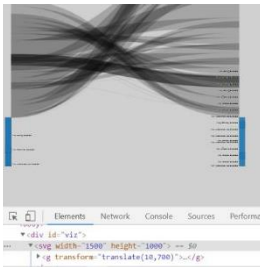
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074457/bin/pone.0165196.s003.xls [download]

### Process:

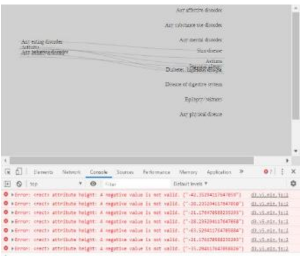
After having converted the matrix like data format into a column format, I prepared the dataset in python and then loaded it with d3.js.

### Prototypes (or "failed versions"):

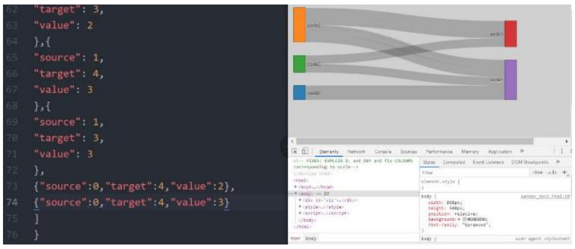
The Sankey plugin seemed like a good idea, but it turned out to be a burden rather than help. The code sample belongs to an old version and it requires a very specific data format plus quite some time to get it right. Also, the error handling is quite misleading (I had thirty version of my dataset plus the effort trying to change the plugin because of error messages suggesting the bugs could be in there). While looking at a normal json-file, the Sankey-format requires an ordered dictionary where the order and number type of every key-value pair matters. It took a fiddle of an older d3 version to find the right solution. Here are some examples of the failed versions (last one is the fiddle example):



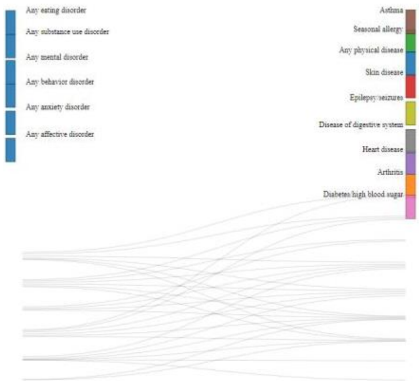
```
<div id="viz">
  ▼<svg width="1500" height="1000"> == $0
    ▶<g transform="translate(10,700)">…</g>
```

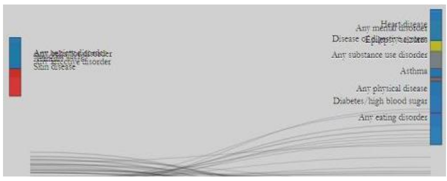


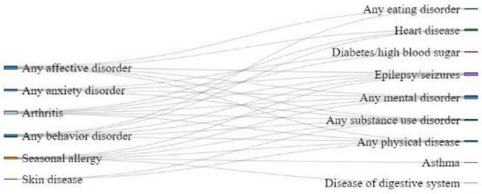
```
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-42.35294117647059")   d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-30.23529411764706")   d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-21.176470588235293")  d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-28.235294117647058")  d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-63.529411764705884")  d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-21.176470588235293")  d3.v5.min.js:2
✖ ▶Error: <rect> attribute height: A negative value is not valid. ("-35.294117647058826")  d3.v5.min.js:2
```



```
"target": 3,
"value": 2
},{
"source": 1,
"target": 4,
"value": 3
},{
"source": 1,
"target": 3,
"value": 3
},
{"source":0,"target":4,"value":2},
{"source":0,"target":4,"value":3}
]
}
```







## Reflective Evaluation:

The Sankey diagram might be a surprising choice, as there are many other ways to visualise hazard ratios (multiple line chart, or multiple boxplots). However,  they normally include the confidence intervals which makes it more complicated for laymen and comparison between the different ratios is extremely difficult (the strength of the links but even more so the number of links from one variable to the others. In the Sankey though, it becomes very clear that some nodes have more and stronger links to others, which can also be seen in the colour and height of the nodes. People with a seasonal allergy for example are more than five times more likely to suffer from an eating disorder than the average citizen and many mental disorders are strongly linked with epilepsy. All these more generic messages of the data are very clearly conveyed and easy to understand, even for the layman.

Nevertheless, there are many limitations to this visualisation, which becomes very apparent when you check it against Tufte's visualisation principles (Tufte, 2006):

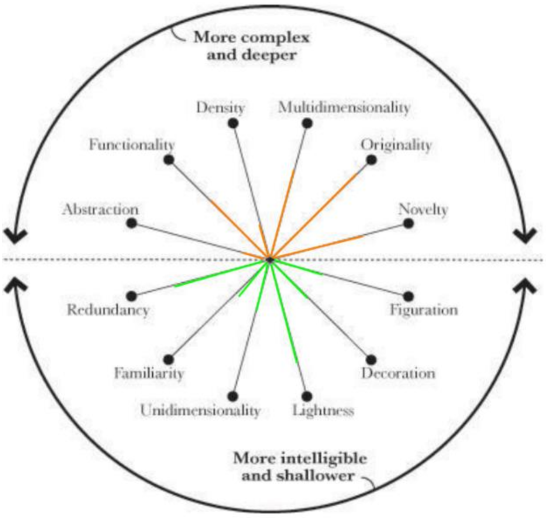
**Is the lie factor is one?** *Maybe. While the data is represented accurately and labelled, the thickness of links is not easy to compare as the units are not clearly visible without the tooltip. A even bigger problem are the nodes, as there is no colour scale and it is unclear what exactly defines the height. Clearly, this visualisation lacks a legend.*

**Do dimensions in data equal dimensions in design?** *Yes. Both the data and the visualisation have nominal categories and the hazard ratios as dimensions. The visualisation does also display the number of links, which is included but not really visible purely by looking at the data.*

**Is Necessary context provided?** *Not enough. The visualisation lacks a legend and possibly an information button like the one in the heatmap, that explains the concept of hazard ratios.*

**High Data:Ink Ratio?** *Not much unnecessary information is displayed, the amount of ink used to show data relative to the total amount of ink used in the graphic is high.*

**Chart Junk?** No. The Sankey is a very minimalistic way of showing the data.



The strengths and weaknesses of the visualisation become clear, when you draw in how it would score on Cairo's visualisation wheel (Cairo, 2012).

As you can see, the Sankey scores well on intelligibility. In comparison to the heatmap it has very limited functionality and it is not very dense. In fact, you don't see any numbers at all in the beginning.  What the wheel does not show is the how intelligibility is affected by the absence of the legend.

The wheel also shows that the visualisation is very original and novel for the average viewer, but the redundant elements and the lightness can make up for that.

# Overall reflective evaluation:

## Colour scheme:

I chose the colour scheme blue-lime green, as lime green is the colour for mental health awareness and blue is the colour many people with mental illnesses prefer. In particular, I used rgb(87, 141, 141) which is the colour of the Mental Health Foundation in the UK and I used the primary colour palette of the US National Alliance on Mental Illness. This is because these colours serve the good cause of promoting awareness while not harming design considerations at the same time. Moreover, they are uncontroversial in this context.

## Typography:

For my visualisation I chose sans-serif types, which was suggested in "Data Visualisation: A Handbook for Data Driven Design"( (Kirk, 2016), page 257), as sans-serif fonts are suitable for shorter sections of text and labels, titles and displays. The labels and numbers are in Veranda to differentiate them from the titles.

## Outlook: what to change:

While the visualisations are kept as simple as possible, they do not good in arousing interest in people who are not natural interested in the matter (McInerny, et al., 2014). An appropriate animation might help to attract and sustain attention, in the Sankey the links expanding from left to right might also help understanding. To further reduce cognitive load in the heatmap, I would propose a clickable colour scale, where only the rectangles that lie withing the selected percentage range are shown and the others are greyed out.

# References:

Kirk, A. (2016). *Data visualisation: a handbook for data driven design.* Sage. Retrieved from https://books.google.de/books?hl=de&lr=&id=wNpsDAAAQBAJ&oi=fnd&pg=PP1&dq=typography+in+data+visualisation&ots=AFxa_sJi9a&sig=xd1p1PUpwI6lANL3d9kRrYraFgQ&redir_esc=y#v=onepage&q=fonts&f=false

Nielsen, J. (2005). Ten usability heuristics. Retrieved from http://lore.ua.ac.be/Teaching/SE3BAC/practicum/acceptanceAndUsabilityTesting/TenUsabilityHeuristics.pdf

Nielsen, J. (2006). F Shaped Pattern For Reading Web Content. Retrieved from http://bit.ly/c9c711

Tufte, E. R. (2006). Beautiful evidence. Graphis Pr.