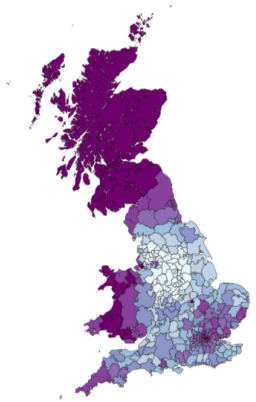
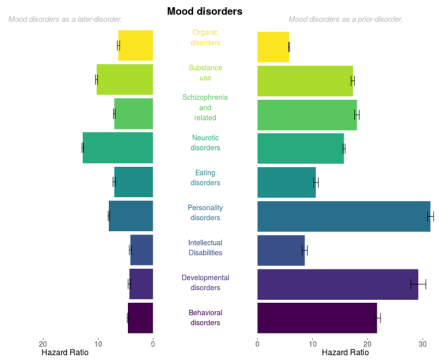
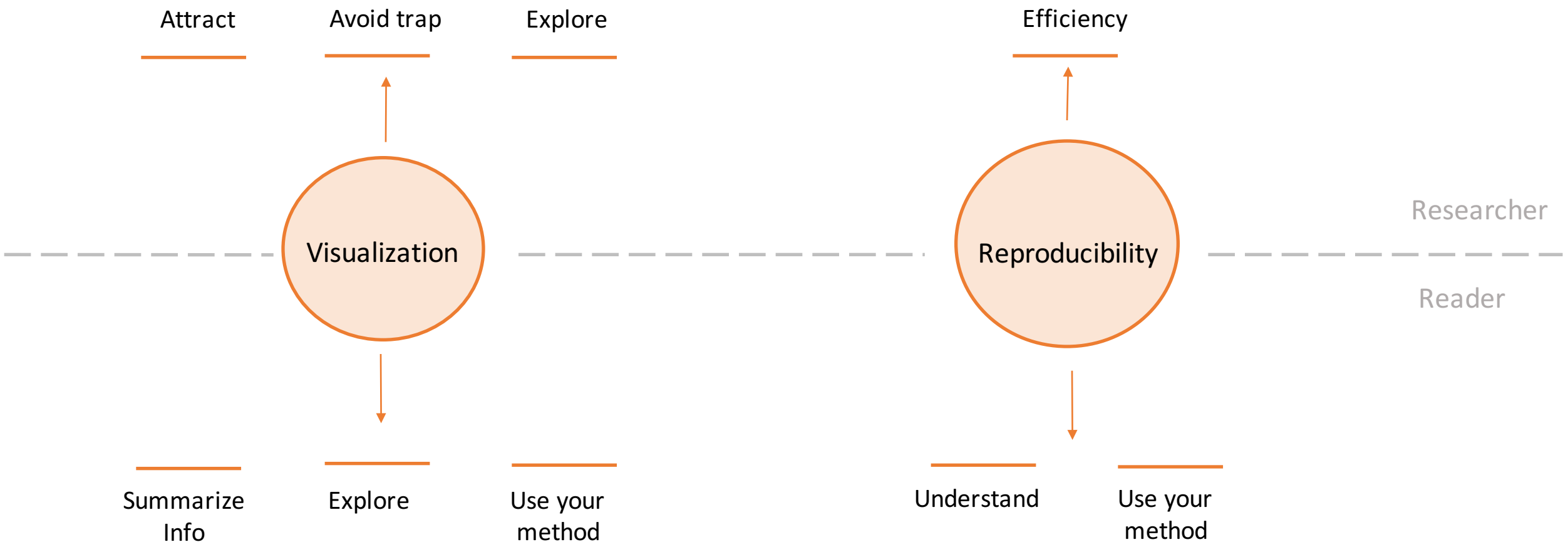
DATAVIZ, REPRODUCIBILITY AND IMPACTFUL RESEARCH

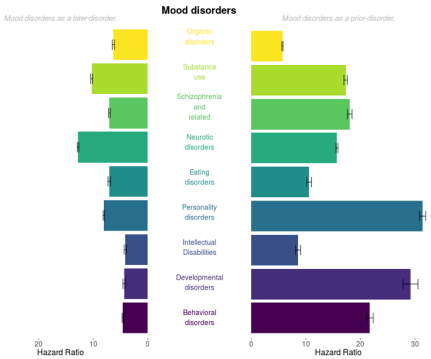




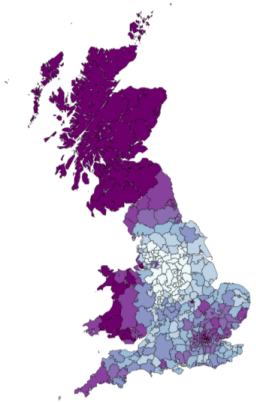

Attract
Avoid trap
Explore
Efficiency
Visualization
Reproducibility
Researcher
Reader
Summarize Info
Explore
Use your method
Understand
Use your method

DATAVIZ

Comorbidity in the UK Biobank



Comorbidity in the Danish register



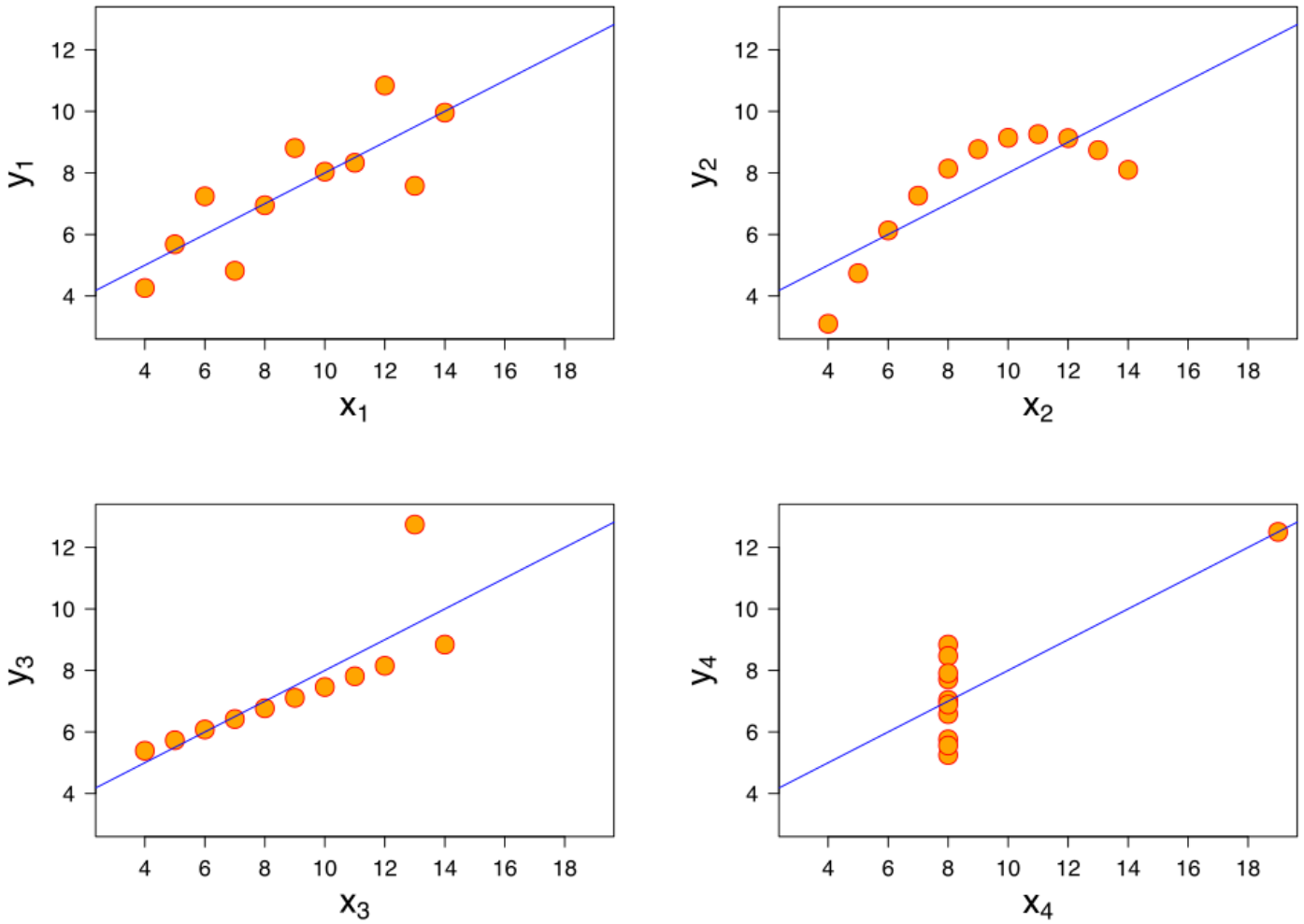
Geography in the UK Biobank

Y = 3 + 0.5x
Cor = 0.8

Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1

Anscombe et al. 1973



Y = 3 + 0.5x
Cor = 0.8

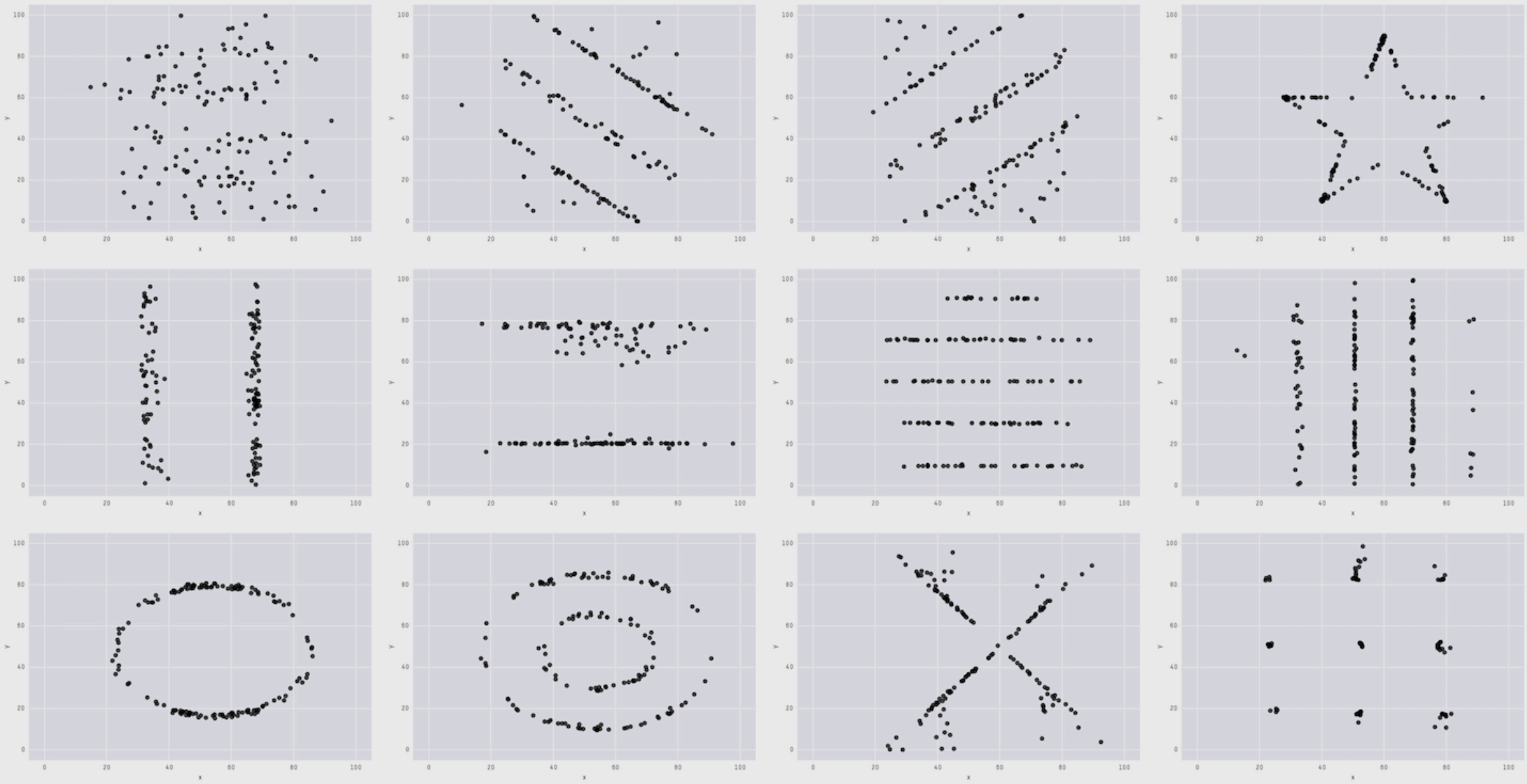
Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1

Plot your data

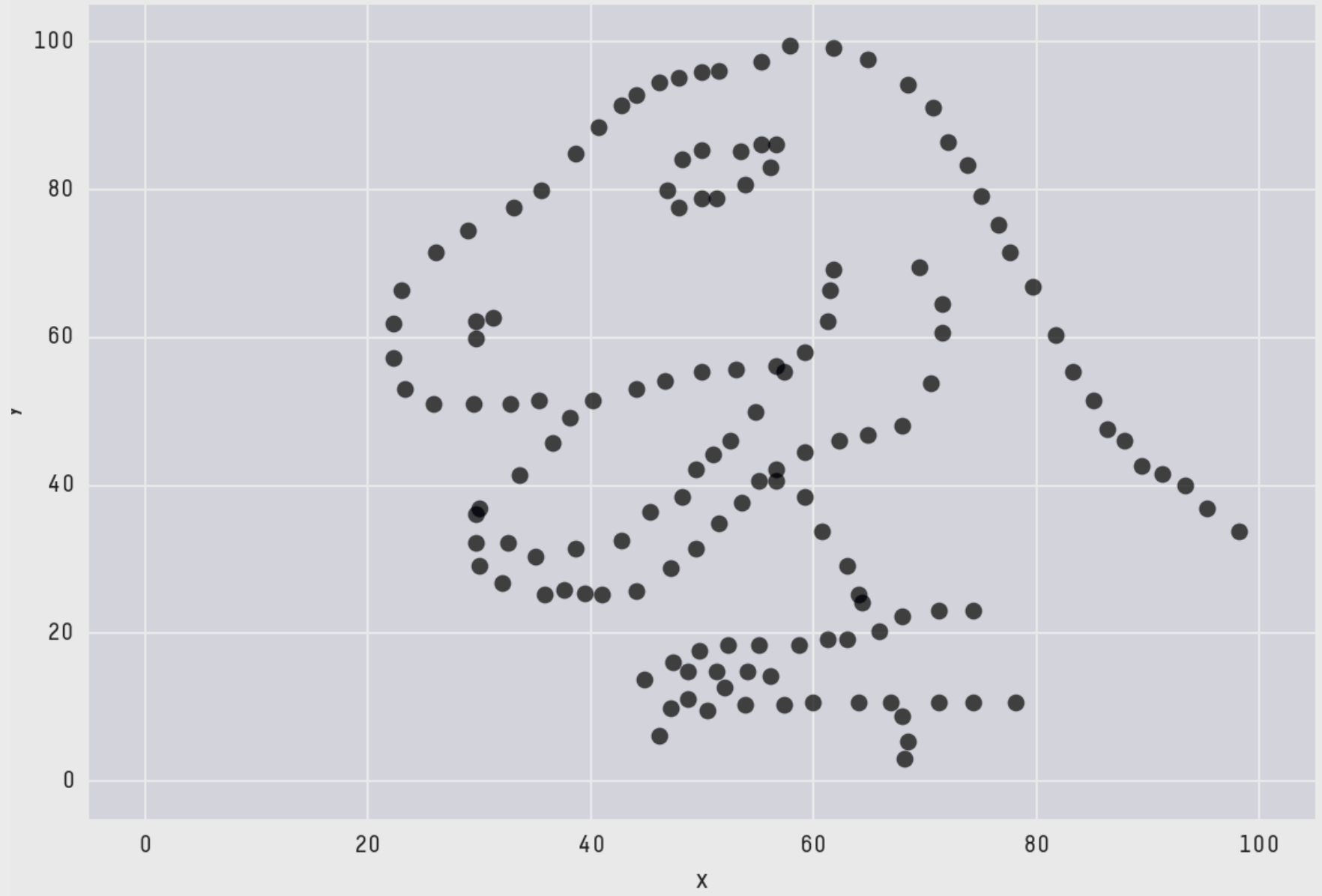
Anscombe's quartet

Anscombe et al. 1973



The Datasaurus Dozen

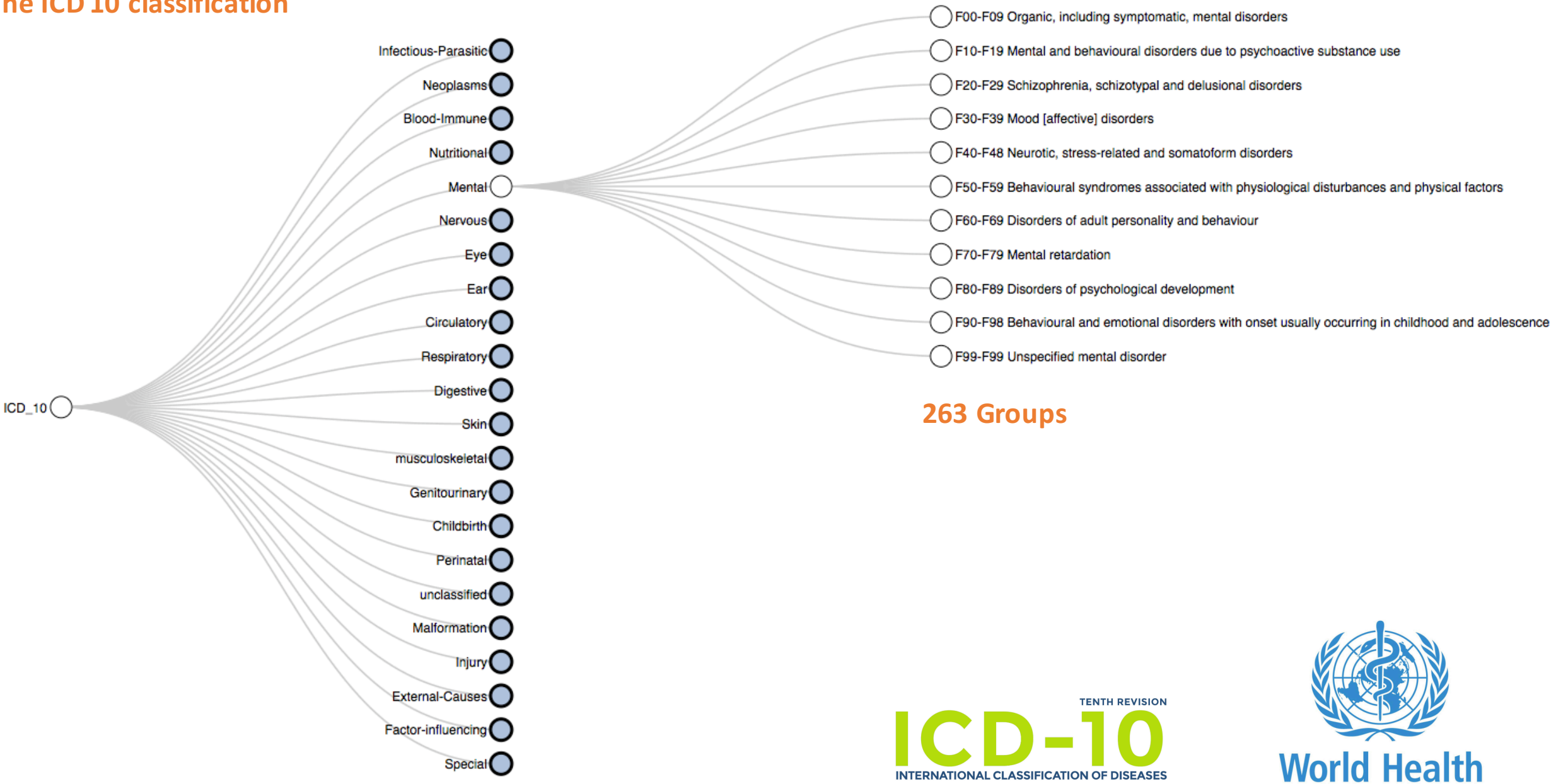
Cairo, 2017
Matejka & Fitzmaurice, 2017



The Datasaurus Dozen

Cairo, 2017
Matejka & Fitzmaurice, 2017

The ICD 10 classification

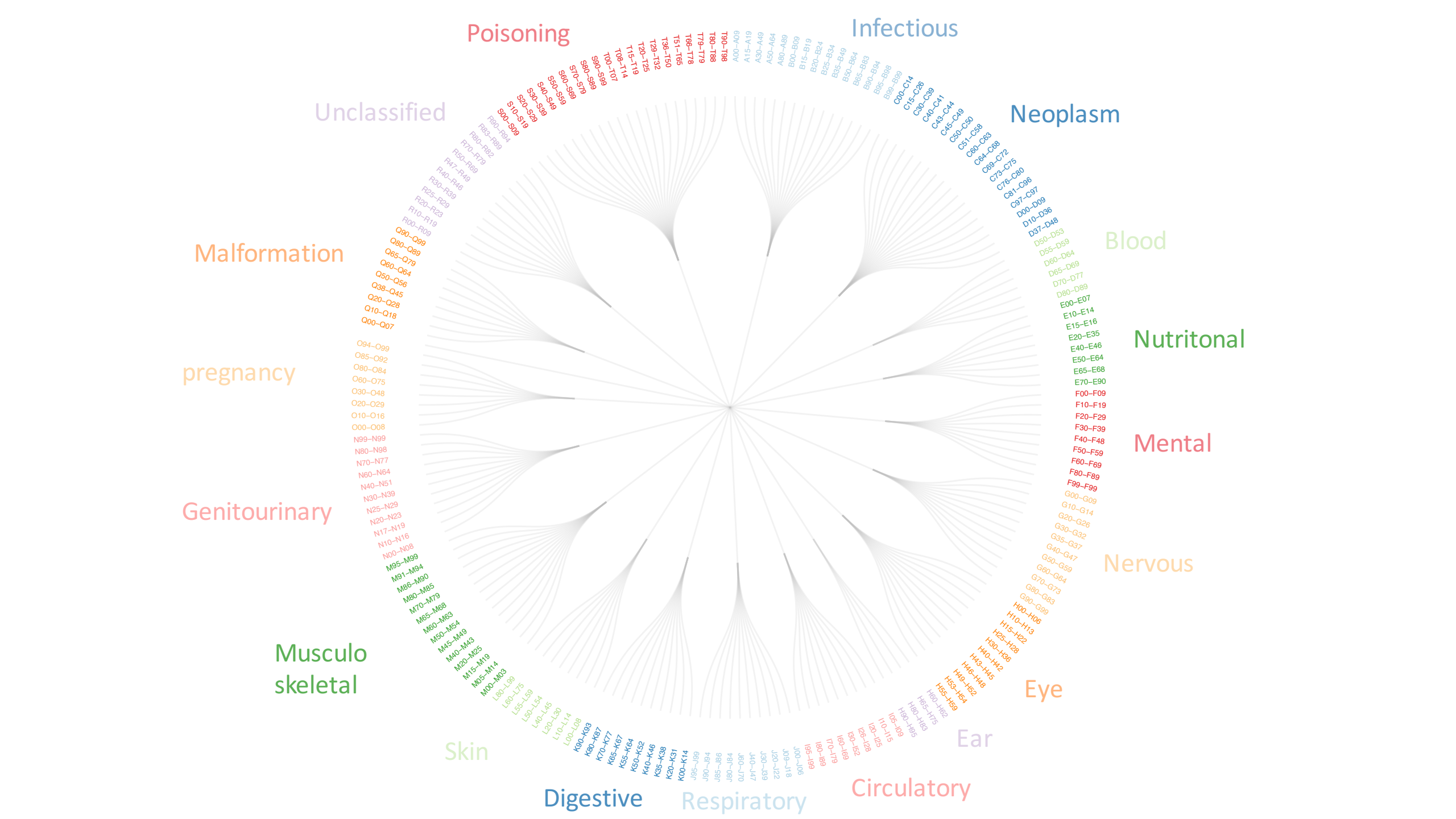


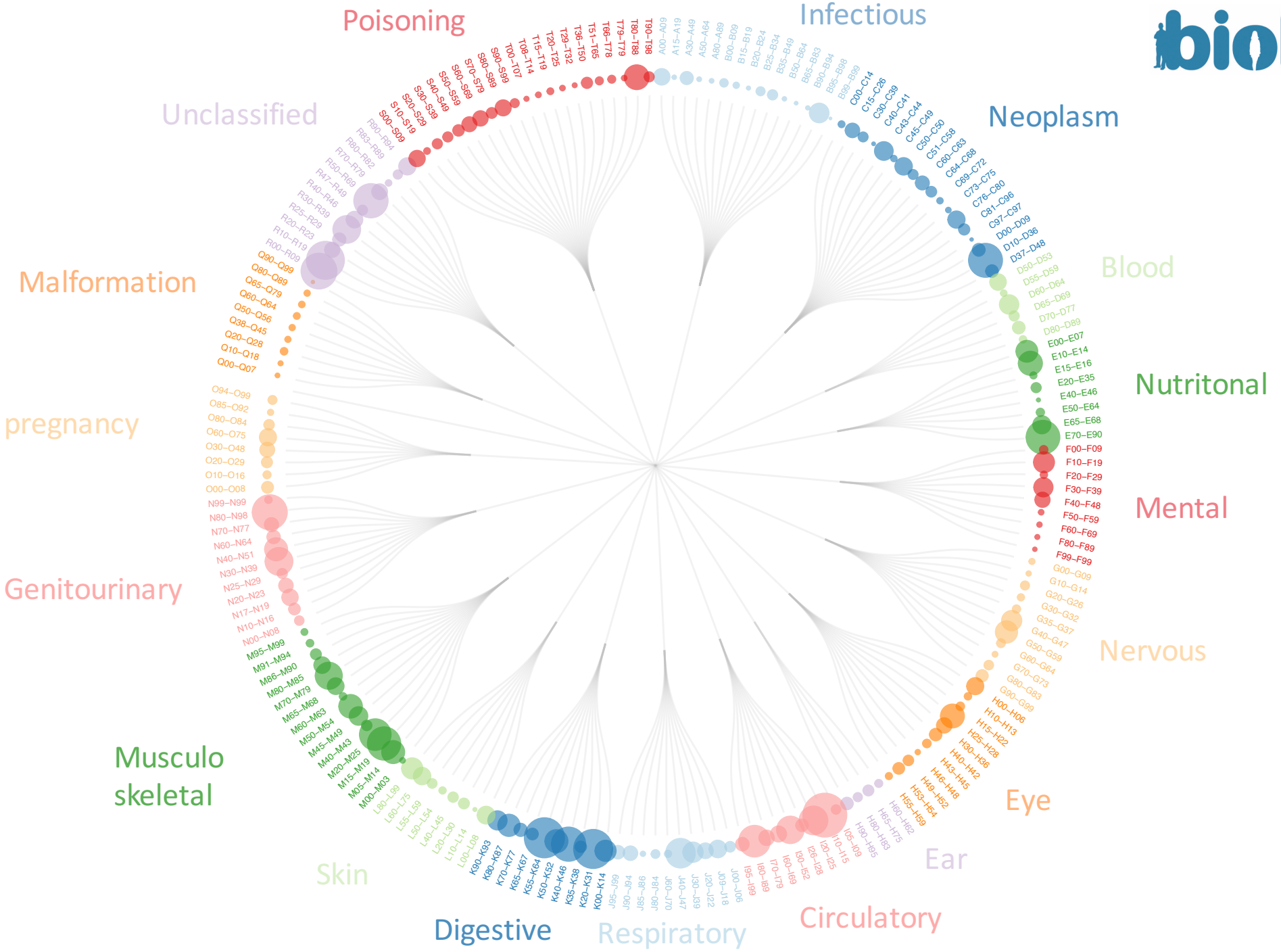
263 Groups

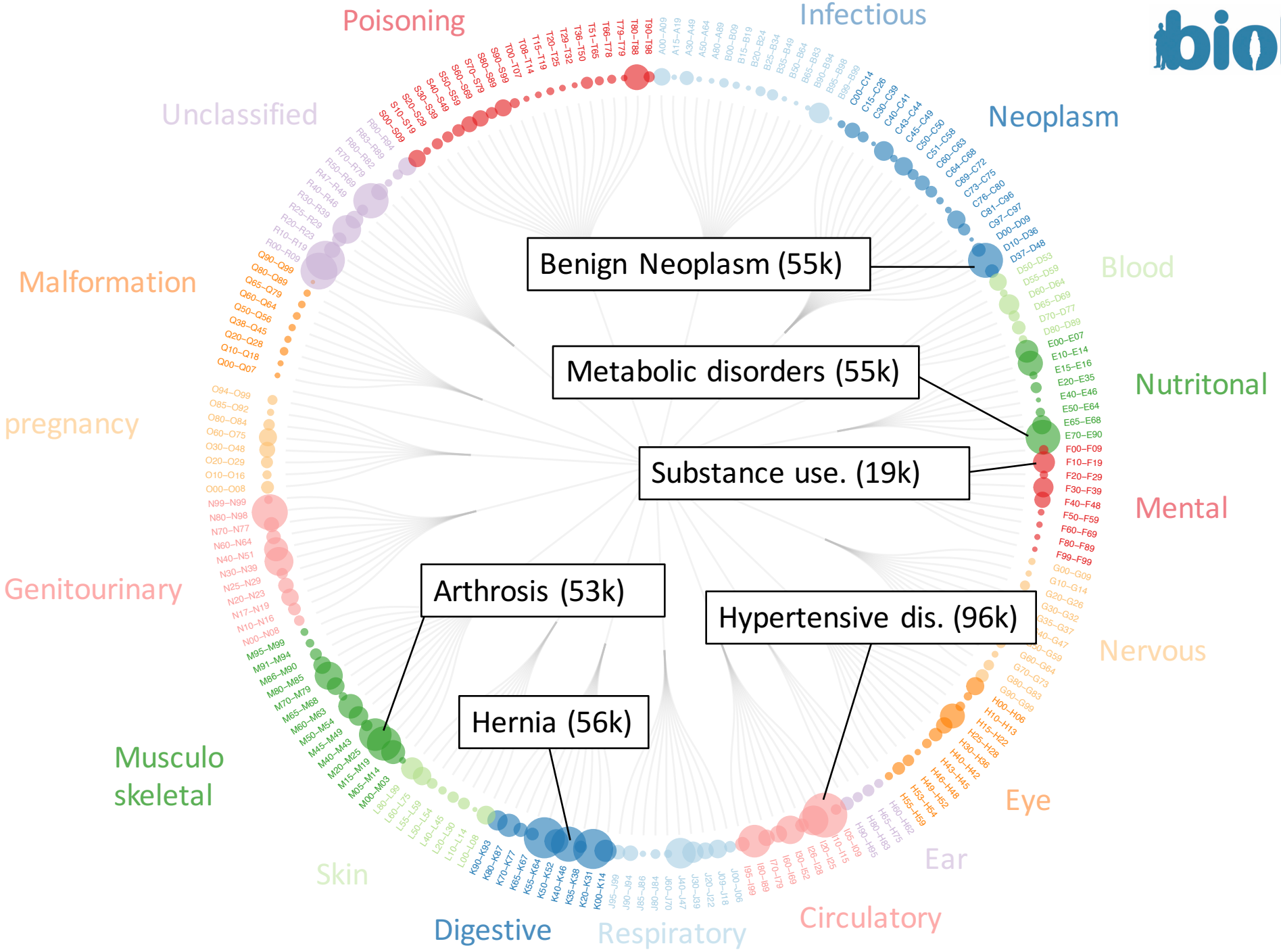
22 Chapters

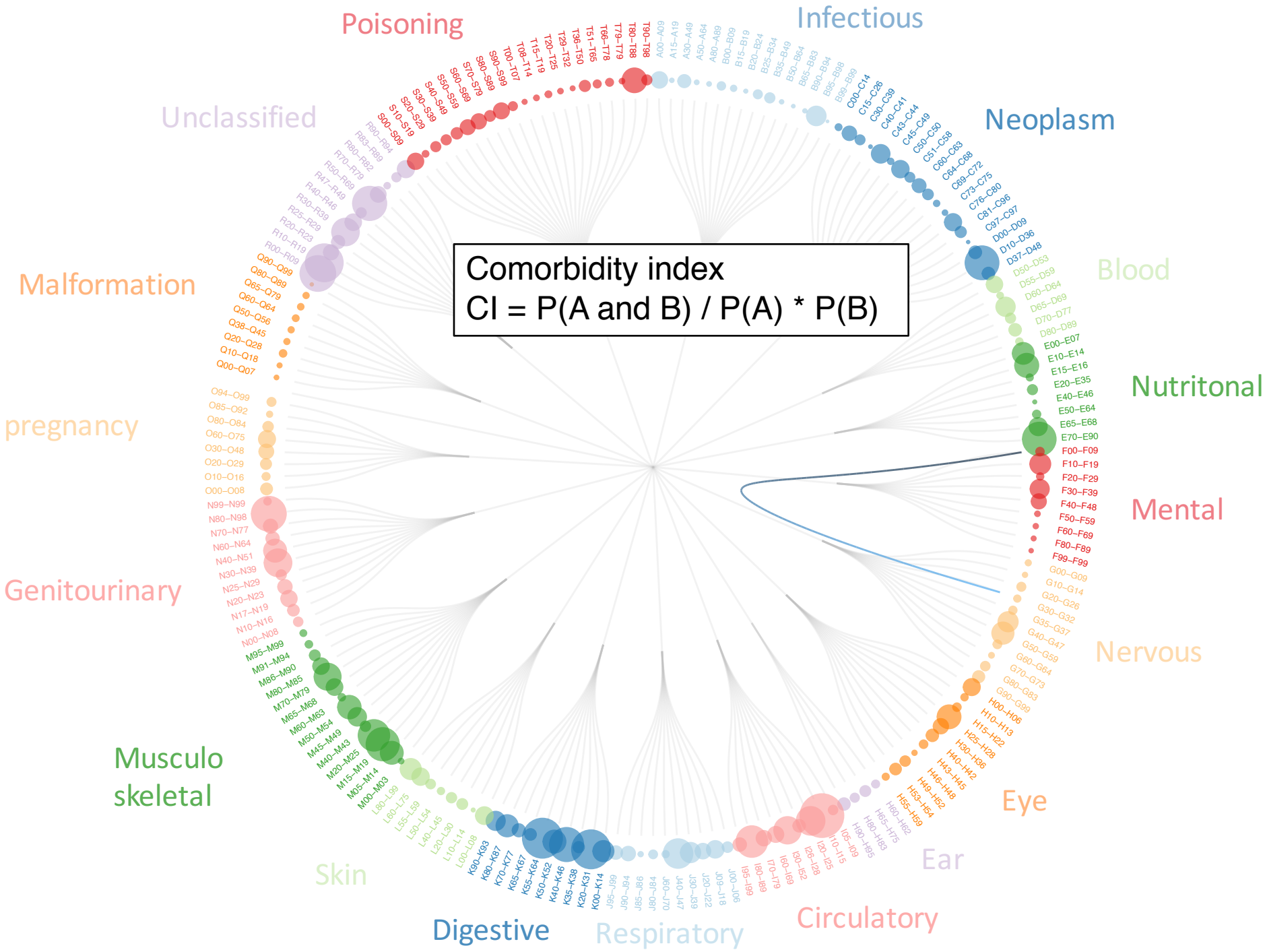
TENTH REVISION
ICD-10
INTERNATIONAL CLASSIFICATION OF DISEASES
ICD-10 is a new code set for reporting medical diagnoses & inpatient procedures.

World Health Organization

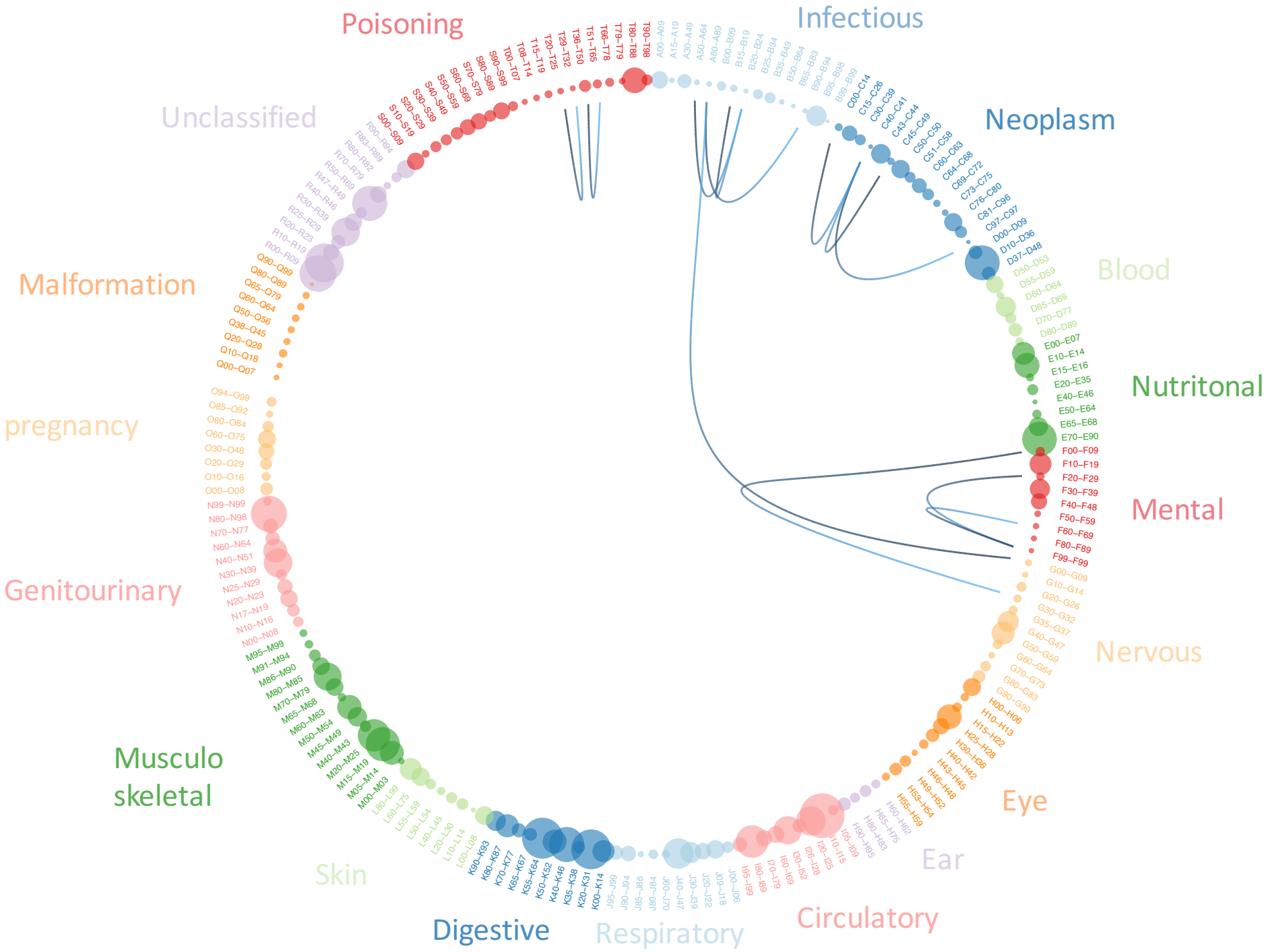

Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Respiratory
Circulatory

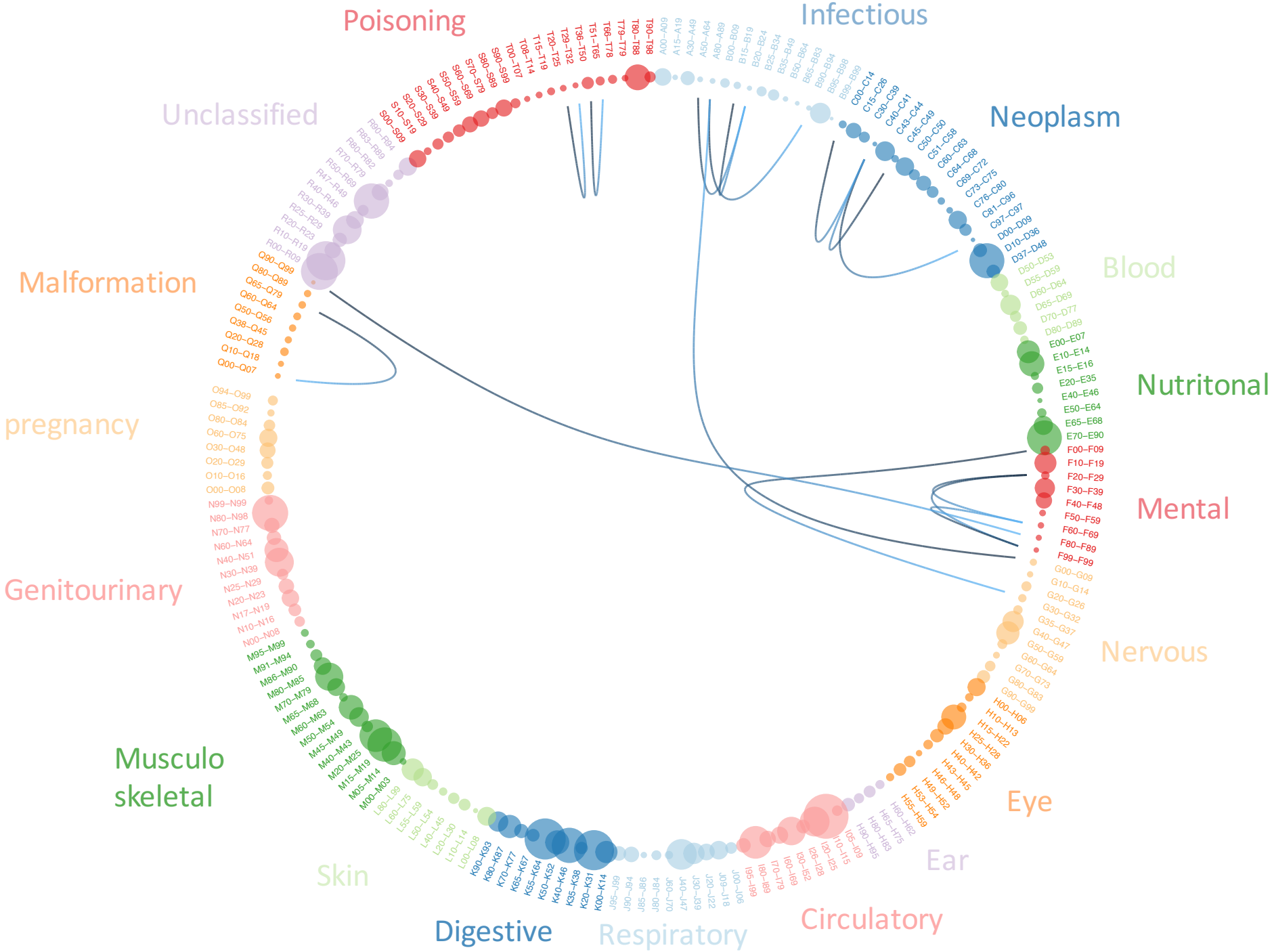

Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Respiratory
Circulatory

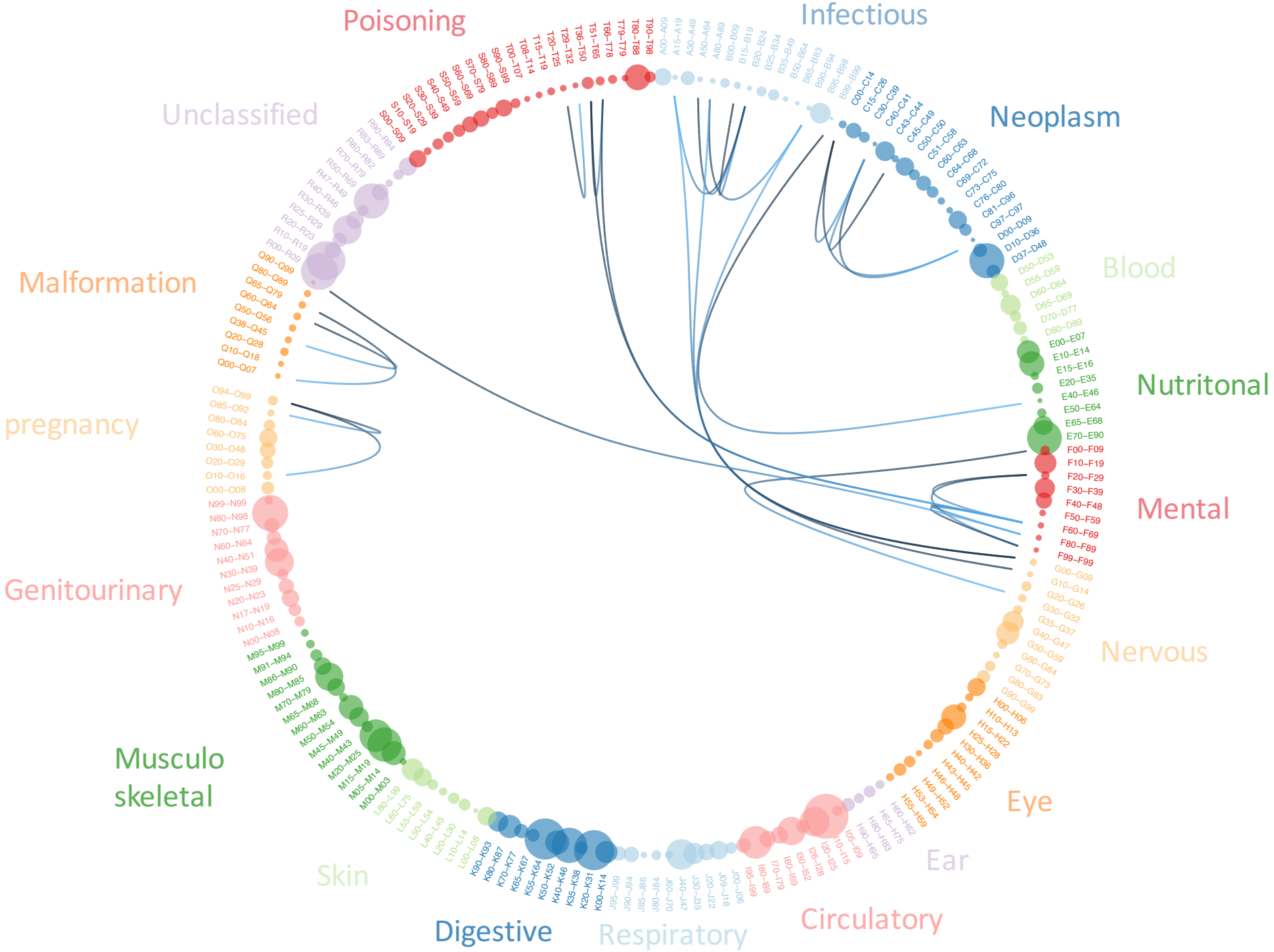

Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Respiratory
Circulatory
Benign Neoplasm (55k)
Metabolic disorders (55k)
Substance use. (19k)
Arthrosis (53k)
Hypertensive dis. (96k)
Hernia (56k)


Comorbidity index
CI = P(A and B) / P(A) * P(B)
Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Ear
Skin
Circulatory
Digestive
Respiratory

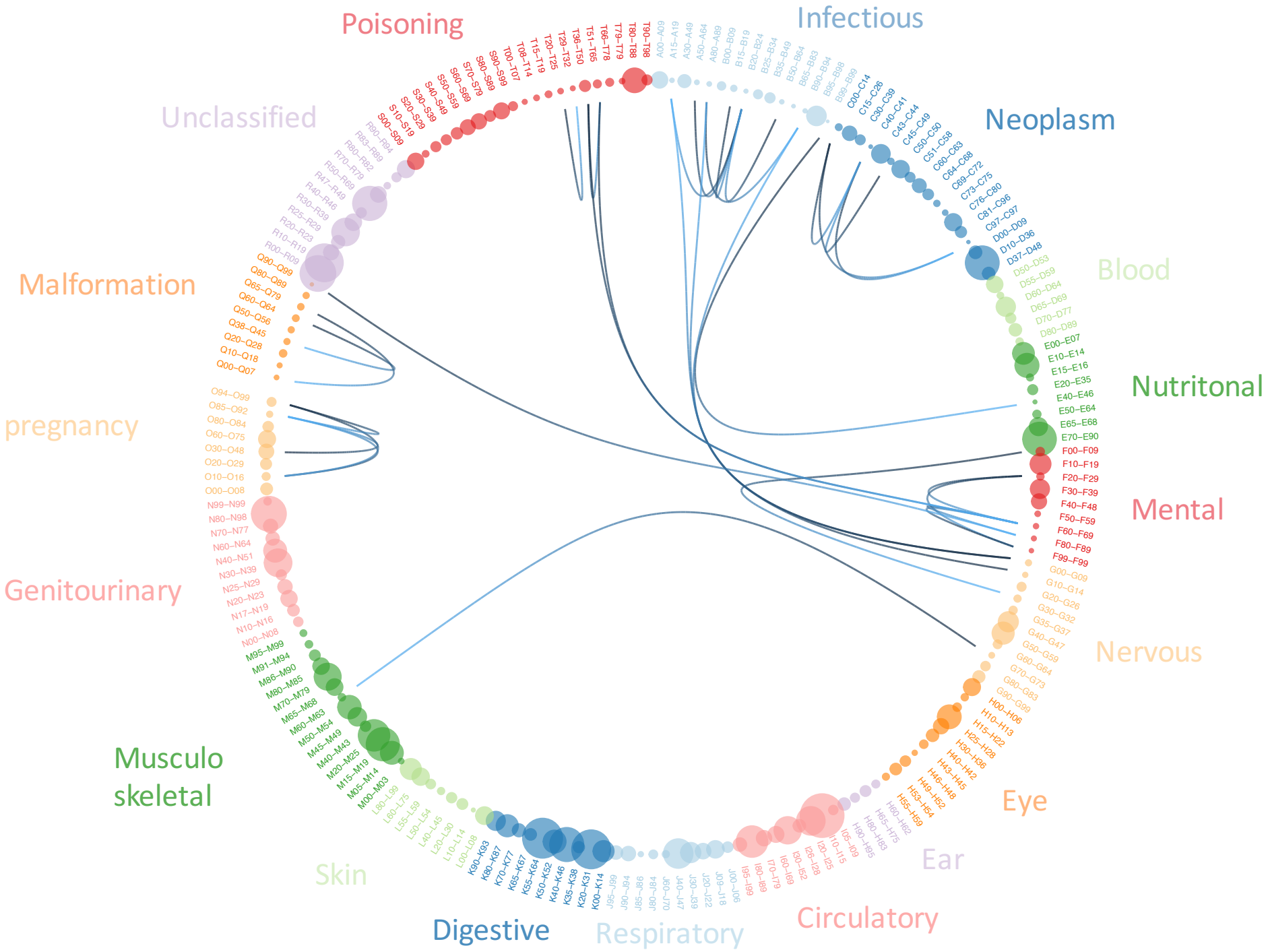
CI > 70


Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Circulatory
Respiratory

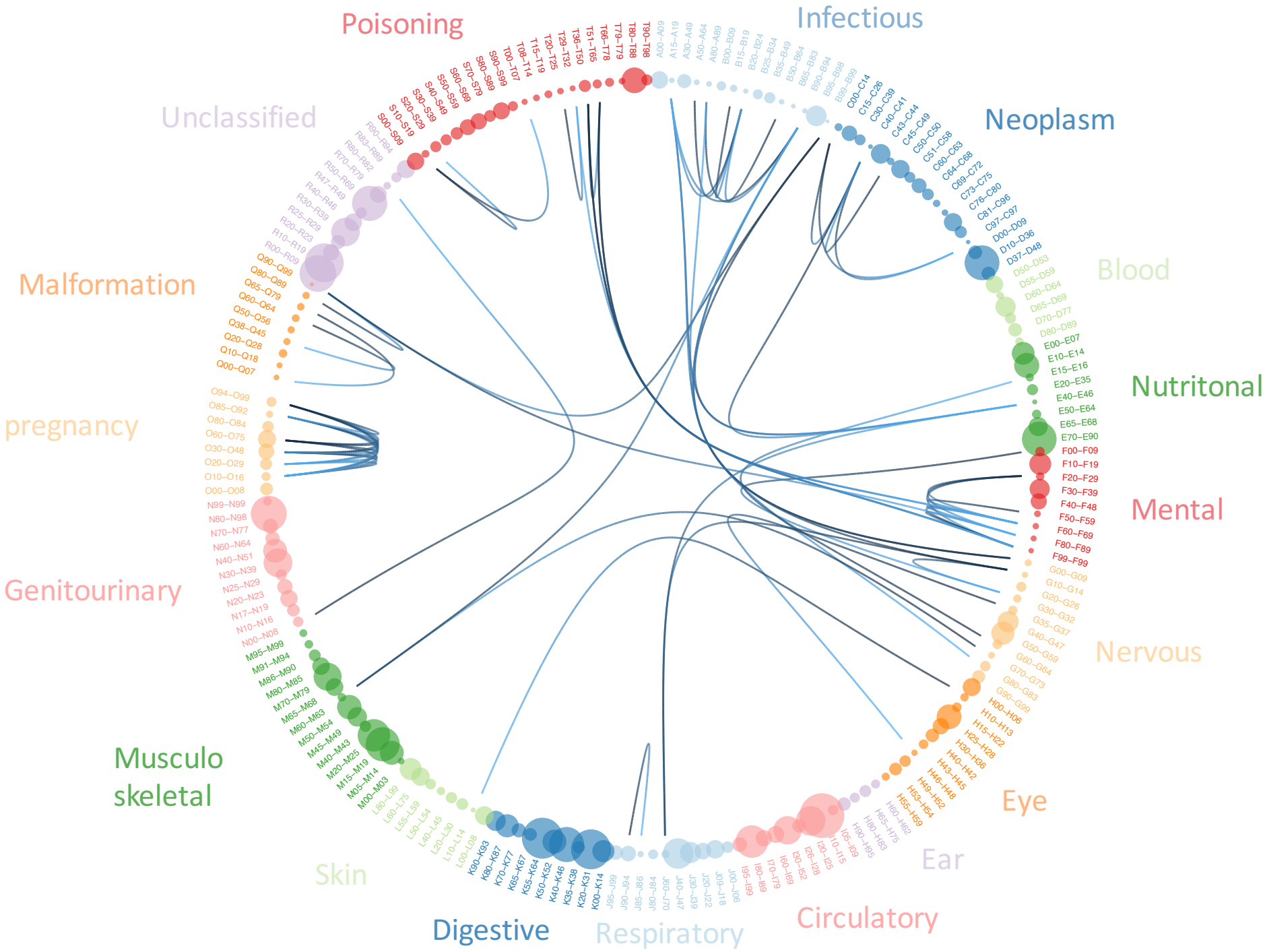

CI > 60
Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
Nutritonal
pregnancy
Mental
Genitourinary
Nervous
Musculo skeletal
Eye
Ear
Skin
Circulatory
Digestive
Respiratory

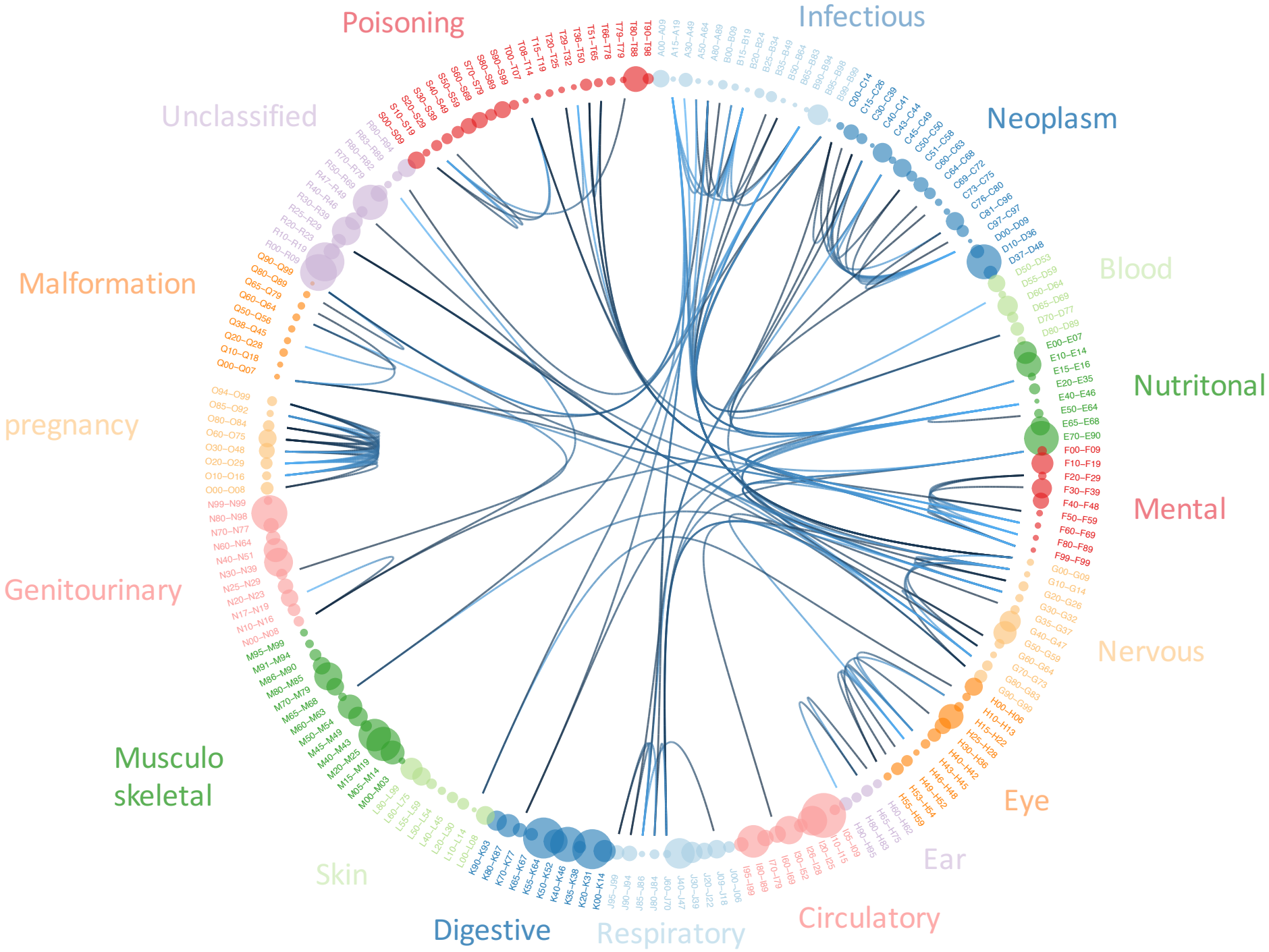

CI > 50
Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Ear
Skin
Circulatory
Digestive
Respiratory

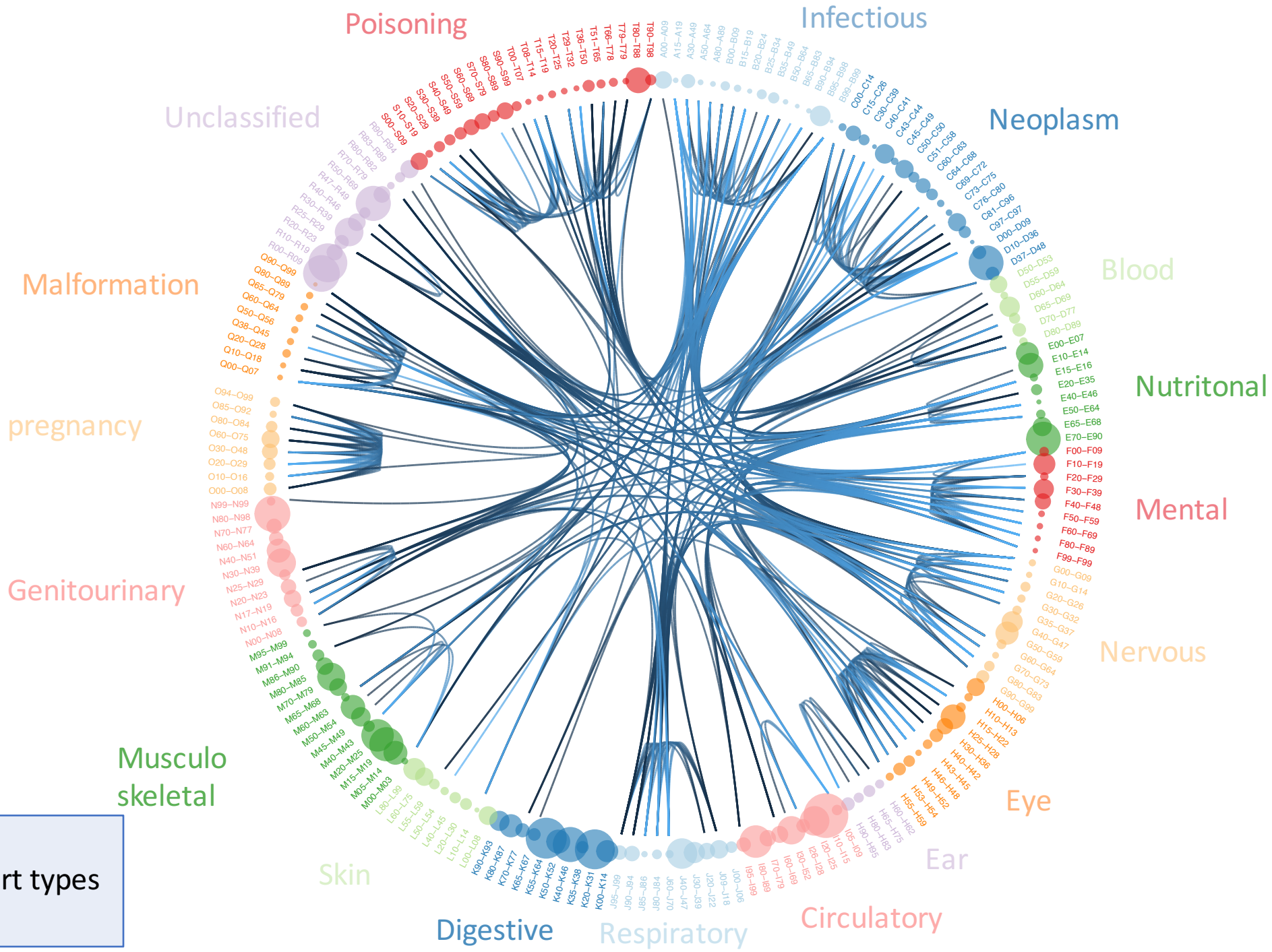
CI > 40


Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Mental
Genitourinary
Nervous
Musculo skeletal
Eye
Ear
Skin
Circulatory
Digestive
Respiratory

CI > 30

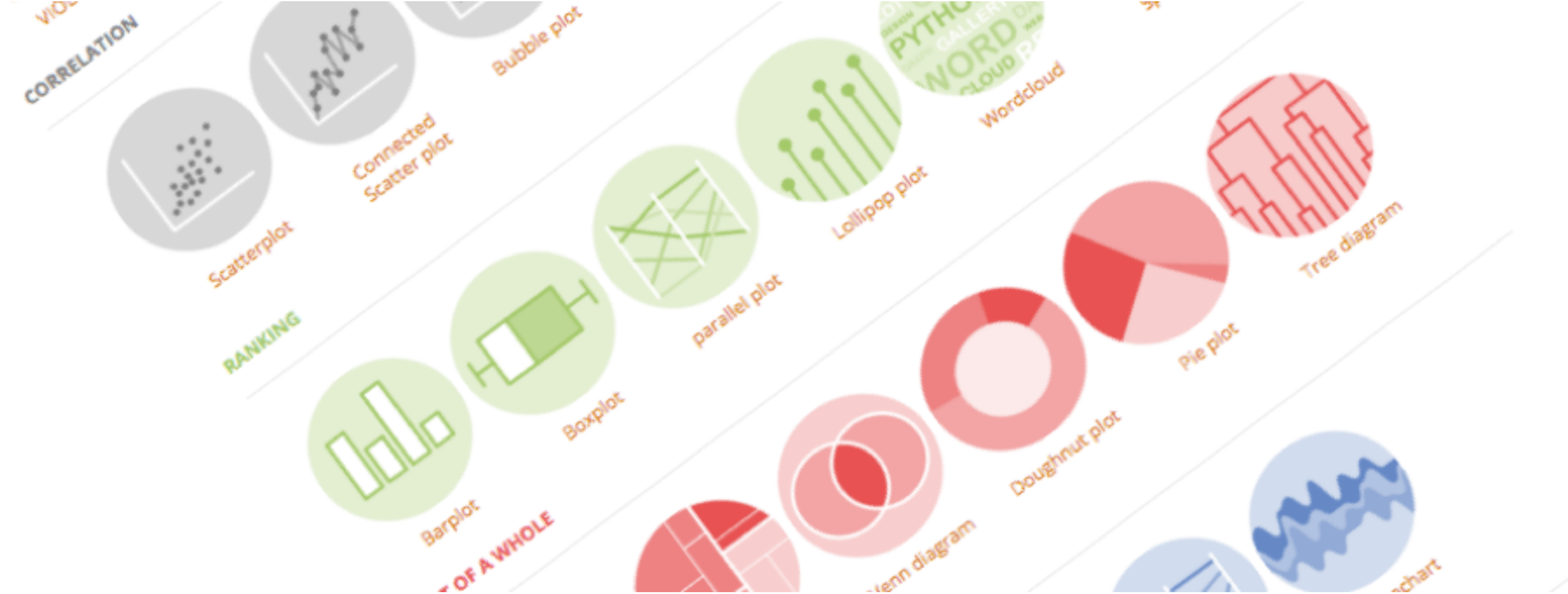

Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Respiratory
Circulatory


CI > 20
Poisoning
Infectious
Unclassified
Neoplasm
Malformation
Blood
pregnancy
Nutritonal
Genitourinary
Mental
Musculo skeletal
Nervous
Eye
Skin
Ear
Digestive
Respiratory
Circulatory



CI > 10

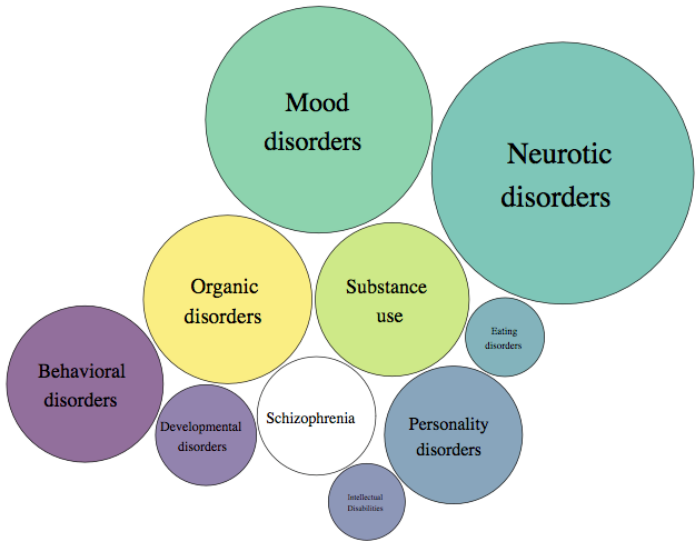
Try other chart types



R-graph-gallery.com

Python-graph-gallery.com

Danish register
(n ~ 5M)

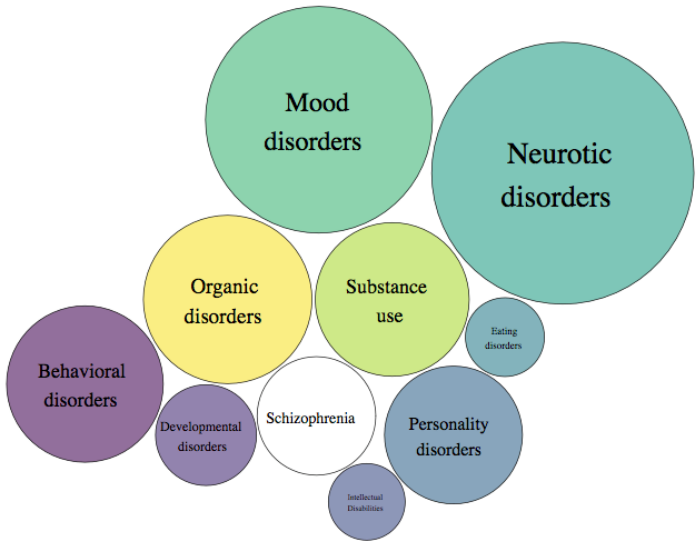


10 Mental Disorder Groups (F chapter of ICD 10)

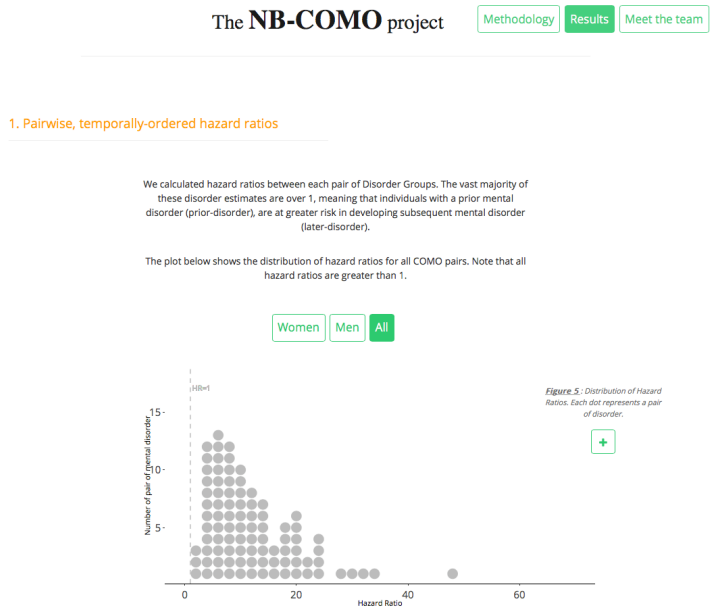
Compute Hazard Ratios:

- 90 pairs
- Males and Female
- Several Models
- Many angles of analysis

Danish register
(n ~ 5M)



10 Mental Disorder Groups (F chapter of ICD 10)



Compute Hazard Ratios:

- 90 pairs
- Males and Female
- Several Models
- Many angles of analysis

Organic | Substance | Schizophrenia | **Mood** | Neurotic | Eating | Personality | Intellectual Dis. | Developmental | Behaviour

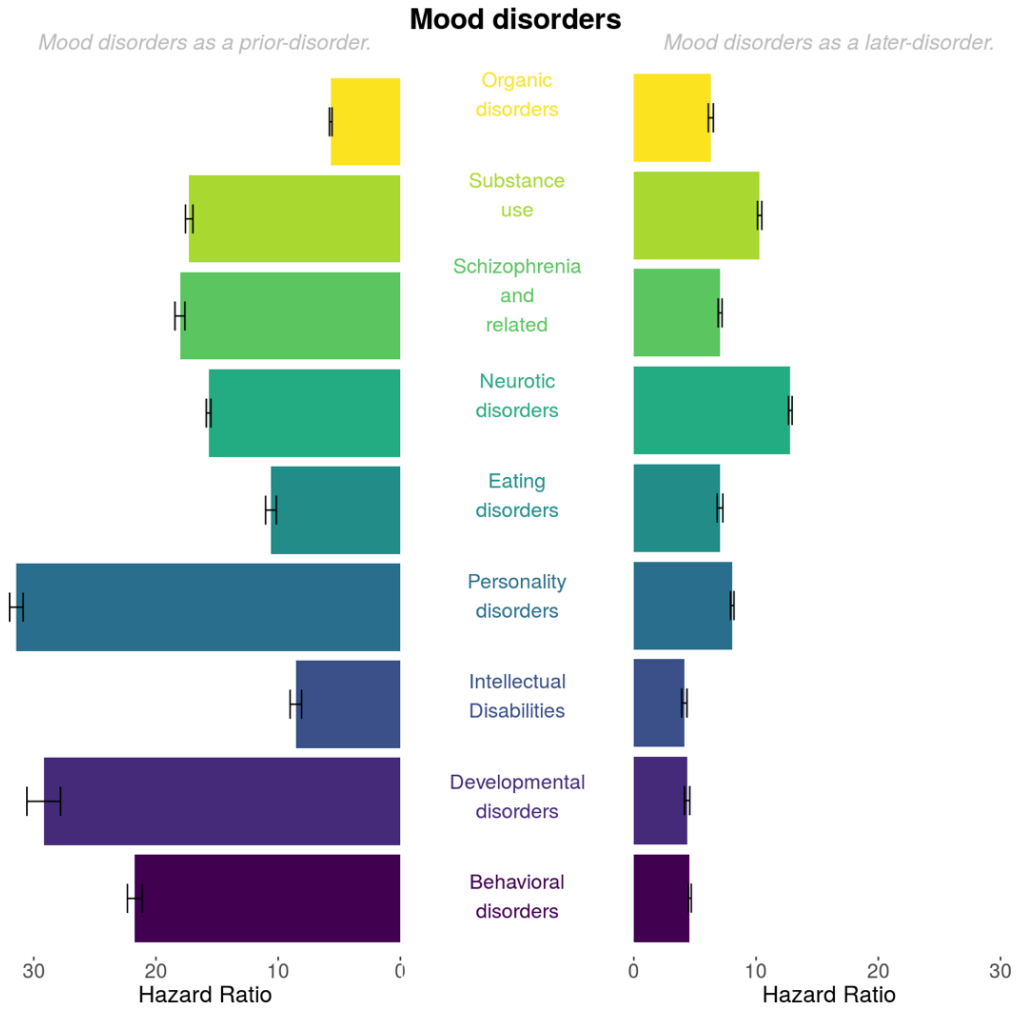


Figure 8 *: Description of the symmetry between Disorder Groups. Hazard ratio between all possible disorders and your selected disorder are represented on the left panel and vice-versa.*

+

Men | Women | All

Allow exploration

Dataviz Pipeline

Data prep

Charts

Interactivity

Button
Website

biobank uk

(N ~ 500k)

↓

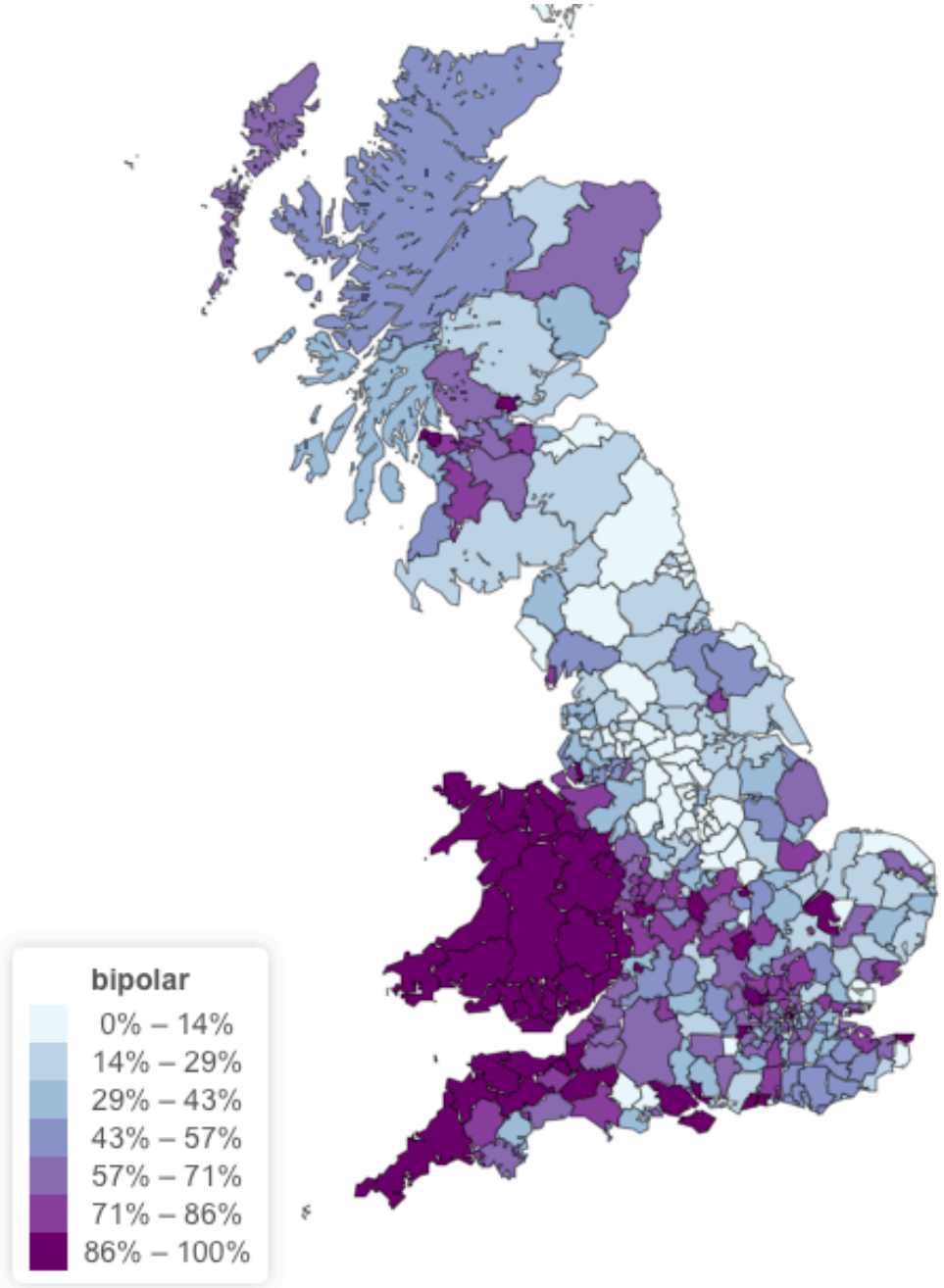
Compute PCs

↓

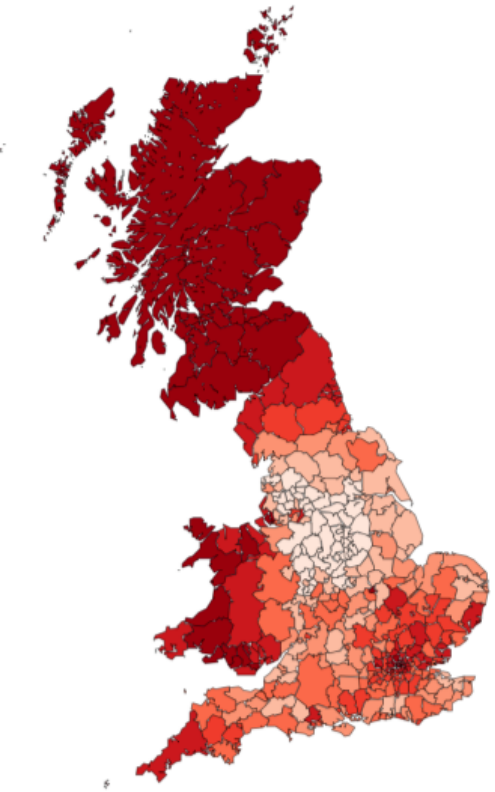
Compute PRS,
corrected or not

↓

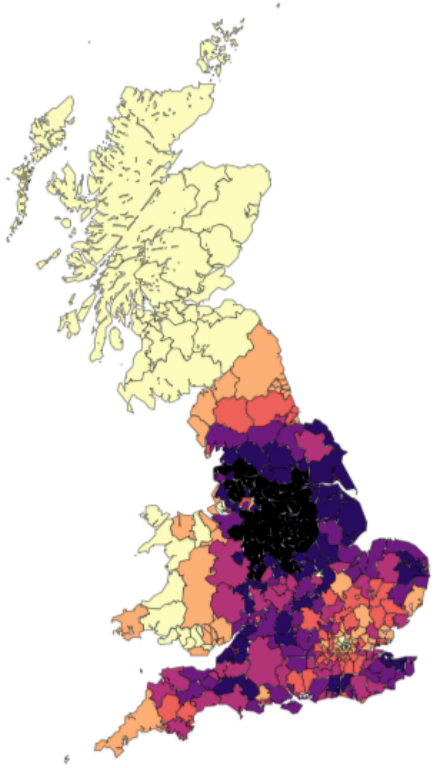
Map



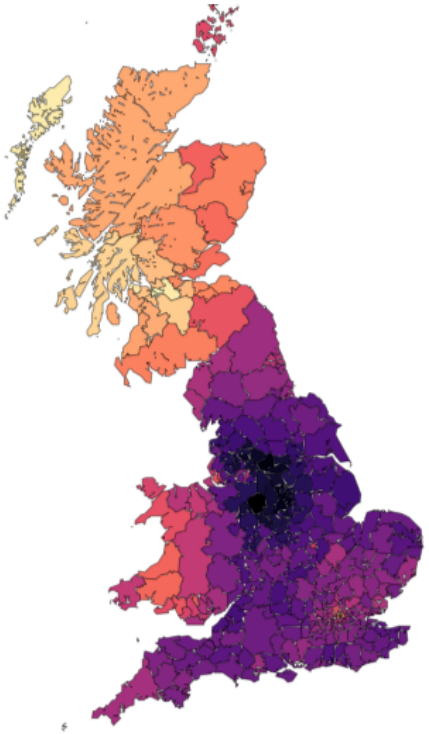
PC2 representation in the UK Biobank



Red



BuPu – 6 bins



BuPu – Numeric

Hexbin

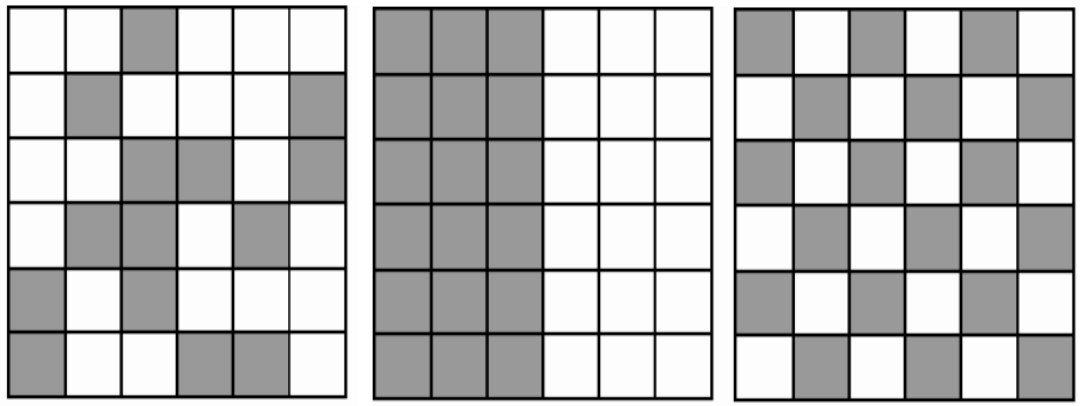
Cartogram

Compute Moran's Value

Spatially Random

Clustered

Dispersed



0

1

-1

Genes & Geography in great britain

Methods | Explore | Compare | Load your data | Share what you see

Load your data

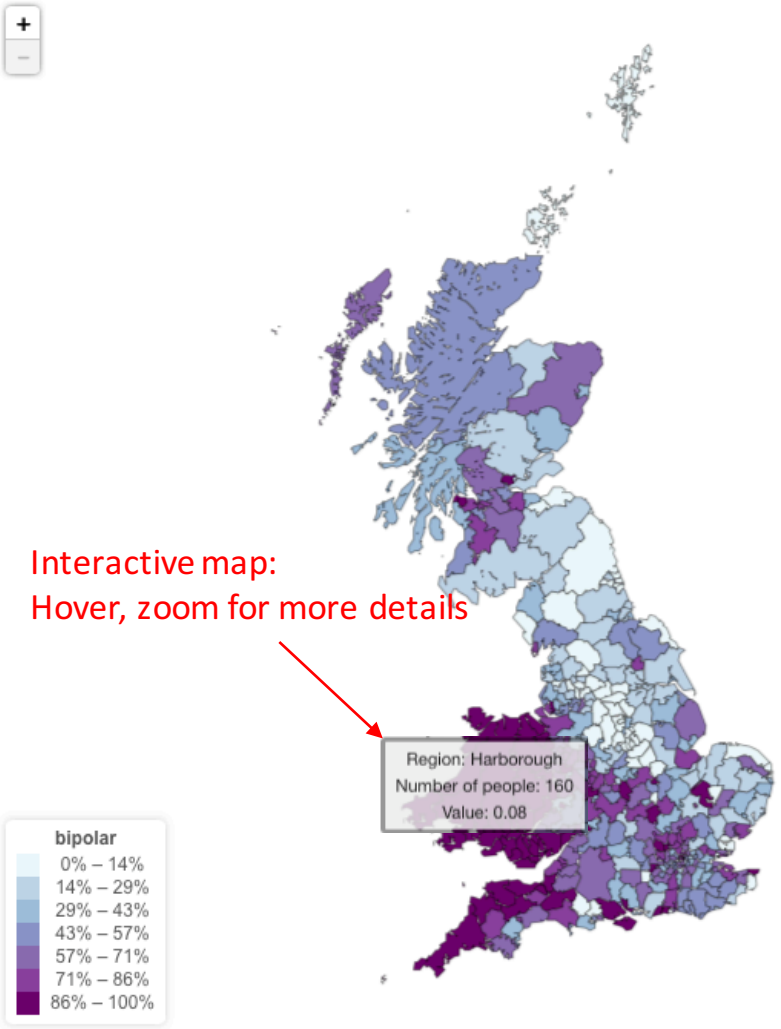
Custom the map

Welcome

This application describes the geographical distribution of several variable of the UK Biobank dataset (n=502630).

More than 100 variables are available for visualization. You can observe them using different geographical units. It is possible to custom and export this map following the surrounding buttons. Use the compare tab above if you want to study the relationship between several variables.

Interactive map:
Hover, zoom for more details



Variable

We propose to represent the geographical distribution of 126 variables. These veriable are split in several groups: Principal Components (PCs), Polygenic Risk Scores (PRS). To understand how these variable have been computed, visit the method section.

Pick up your traits

bipolar

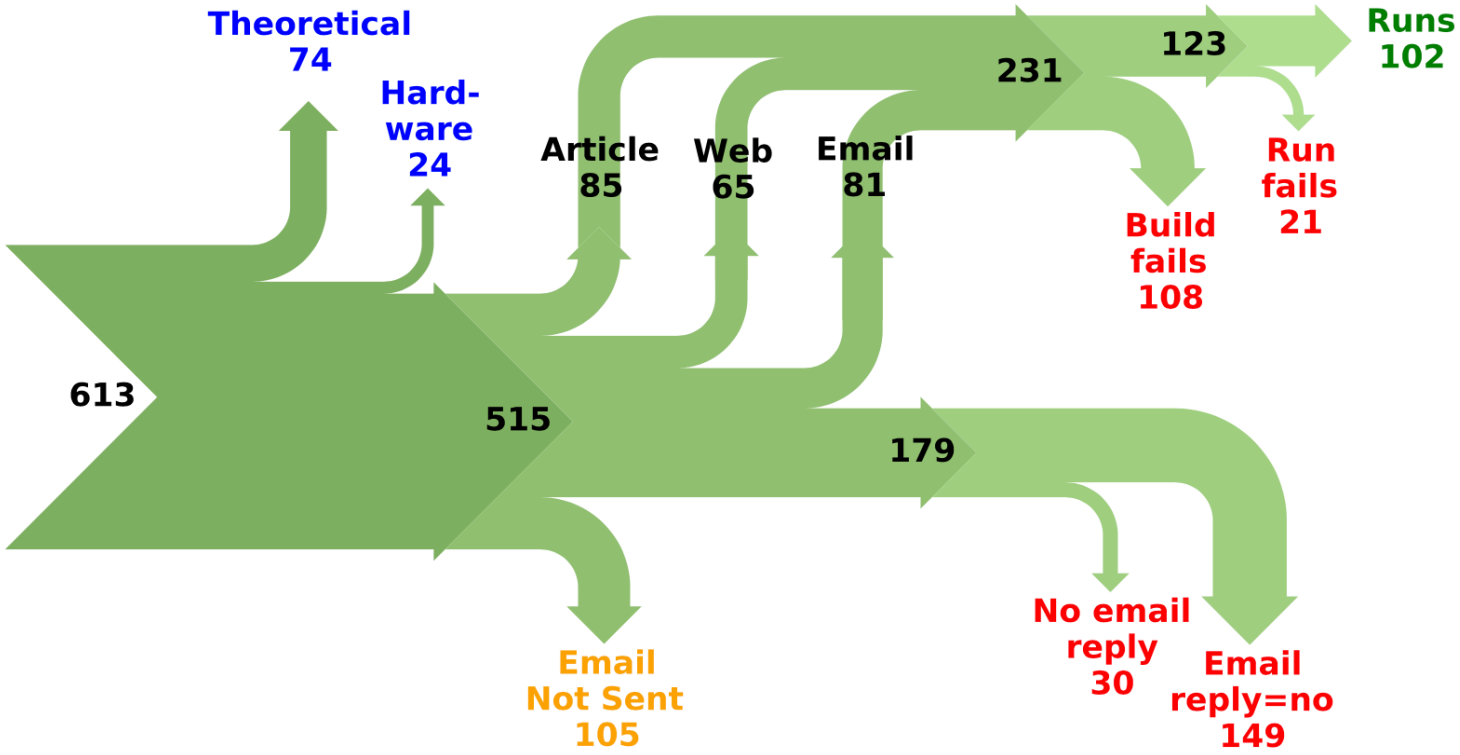
Moran's I value

Moran's I value is a measure of spatial autocorrelation, i.e. a correlation in a signal among nearby locations in space. Click on the plus button below to see what variables are the most clusterised

Figure 1: Geographical distribution of bipolar in the UK

Allow re-utilization

REPRODUCIBILITY IN DATA ANALYSIS



Collberg et al. 2014

An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden, Jennifer Seiler and Zhaokun Ma

PNAS March 13, 2018. 115 (11) 2584-2589; published ahead of print M
https://doi.org/10.1073/pnas.1708290115

Edited by David B. Allison, Indiana University Bloomington, Bloomingto
Susan T. Fiske January 9, 2018 (received for review July 11, 2017)

Article | Figures & SI | Authors & Inf

Table 1. Responses to emailed requests (n = 180)

Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	20	11
Asked for reasons	20	11
Refusal to share	12	7
Directed back to supplement	6	3
Unfulfilled promise to follow up	5	3
Impossible to share	3	2
Shared data and code	65	36
Email bounced	3	2
No response	46	26

https://twitter.com/cj_battey/status/974045242444820480

An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden, Jennifer Seiler and Zhaokun Ma

PNAS March 13, 2018. 115 (11) 2584-2589; published ahead of print M
https://doi.org/10.1073/pnas.1708290115

Edited by David B. Allison, Indiana University Bloomington, Bloomingto
Susan T. Fiske January 9, 2018 (received for review July 11, 2017)

Article | Figures & SI | Authors & Inf

Table 1. Responses to emailed requests (n = 180)

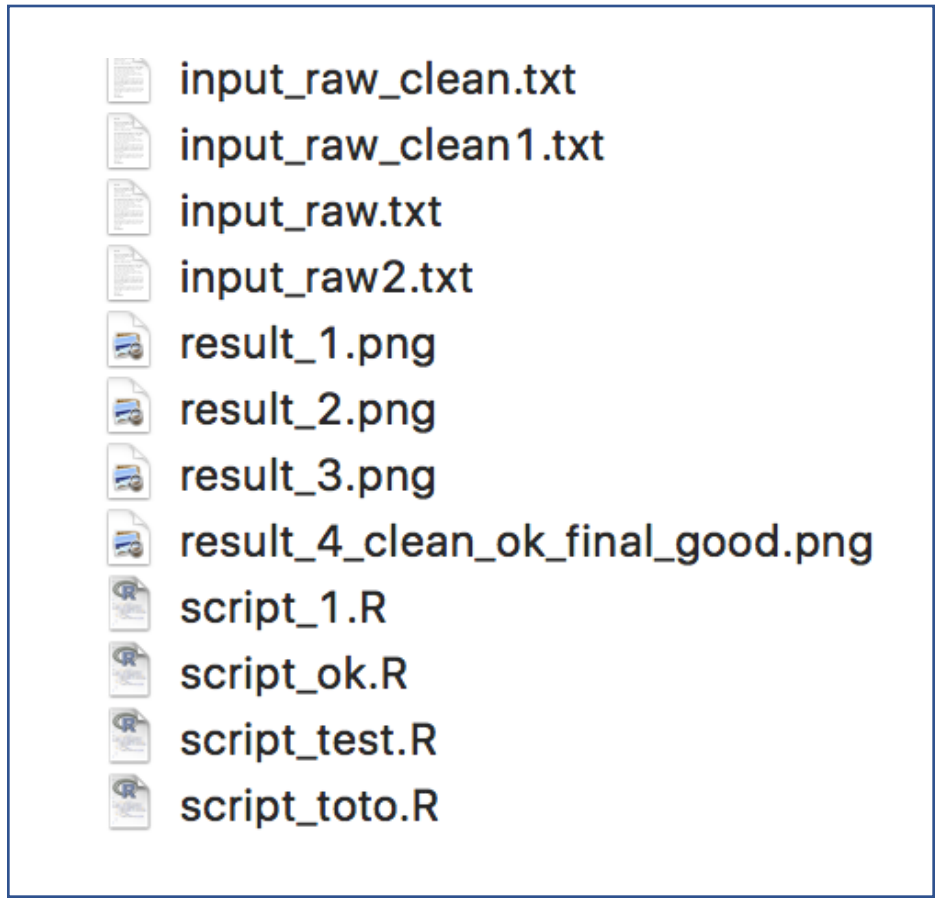
Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	20	11
Asked for reasons	20	11
Refusal to share	12	7
Directed back to supplement	6	3
Unfulfilled promise to follow up	5	3
Impossible to share	3	2

When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

Share your data

https://twitter.com/cj_battey/status/974045242444820480

Can you run your analysis again ?



A few best practices:

- Don’t modify input files manually
- Don’t copy and paste code / results
- Clean code (comment, use functions)
- Share your code
- Use version control

R-OpenSci Reproducibility guide

R Markdown

from RStudio

```
53
54 # An interactive manhattan plot
55 ***
56
57 Using `HTML` outputs you can embed some interactive graphics. For example, the
   plotly library can transform any of your ggplot2 graphic in an interactive
   chart:
58
59 ```{r, message=FALSE, warning=FALSE, echo=FALSE}
60 # Libraries
61 library(plotly)
62 library(tidyverse)
63
64 # Prepare the dataset
65 don <- gwasResults %>%
66 ```
67
```

Title

Text

Code

R Markdown

from RStudio

```
53
54 # An interactive manhattan plot
55 ***
56
57 Using `HTML` outputs you can embed some interactive graphics. For example, the
   plotly library can transform any of your ggplot2 graphic in an interactive
   chart:
58
59 ```{r, message=FALSE, warning=FALSE, echo=FALSE}
60 # Libraries
61 library(plotly)
62 library(tidyverse)
63
64 # Prepare the dataset
65 don <- gwasResults %>%
66 ```
67
```

Title

Text

Code

Program in Complex Trait Genomics

1 Loading data

2 Show data in a table

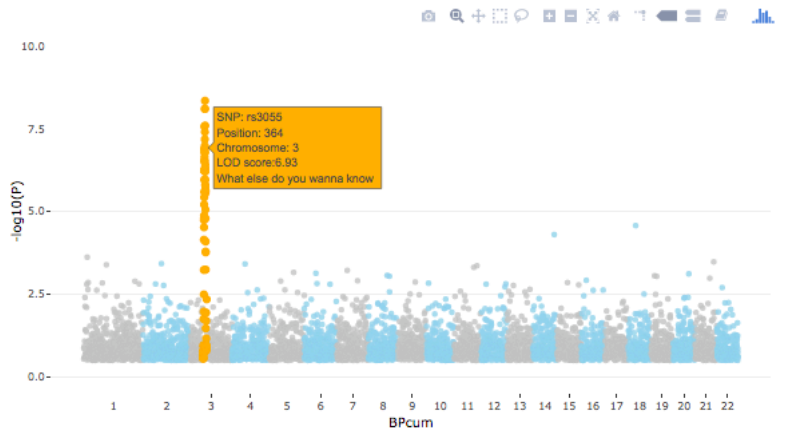
3 An interactive manhattan plot

3 An interactive manhattan plot

Using `HTML` outputs you can embed some interactive graphics. For example, the plotly library can transform any of your ggplot2 graphic in an interactive chart:

HIDE

```
# Make the plot
p <- ggplot(don, aes(x=BPcum, y=-log10(P), text=text)) +

    # Show all points
    geom_point( aes(color=as.factor(CHR)), alpha=0.8, size=1.3) +
    scale_color_manual(values = rep(c("grey", "skyblue"), 22 )) +

    # custom X axis:
    scale_x_continuous( label = axisdf$CHR, breaks= axisdf$center ) +
    scale_y_continuous(expand = c(0, 0) ) +     # remove space between plot area and x axis

    # Add highlighted points
    geom_point(data=subset(don, is_highlight=="yes"), color="orange", size=2) +

    # Custom the theme:
    theme_bw() +
    theme(
      legend.position="none",
      panel.border = element_blank(),
      panel.grid.major.x = element_blank(),
      panel.grid.minor.x = element_blank()
    ) +
    ylim(0, 10)

ggplotly(p, tooltip="text")
```



Interactive report

Is your code reusable ?

GitHub

Holtz Yan
holtzy

Data Analyst with a + in data visualization

Edit bio

Queensland Brain Institute
Brisbane, Australia
yan.holtz.data@gmail.com
https://holtzyan.wordpress.co...

Overview | Repositories 28 | Stars 22 | Followers 88

Pinned repositories

The-Python-Graph-Gallery
A website displaying hundreds of charts made with Python
Python ★122 15

R-g
A webs
R

Pimp-my-rmd
A few tips about R markdown
HTML ★11 2

epu
A clea
HTM

GenMap-Comparator
An application to compare genetic maps with D3 & Shiny
R ★8 5

Pul
Reproc
resista
HTM

452 contributions in the last year

May Jun Jul Aug Sep Oct Nov D
Mon
Wed
Fri

Learn how we count contributions.

holtzy / the-NB-COMO-Project

Unwatch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

A Shiny app describing comorbidity in the Danish Health Register

Edit

Add topics

20 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

holtzy Change 2 words — Latest commit e1a433e 21 days ago

DATA	Change 2 words	21 days ago
rsconnect/shinyapps.io/holtzyan	Change 2 words	21 days ago
www	John Correction	5 months ago
.DS_Store	John Correction	5 months ago
.Rhistory	First commit	5 months ago
README.md	Last version before sending to collaborators	5 months ago
global.R	Change 2 words	21 days ago
server.R	Change 2 words	21 days ago
ui.R	Change 2 words	21 days ago

README.md

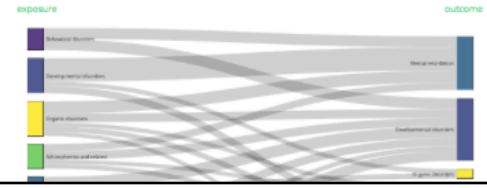
The NB-COMO Project

Overview

The NB-BOMO project aims to explore the patterns of comorbidity within treated mental disorders. It explores different ways to capture the complex patterns of comorbidity, notably through data visualization techniques. The first part of this project explores COMO within the Danish National Patient Registry (DNPR), one of the world's oldest nationwide hospital registries.

This repository gives the code of a web application that allows to interactively explore our results. It goes along with our peer reviewed publication (work in progress).

Here is a screenshot of one of the multiple visualizations proposed in the website:



Acknowledgment

John McGrath

Naomi Wray
Peter Vissher
Jian Yang

Oleguer Plana-Ripoll
Abdel Abdellaoui

Contact

@R_Graph_Gallery

github.com/holtzy/Talk

Yan.holtz.data@gmail.com

R-graph-gallery.com

Program in Complex Trait Genomics

NIELS BOHR PROFESSORSHIP PSYCHIATRIC EPIDEMIOLOGY

The University Of Queensland AUSTRALIA | IMB Institute for Molecular Bioscience

The University Of Queensland AUSTRALIA | QBI Queensland Brain Institute

The Scientific Paper Is Obsolete

Here's what's next.

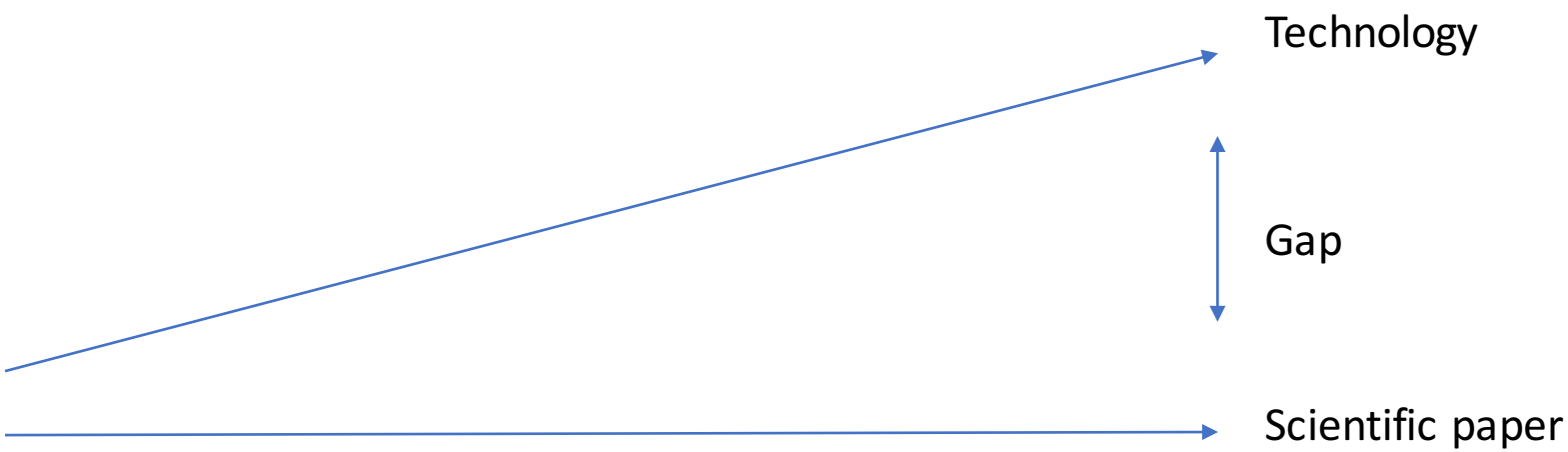
theatlantic.com

Observable is a better way to code.

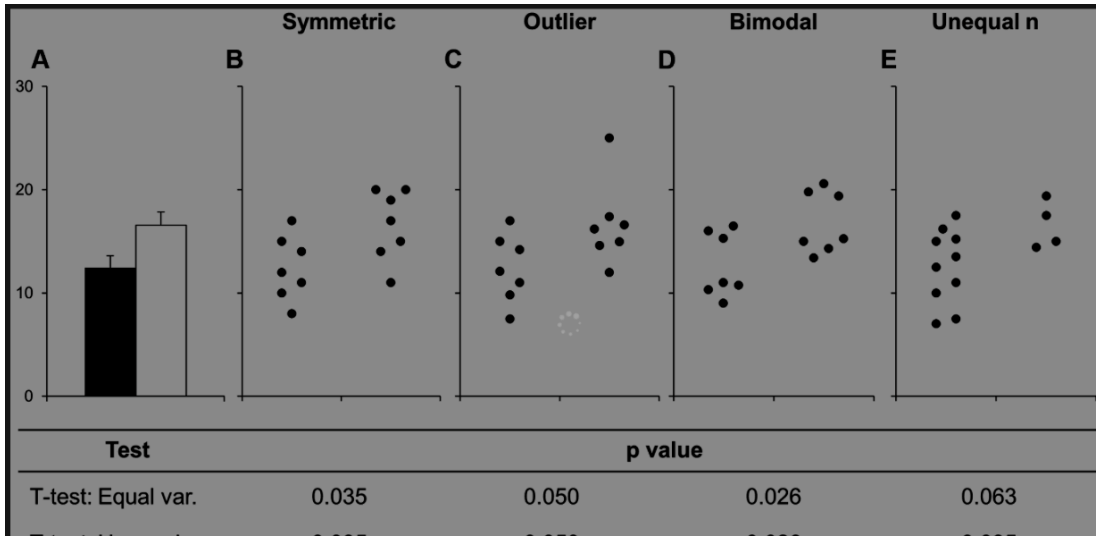
Discover insights faster and communicate more effectively with interactive notebooks for data analysis, visualization, and exploration.

Sign up for free | Try the scratchpad

https://beta.observablehq.com



Data viz: Avoid mistake and misleading figures



OPEN ACCESS

PERSPECTIVE

Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

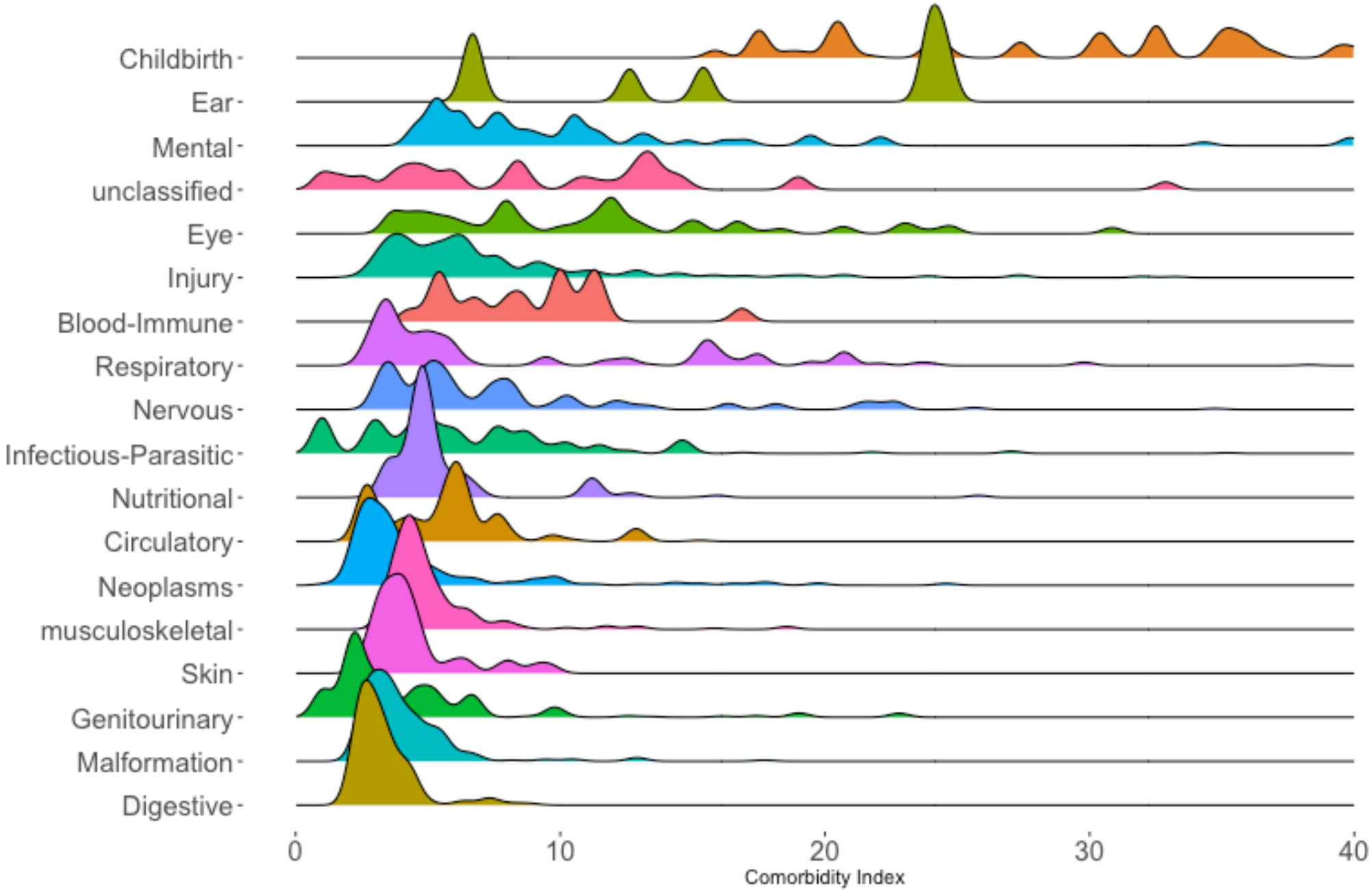
Tracey L. Weissgerber, Natasa M. Milic, Stacey J. Winham, Vesna D. Garovic

Published: April 22, 2015 • https://doi.org/10.1371/journal.pbio.1002128

Abstract

Figures in scientific publications are critically important because they often show the data supporting key findings. Our systematic review of research articles published in top physiology journals (n = 703) suggests that, as scientists, we urgently need to change our practices for presenting continuous data in small sample size studies. Papers rarely included scatterplots, box plots, and histograms that allow readers to critically evaluate continuous data. Most papers presented continuous data in bar and line graphs. This is problematic, as many different data distributions can lead to the same bar or line graph. The full data may suggest different conclusions from the summary statistics. We recommend training investigators in data presentation, encouraging a more complete presentation of data, and changing journal editorial policies. Investigators can quickly make univariate scatterplots for small sample size studies using our Excel templates.

Show violin plot average odd ratio per group


Childbirth
Ear
Mental
unclassified
Eye
Injury
Blood-Immune
Respiratory
Nervous
Infectious-Parasitic
Nutritional
Circulatory
Neoplasms
musculoskeletal
Skin
Genitourinary
Malformation
Digestive
0
10
20
30
40
Comorbidity Index