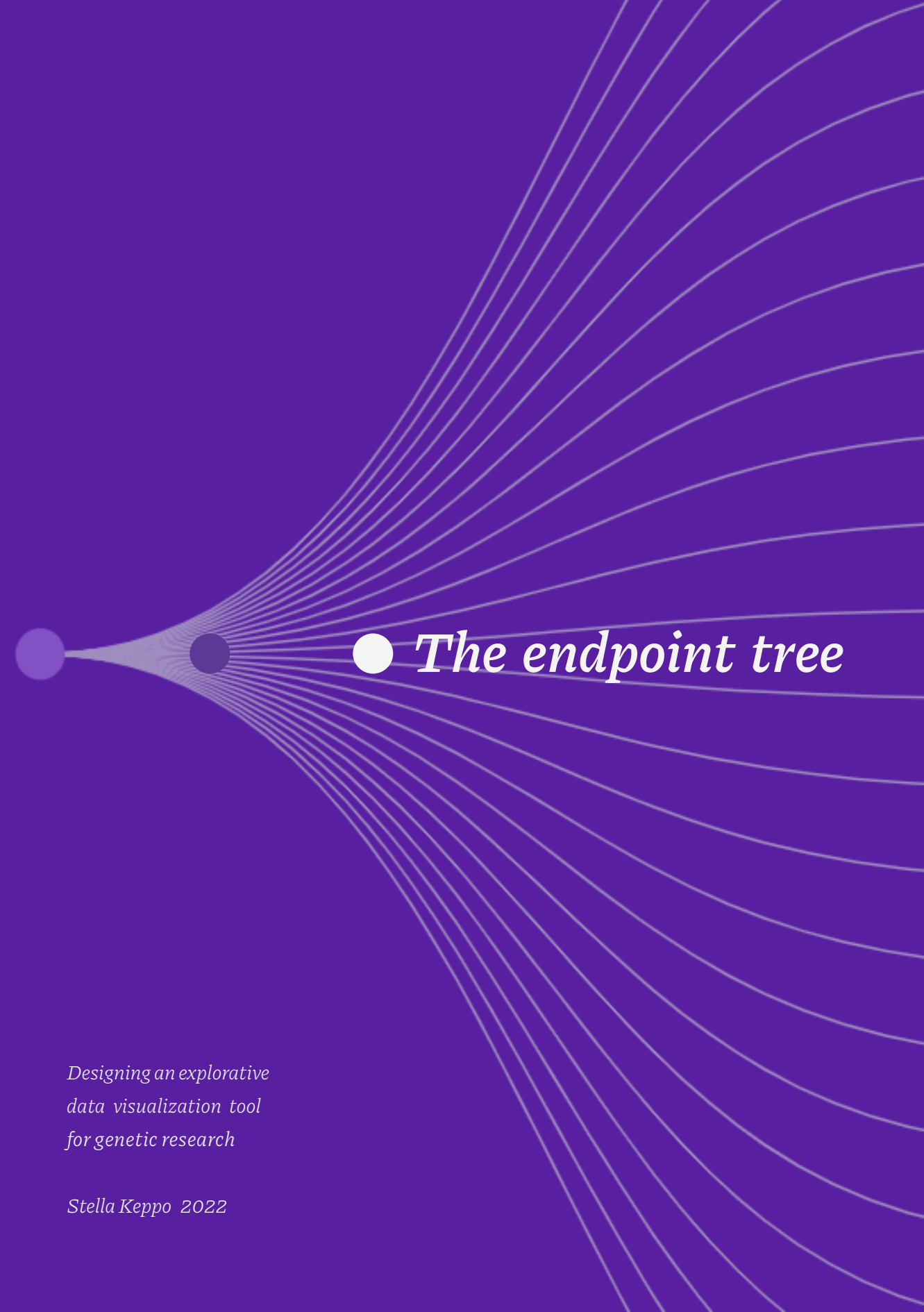


# The endpoint tree

*Designing an explorative data visualization tool for genetic research*

*Stella Keppo 2022*

*Stella Keppo*

*The endpoint tree: Designing an explorative data visualization tool for genetic research*

*data visualization*
*information design*
*interaction design*
*scientific data visualization*
*visual communication design*
*user - centered design*
*data visualization tools*

*Master's programme in Visual Communication Design, VCD*
*Aalto University*

*Thesis supervisor: Rupesh Vyas*

*Thesis advisor: Nicola Cerioli*

*Collaborative partner: FinnGen*

*Date: 28.04.2022* | *Number of pages: 113+13*

*Production URL's:*
*v1 - https://geneviz.aalto.fi/endpoint_browser/v1*
*v2 - https://geneviz.aalto.fi/endpoint_browser/v2*

## Abstract

*This thesis is done in collaboration with FinnGen, a global research project aiming at better understanding the relationship of human genome and health. FinnGen is utilizing vast amounts and varieties of data and is thus using scientific data visualization tools and collaborating with designers in order to remodel, design and implement these tools. This study explores user-centered data visualization practices in the context of browsing phenotypic data within FinnGen. Phenotypic data in this case refers to indicators of health events or outcomes defined using different data from health registries. These indicators in FinnGen are referred to as endpoints, and they are accessed through a phenotype-level data exploration system Risteys. Risteys is a public online platform where researchers acquire information about different endpoints including facts and figures about their definitions, statistics and their relationships with other endpoints. This study asks what kind of challenges users are experiencing while exploring phenotypic data in Risteys, and how these challenges could be answered by data visualization.*

*An understanding of best practices related to scientific data visualization tools within the topic of the study was formed through a literature review. A compact review of limits relating to reading and using data visualization tools imposed by cognitive attributes was also made. After acquiring this knowledge base, a small-scale user study was conducted with a pool of expert researchers to gather information on possible challenges relating to phenotype browsing in Risteys. Based on key findings from the user study, an interactive prototype was designed and produced to answer some of the challenges that the users were found to share.*

*This study establishes a targeted knowledge base on designing and producing an exploratory scientific data visualization tool from a visual communication design standpoint, and contributes to the ongoing discussion about data visualization processes within the confluence of the scientific community and the visual communication design field.*



## Tiivistelmä

*Tämä opinnäytetyö on tehty yhteistyössä FinnGen -tutkimushankkeen kanssa. FinnGen on kansainvälisen laajuuden geenitutkimushanke, jonka tavoitteena on lisätä ymmärrystä sairauksien syistä sekä parantaa niiden diagnosointia, ennaltaehkäisyä ja hoitomuotojen kehittämistä. FinnGen hyödyntää tieteellisiä datavisualisointityökaluja suurten ja monipuolisten datasettien visualisoinnissa ja tekee yhteistyötä visuaalisten muotoilijen kanssa uudistaakseen ja kehittääkseen näitä työkaluja. Tämä tutkimus tarkastelee käyttäjälähtöisen datavisualisoinnin käytänteitä fenotyyppi-informaation selaamisen viitekehyksessä. Fenotyyppi-informaatio tässä tapauksessa viittaa terveyteen liittyvien tapahtumien indikaattoreihin, jotka on määritelty pohjautuen erilaisten terveysrekisterien tietoihin. Näitä indikaattoreita kutsutaan FinnGenissä lääketieteelliksi päätepisteiksi, sairauspäätepisteiksi tai päätemuuttujiksi. Näihin päätemuuttujiin liittyvää tietoa haetaan FinnGenin Risteys -fenotyyppidatajärjestelmästä. Risteys on julkinen verkkoympäristö, jonka kautta tutkijat saavat tietoa päätemuuttujien määritelmistä, niihin liittyvistä tilastoista sekä niiden suhteesta muihin päätemuuttujiin. Tämä opinnäytetyö tutkii käyttäjien työnkulkua ja heidän kokemiaan haasteita liittyen fenotyyppidatan tutkimukseen Risteys -järjestelmässä, sekä kysymystä siitä, miten näihin haasteisiin voidaan vastata käyttäjälähtöisen datavisualisoinnin praktiikan näkökulmasta.*

*Kirjallisuuskatsauksen perusteella luodaan pohjustava ymmärrys koskien datavisualisoinnin parhaita käytänteitä tutkimuksen aihetta silmällä pitäen. Tiivis katsaus luodaan myös kognitiivisten prosessien luomista rajoitteista ja vaikutuksista koskien datavisualisointien lukemista. Tämän jälkeen suoritettiin käyttäjätutkimus, jonka tavoitteena oli kerätä tietoa käyttäjien mahdollisesti kokemista haasteista Risteys-järjestelmän kanssa työskentelystä; Käyttäjätutkimuksen tulosten perusteella suunniteltiin ja toteutettiin interaktiivinen prototyyppi, joka vastaa näihin haasteisiin.*

*Tämä opinnäytetyö luo kohdennetun tietopohjan koskien tieteellisen datavisualisointityökalun suunnittelu- ja toteutusprosessia visuaalisen kommunikaation suunnittelun näkökulmasta, ja siten ottaa osaa keskusteluun datavisualisoinnin prosesseista tiede- ja muotoiluyhteisöjen rajapinnassa.*

*Index*

# *Introduction*

## *Background and motivation for this study*

This research is done in collaboration with FinnGen, a global research project aiming to better understand the relationship of human genome and health by combining genome information with healthcare data from national registries. The goal of analysing this relationship is to have a more profound understanding of diseases and their treatments. FinnGen brings together a vast amount of institutions with University of Helsinki as the responsible organization[1]. The role of Aalto University's team in this project has been to design, develop and remodel data visualization tools that provide new insight for the researchers working within and in collaboration with FinnGen. I have been working in this data visualization team as a research assistant since spring of 2021, and both my experiences and observations during this time have sparked the interest in the topic of scientific data visualization tools and the processes included in their design and development. This thesis will discuss concepts from theories concerning cognitive processes relevant to visual design such as perception and attention as well as add input from realms of information design and user-centered design practices in order to establish a comprehensive knowledge base and explore how these persectives collide in the design process of a novel visualization tool. From these premises, this study will add to the discussion of a theoretical and practical base for the cross-disciplinary practice of designing data visualizations within genetics research.

The structure of this study consists of building a theoretical base for the case study through a literature review, conducting a user study to explore how expert users search for phenotypic data from FinnGen's phenotype browsing system Risteys[2] and building a prototype to support the findings of the user study. The documentation of the production process aims at establishing a more comprehensive view on the processes and practices from the visual communication design point of view.

The gap regarding analyses of data visualization tools in genetic research from an academic user-centered design and information design research perspectives has also served as motivation for writing this thesis. Contrasting this observation however, data visualizations have been dealt with increasing interest over the years in scientific publications, as also mentioned in the annual review of biomedical data science: "....data visualization has been a major research focus in computer science for decades.... yielding many resources that could accelerate discovery in biomedical research. Unfortunately, relatively few scientists currently use these resources"[3]. The same revision continues to state that "many biomedical data sets (often difficult or expensive to acquire) are inspected using poor visualization methods, even though better alternatives are known"[4]. This thesis argues that a design professional collaborating with scientists conducting a user study and exploring visualization solutions can lead to a better user experience and workflow through addressing possible usability issues as well as contribute to the strength of a design framework by building effective visual encodings and hierarchies. Within a parallel scope, it has been suggested that having a usability expert included in the process of system development in the healthcare field can shorten the iterative

*1. This information and more from www.finngen.fi/en*

*2, FIMM, Institute for Molecular Medicine Finland. (n.d.). Risteys FinnGen R9. Risteys. Retrieved April 6, 2022, from https://risteys.finngen.fi/*

*3,4. O'Donoghue et al., 2018*

design cycle and lead to fewer errors made by the users while using the systems[5].

5. Peute[a]et al., 2008

Implementing a visual communication design perspective into the design of visualizations and their interfaces does not only contribute to the look and feel of the displays, but with that extends to potentially having a direct effect on the user satisfaction and usability of said applications. Among others, Norman[6] has suggested that attractive things make people feel good, and thus make them better at thinking creatively and being more prepared to examine alternative solutions to solve a problem. This argument is originally discussed by Tractinsky in the context of positive affect; Both argue that a positive affect resulting from aesthetic appeal could play a role in improving users task performance[7]. Establishing a space where users can have greater potential for exploration and examining the data is of core interest for designing an interactive data visualization tool. These remarks serve as a starting point in arguing that values emphasised in design research and the knowledge of best practices in data visualization can yield to better end results within the field of scientific visualizations. As well as affecting the resulting visualization, these cross-disciplinary collaborations can be beneficial in establishing and building mutual understandings in developing better practices for the field.

6. Norman, 2004, page 19

7. Tractinsky et al., 2000, page 15

This thesis follows a common convention of terminology to refer to concepts of data visualization. The industry offers an array of terminology to refer to different genres inside the field, but this thesis will discuss interactive data-driven visualizations specifically rather than graphics that are used to illustrate more conceptual information such as events or natural phenomena; Thus, scientific data visualization in this context is also separated from scientific illustrations or models. Interactive data visualization tools can, in the context of this thesis, also be referred to as exploratory graphics. In exploratory graphics, the possibility to explore and analyse the data surpasses the need to create a rigid visual representation of the data establishing a narrative, as the very nature of interactivity implies a view that can be changed as a response to user input. This agile nature of interactive visualizations can be considered an overarching subject matter that affects the design process of such visualizations.





## *Foreword on collision and collaboration between science and design perspectives in the process of designing a scientific data visualization tool*

Data visualization is a highly cross-sectional field with a relationship to both sciences and design. This connection is evident in both where and how data visualizations are used as well as in the processes included in the design and production of the visualizations themselves. The inter-disciplinary elements of working in the field and producing custom interactive visualizations include working with data, interfaces, user interactions and programming methods. This intersecting nature provides fertile ground for direct collaborations between scientists and

designers or artists that might lead to results previously unknown or unexplored.

The benefits of deploying well-designed data visualization in the context of genetic research goes beyond the visualizations and their implications for the users, but extends to the organisation of the data itself. Especially considering cross-disciplinary project environments such as FinnGen where a lot of data flows through different professionals, platforms and tools, well-designed explorative visualizations might help researchers pick up previously undetected issues within the data or issues relating to the previous data analyzing methods; This is the very nature and a profound motivation of using data visualization in general. The phenomena is recognized in part by Shneiderman, Plaisant & Hesse: "Visualization can reveal data quality problems, which are common when repurposing clinical data for secondary analysis"[8]. Data is often passed through multiple instances and sometimes small inconsistencies or oddities can only be spotted when the data is made visible. Successful user interface and user experience design outcomes are also becoming of importance here; By encouraging the users to explore the data in different ways and enabling them to discover any abnormalities present in the data, these user experience considerations become an essential part in what we can consider as a successful outcome in the process of designing a scientific data visualization tool.

*8. Shneiderman et al., 2013*



## RQ1

***Are users experiencing challenges with FinnGen's phenotype browsing system Risteys?***

## RQ2

***How can a user-centered data viz design practice answer to challenges related to phenotype browsing?***

## *Best practices in designing data-driven visualizations relevant for the scope of this study*

This study will not cover the entirety of the extensive area known as best practices in data visualization; However, it will discuss some points of consideration, concepts and justifications relevant for the case study of phenotype browsing and the more general field of interactive, explorative data visualizations.

### *Selection design*

User selections are an inherent part of many interactive data visualization tools. Selection in this context means that one or more items of the data are selected through interaction methods with an intent to separate them or gain more information about them in relation to other data in the set. Selection design includes considerations on what kind of elements can be selection targets, through which interaction methods they should be able to be selected[9] and how the selection should be handled visually. Different interaction methods for selections can include buttons and input elements, direct mouse clicks on objects, key presses, mouse hovers and different brushing methods where a drawn area on the visualization returns a selection. Modifying and clearing selections as well as considering the possible need of multiple selections at once need to be considered as a part of selection design. Understanding the user's tasks helps determine the selections that are needed as well as to enable the comparisons, grouping or other tasks that the users might want to perform on the selected data items or groups.

*9. Munzner, 2014, chapter 11*

As well as considering the selection patterns and interactions themselves, the visual indications of selections should be considered. Nearly any visual encoding method could be used including a motion channel, but their ample separation of the encoding used in the non-selected elements should be ensured. This could be by either tuning the features of the selected objects or groups, the ones that are not selected or both; In other words, this could mean highlighting the selected items in different ways, dimming the non-selected items or using both in conjunction to achieve higher visual separation. The visual encoding methods used for the separation of elements to support visual searches are also called visual *popouts* or *tunable features*[10]. These will be discussed further in a later chapter reviewing cognitive tendencies in the context of visual searches.

*10. Discussed among others by Munzner, 2014 and Koponen & Hildén, 2019*

Filtering could also be considered a part of selection design; The users might want to filter the dataset in order to reduce either the items on display or the amount of attributes in each data item. Filtering is intended to manipulate the view of the dataset, and thus inevitably leads us to consider the layering or arranging the states in the manipulated view to support the users efforts in making comparisons between the pre-selection state and the selection display. On top of showing the change resulting from the filter as a jump cut, animated transitions or different zooming methods could be considered. The states could also be superimposed or split into multiple views[11], sometimes also referred to as small multiples. Displaying data as small multiples means separating the dataset into sections based on a variable of the dataset, and then displaying those resulting

*11. Munzner, 2014*

datasets as separate displays in close proximity to one another. Superimposing, on the contrary, refers to a composite view of different layers of data compiled into one image, where the focus is displayed within the spatial frame of the background that provides the context. The selection of the layout indicating the change caused by user selections is a balancing act between the available screen space, avoiding visual clutter and the need to support the user in making effective comparisons, and regarding the user-centered viewpoint of this study, the user's needs will effectively be a deciding factor between these possibilities.

*Designing with flexibility*

Within the context of this study, the concept of flexibility in terms of scientific data visualization refers to a need to build tools that can provide a dynamic view of different datasets or parts of a dataset by user demand as well as to provide a possibility to use and modify the display of rendered data in meaningful ways through interaction techniques. Meaningful ways of manipulating the visual representations of data for the user might include the flexibility to filter and explore the data, modify the visual aspects of the display to best show the current dataset as well assisting in the process of communicating the data to external groups of people when needed. When discussing big datasets, the concept of scoping the visual display according to the user's core interests before the data is rendered can also play a role in customizable and flexible visualizations. On the grounds of these remarks, aiming for flexibility in this context can imply the deployment of a wide range of different design solutions both in the structure of the visualization tools as well as their interfaces.

Considering flexibility also includes considerations on recognizing the need to visualise datapoints beyond exact data matches. By filtering, searching or selections, the user is setting the boundaries on what they are looking for, but sometimes they might still miss interesting or relevant information from right outside their search if it is not displayed at all; It is more difficult for the user to take into account information that is eliminated from the display and no longer visually detectable to make comparisons compared to information that would be still available through an alternative visual encoding. This issue is also recognized by Shneiderman, Plaisant and Hesse in a medical context as they note: "....visual search strategies to find patients with certain sequences of events such as suffering a stroke, receiving Coumadin (blood thinner), and then having severe headaches within 72 hours requires advanced visual analytics tools for search and to display the search results of exact matches and near misses"[12]. On top of considerations on query boundaries, this leads us into considerations on how and when to enable comparisons between different datasets and not only between the datapoints inside the currently analysed dataset. Cross-chart comparisons are not discussed in further detail here, but are worth mentioning in the context.

*12. Shneiderman et al., 2013*

*Data density, clarity and complexity*

Another consideration often accepted as best practice in data-driven visualizations is the concept of data density introduced by Tufte[13]. Data density refers to the information density of a visualization and thus calls for optimising space to allow the display to present high resolutions of data. Tufte has specified this resolution threshold to be at least 200 numbers represented per square centimeter[14]. It is to be noted, as Tufte also recognises, that the goal of promoting data density is not to compress the entirety of scientific data visualizations into their most data-dense format but rather to establish a goal of intensifying the resolution in which the data is shown. In their judgement, Tufte does not really concern themselves with the media and the possibility to layer or transform information without the constraints of a static ink dot when dimension and location as attributes linger on the borderline of being arbitrary. Data density as a concept still has its benefits especially when discussed together in consideration of attention and perception. Tufte defends the concept by arguing that data-dense displays allow the reader of the visualization to more easily "make a contrast, a comparison, a choice"[15]; Enabling these reading methods for the users should be the goal for the design process of any interactive data visualization tool.

Goals related to data density can be linked to goals of clarity and ease of reading that could be associated as characteristics of a successful visualization. It should be emphasized that the possibly complex data at hand does not need to be simplified in order to clarify it; In Tufte's words, "Quantity of detail is an issue completely separate from the difficulty of reading"[16]. Clarity, then, is not to be confused with simplicity, that is rather a stylistic choice than a given standard to work towards. Accepting complexity and defining a meaningful scope for it has also been referred to as *appropriate complexity*[17] in the context of the data itself; With any data, but especially with scientific data, the complexities of it should be respected everywhere where they have the possibility to carry meaningful information.

Considerations on visual aspects of the data displays presented can move the visualization to the direction of clarity. For example, unnecessary emphasis on non-data related visual attributes should be avoided if the emphasis does not aid the user or result in any meaningful readings to. One such practical use case can be found from gridlines. They should not be heavier than necessary in order to be able to guide distinction[18], and should not advocate a border or a barrier where there is none. Another such example can be found from labelling the data. Labels are a part of the data[19], and have great potential to reduce the amount of ambiguity in the visualization. Thus, working towards clarity does not have to advocate removing labels altogether, but rather considering design techniques to lighten the visual weight of the typography as well as allowing a dynamic display of the labels by enabling hiding or showing all or a portion of the labels. Visual proximity of a label or a legend in relation to the data marker encourages ease of reading despite adding material to the display.

13. Tufte, 1984

14. Tufte, 2006, page 48

15. Tufte, 1991, page 50

16. Tufte, 1991, page 51

17. Norman, 2010

18. Tufte, 1991, page 55

19. Tufte, 2006, pages 118-119

Legibility considerations will also play a part in aiming for clarity in data visualization. Legibility in terms of typeface selection includes paying attention to both the ample visual means for glyph differentiation and other attributes of the type such as glyph proportions, spaciousness, line width and contrast. Versability of the chosen typeface should also reach a limit where it can be used to build meaningful reading hierarchies. These legibility criteria can also in part improve the reading of the graphic by considering sensory differentiation of visual variables beyond type[20]. This increases user's ability to compare between visual coders and thus can enhance the reading of the visualization making it clearer for the user without the need to simplify the content.

20. Bertin, 2011, page 13

*Designing with consistency*

Consistency is a crucial factor when discussing principles of data-driven visual design. In this context, consistency is parallel to continuity, meaning that visual coding methods stay the same when coding similar information and thus do not change arbitrarily. Consistency can also refer to interface aspects, where users would expect similar elements and operations to produce similar results[21]. Changes in the consistency of visual encoding should only be motivated from changes in the content, and should not code differences where there are none. A change within the same attribute of the data that is encoded can be signified by a change in the visual variable that is currently used for that specific attribute rather than changing the encoding itself to signify change. Where a change in visual form communicates a change in data, it also prepares the data for visual comparisons; Without comparison there is no visualization to be made, and changes without roots in the data can mislead and confuse the user of the visualization. Therefore, consistency as a design principle is of primary importance in any data-driven visualization.

21. Sharp et al., 2019, page 29

*Coding data in colour*

Colour in data visualization is best discussed through its physical attributes of hue, saturation and luminance as well as the different use cases for color in data visualization. According to Tufte, fundamental uses of colour in information design include labelling, showing measure, representation and adding to the visual beauty of the image of information[22]. In all of these areas, the best solutions for each visualization project should be weighted in according to their specific circumstances. It should be kept in mind that readers of the visualization can assume meaning to colour in each use case even if there is none[23], so it should be made clear why colour is used and what information or aspect from the data it is encoding. This is especially the case when using multiple different hues, since the user can, and likely will, assume a change in the quality of the data according to the change of the color encoding.

22. Tufte, 1991, page 81

23. Discussed also in Koponen & Hildén, 2019

When colour is used to code measure in the data visualization, some widely accepted best practices regarding the use of different color scales should be taken into account. When coding univariant data via a colour scale, it is best to harness

its character as a natural quantifier and use a sequential scale to encode orders in the data. On the other hand, if the values are divergent, a divergent colour scale should be used. Both divergent and sequential colour scales provide an intuitive way of encoding data due to decrease or increase of the luminance or saturation that act as a magnitude channel coding implicit perceptual ordering. In the scales where different hues with equal luminance are used, this hierarchy is arbitrary[24] and as such needs to be explained through a legend or other means. When using colour as a magnitude channel to code attributes of data through luminance or saturation, it should be noted that the number of possible steps in the scale that can still be discriminated from one another is relatively small, and the use of luminance as a coding channel should be re-considered if the number of separable bins gets closer to or over five[25] different ones. Luminance contrast still stands as a better choice for encoding fine detail over only using different hues in general, since "luminance contrast is required for edge detection in the human eye"[26]. Possibly the safest choice for achieving the best possible visual salience in colour coding would be to differentiate with both hue and luminance, but certainly not to rely on hue alone.

*24. Discussed also in Maciejewski, 2011 and Koponen & Hildén, 2019*

*25, 26, 27. Munzner, 2014*

There are several unstabilities regarding colour usage in the context of digital displays and data visualizations that are worth mentioning. These include considerations on different device settings such as dark display modes, brightness and colour temperature settings as well as the varied ways in which displays reproduce colour information. Environmental and other external factors, such as light conditions and especially the most common forms of colour vision anomalities also play a part in perceiving colour. Preferring luminance contrast to code data instead of relying on the users to detect hue changes also works in favour of taking account these colour vision anomalities. Summarizing these considerations, the perception of colour will always be a sum of different factors and can not be estimated or designed with perfect accuracy.

Considerations regarding colour should also extend to how it relates to other visual coding methods on the data display. Colour in visualizations is tied to its spatial representation meaning that the shape and size of the visual marker that deploys colour both effect and enable its appearance; One example that can be given is that colour is harder to perceive the smaller a visual marker containing the colour gets[27]. On the other hand, if markers or areas that use colour encoding get bigger and occupy a relatively large area of a display, their potential to disrupt the visual balance of the complete view should be considered[28]. On top of this, deploying multiple hues predisposes the visualization for illusions such as simultaneous contrast and cognitive countours[29].

*28. Tufte, 1991, pages 82-83*

*29. Tufte, 1991, pages 92-95*

Regarding the reading of colour as a visual encoder, semantic colour schemes should be used where possible, implying that if there exists a natural representation of the element that is encoded, it shouldn't be overlooked as they posess a "certain definite authority"[30] as the origins of the color association. Using naturally intuitive color codings has the potential to ease the reading of the visualization since the user would not have to

*30. Tufte, 1991, page 90*

memorize or keep checking the codes for the color dimension. On top of purely semantic meanings, there is an emerging area of conversation about a wider range of affective potentials of colour in scientific data visualization[31] that pays closer attention to expressive qualities of colour and their potential benefits for the visualization from a user-centered standpoint. Even if this perspective raises questions on objectivity and the responsibilities of the designer, affective dimensions of colour cannot be completely dismissed and they should be taken into consideration at the very least when deciding the strengths of the hues; bolder and brighter hues will be sure to attract attention first compared to more demure ones[32].

*31, 32. Samsel et al., 2018*

Lastly, the effective use of colour in data visualization often depends on the background it stands against. Being able to highlight, guide or separate through colour depends as much from where they are separated from than their own attributes; Different neutral shades with little to no hues other than those on a greyscale are a popular choice for enabling other hues to shine in data visualization.

*Animation and movement*

Motion design is only one of many visual encoding channels used for interactive visualizations, but it is an exceptionally powerful tool to grab visual attention and as such will be briefly recognized separately from other visual coding channels. Animations have the potential to both distract and guide the users of an interactive visualization, and they should be used sparingly and with purpose. Motion can be difficult to ignore, and should only be used to guide attention where it is needed and not make the users choose their focus between many moving items, unless their reading as a group according to their direction or other type of animation would benefit the reading of the visualization. This refers to similar type of movement or a direction of movement binding the items to a singular group of association. Another possible benefit of animation strategies is the use of movement to target change blindness. Change blindness means that visual changes can be hard to perceive if attention is not being paid to those changes[33]; Movement can be used to minimize the possible effects of change blindness especially when another possibility to compare changes caused by interaction is not provided. It has been suggested that well-designed animations can indeed improve perception of change[34], since movement during change updates provides a focal point for attention that has the potential to help the user to follow up with the change instead of working towards detecting it through comparison or relying on their working memory. Movement as a narrative element in data visualization is not discussed here, as the framework of explorative data visualization tools aims to support an open area of exploration rather than guidance through a certain narrative.

*33. Itti et al., 2005, chapter 13*

*34. Heer & Robertson, 2007*

## *Visual structures*

This chapter will briefly discuss some concepts relating to the visual structures of exploratory graphics relevant for the production of this study. Such visual structures include interactive trees and networks that represent connectivity between the datapoints, as well as their related interface aspects.

### *On trees and networks*

With a tree structure, this study is referring to the visual representation of a structure that implies the existence of subgroups that stem from a common source. This type of structure could also be referred to as a hierarchical structure, and as such trees especially hold a position as well-established representations for different kinds of taxonomies. They are often represented with a node-link diagram, where a node represents an individual item in the tree or a network accompanied by a link (also called an *edge*) that communicates information about the relationships between the nodes. This is by no means the only way to represent hierarchical, tree structure data; Other visual arrangement methods include treemaps, lists and radial arrangements such as sunburst charts[35] among others.

*35. Munzner, 2014, chapter 9*

Other considerations on the visual form of the tree structure relate to whether the data can be represented by a single tree with a singe root node or whether splitting the representation to a multi-tree structure could be beneficial. The depth of the tree, in other words the amount of levels in the hierarchy, will also play a part in determining the most appropriate visual form. Relating to the depth of the tree, the hierarchy information can be controlled through a collapsible visualization structure where the user can expand and collapse level information in the tree. A collapsible structure might require considerations on the navigation aspects as well as what kind of a view the user is given on the paths or connections relating to the focus location or a node in the tree. With a hierarchical structure, the variations in visual weight of individual nodes, links or levels can cause an interpretation of changes in hierarchy, so they need to be paid close attention to.

### *Critique of trees*

The hierarchical top-down structure of a tree imposes a limitation on its flexibility as a visualization method, as they don't respond well to complexities or inconsistencies of the connections in the information that is being represented. It has also been argued that trees can be associated with being centralistic and thus expressing an "unequivocal concentration of power"[36]. Hierarchical structures can be perceived as implying a power structure even when there is none, since a more central or upper node or level could imply an indication of being more relevant or of greater importance in relation to the structure even if this would not be the case. Another way to describe the limitation of a tree structure could be to refer to them as being stagnant and rigid, and as such might steer towards establishing hierarchy and structure by compressing information into a tree when it is not necessarily the most accurate representation on the

*36. Lima, 2011, pages 44-45*

underlying structures at hand. It should be acknowledged, too, that taxonomies themselves are imposing a tree structure to knowledge that is gathered from a multivariate world, and that the issues possibly underlying in a tree architecture stem from not only the visual representation, but the data itself and how it is gathered, categorised and compressed. This is not to say that taxonomies would not be helpful in arranging information and especially locating and storing information, but their limitations in this sense have to be addressed nevertheless.

*Edges as information containers*

It is to be noted that not all tree structures encode the parent-child relationship of nodes with an edge between them; Some simply display the relationship without imposing any additional variables describing that relationship (*see figure 1*). This poses implications to the consideration on whether there will attributes that imply relative distances between the items in the data or whether they will simply exist in a structure where only their parent-child relationship is significant. Area codings utilized in tree representations excluding nodes are not completely neutral either, but instead will spark other avenues of consideration including their size, colouring or placing their borders in relation to other nodes.

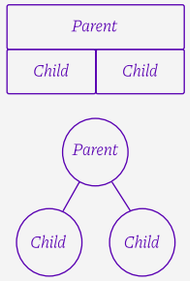



*Figure 1. Simplified example of a tree structure without edges (above) and with edges (below)*

A line as a visual encoder is a spatial element like any other shape used as a visual marker, and its length and width can code attributes of the information on top of any other coding methods that might be added. Lines in node-link diagrams are not often utilized to their full potential as information containers, as the spatial vertical or horizontal positioning of the nodes and links is more often a result of a layout algorithm in action rather than a result of any manual consideration on the attributes or relationships that are encoded with the links. On top of spatial positioning, the internal hierarchy of the nodes, the categories of their connections or other attributes relating to their connectivity might get coded with gradation of different values including colour, hue or a type of a connector line [37]. This method was also utilized and described by Munzner in relation to the TreeJuxtaposer tool: "Unmarked edges are rendered in lighter shade of grey as they are further away from the root. The resulting brightness gradations provide a redundant coding of topological information"[38]. In the context of this case study, it is also recognised that separate trees might have relationships with one another, so that necessarily affects the design space as well.

*37. Lima, 2011, pages 83-88*

*38. Munzner et al., 2003*

*Navigation within the visualization showing connectivity*

With navigation inside a visualization exploring connectivity in the data, this thesis refers to maintaining the user's orientation while exploring the dataset and related interface components and actions of said exploration. There are some suggested strategies for supporting user's navigation that will shortly be discussed here.

In their paper of a tree visualization tool TreeJuxtaposer[39], Tamara Munzner discusses a global distortion approach where

*39. Munzner et al., 2003*

a full dataset is present and visible at all times. This approach is promoted for it guarantees visibility of the data and thus supports navigation. A colourful range of expressions can be found to describe the same effect such as *stretch and squish* or *rubber sheet navigation*[40], but the main takeaway is that all of the data is to remain visible, although possibly distorted relating to the point of interest and the view area so that none of the information is completely cut out of the view. Along the lines of similar considerations, Furnas discusses a *fisheye view*[41] meaning setting the initial scope of the data presented depending on a meaningful framing of the data for each user, while maintaining a context of the global structure. This refers more to the direction in which the user is orienteering in the dataset, the direction being from the overview to selected scope rather from selected scope to overview. In a global distortion approach the amount of distortion is to be adjusted based on relative importance relating to the focus scope, also discussed as *degree of interest* (also reffered to as *DOI*)[42]; Meaning prioritizing interesting or most relevant elements in the structure in relation to the distance of the fisheye, or the chosen location in the system. As described by Furnas: "Fisheye views provide a balance of local detail and global context by trading off a priori importance against distance"[43]. Degree of interest approaches can help tackle visual clutter or the issue of far away items not being visually detectable resulting from the distortions in a guaranteed visibility approach.

*40. Munzner, 2014*

*41, 42, 43. Furnas, 1986*

Other alternative angles to maintaining orientation and context could also be observed through using a minimap or a path that maintain a context of a bigger whole while culling some of the context in compromise of a view area size. One practical example can be found from a geometric zooming approach where the user navigates through panning the zoomed in view where the elements that are closer occupy a relatively bigger size of the view area compared to elements that are further away or, in the case of 2D viz, completely out of frame.

All of these perspectives balance between visibility and context, as well as asking the question of how much of that context is necessary for the user to maintain their sense of orientation and the understanding of interconnectedness of the nodes in a network visualization. Visibility and occlusion will also relate to these questions, since often avoiding visual occlusion includes a tradeoff with the available screen space. Determining best strategies will, again, be dependent on users tasks in the environment.

*Interface aspects*

When designing interface elements through a user-centered design framework, the designer should consider the existing knowledge that the users might have on different signs and their possible meanings, in other words to consider the ontologies that the users might be familiar with in order to extract meaning from the signs. These realms of knowledge can be related to previous experience from using websites, applications or institutional systems as well as shared cultural knowledge realms or other real-life references. A paper[44] suggests that the

*44. Islam & Bouwman, 2016*

user's experienced level of complexity in terms of interpreting interface signs varied between these domains. The ensemble of signs what is known as an interface, however, is rarely a single sign without context. The meaning of the signs in an interface is an interpretation of each sign in relationship to the context in which it is represented together with other signs. Individual signs' reading in terms of meaning can be strengthened by applying a *dual-coding theory*[45] where reasonably possible; A dual-coding theory suggests that double-coding signs through both words and images (*verbal* and *non-verbal channels*) makes them more memorable.

*45. Sadoski & Paivio, 2012*

Level of communication between the user and the system will also be of significance in terms of navigation, reading of the interface signs and the overall user experience. This communication is not a responsibility of the machine or the system itself, but is enabled by the designer through visualization interface[46], and should be considered an integral part of the interaction design. People and machines communicate differently; This juxtaposition has also been aptly described as a "species clash"[47]. Narrowing this gap and thus minimizing frustrations from lacking the means of communication should be among the goals for an agile user-centered framework for scientific data visualization design.

*46. Norman, 2010, pages 111-119*

*47. Norman, 2010, page 6*

### *Design considering cognitive tendencies*

Forming an understanding of designing effective interfaces and visualizations requires us to turn our attention to cognitive attributes and their implications for visual design. Efficiency in the context of this thesis follows the definition by Bertin: "Efficiency is defined by the following proposition: If, in order to obtain a correct answer to a given question, all other things being equal, one construction requires a shorter period of perception than another construction, we can say that it is more efficient for this question"[48]. Of course, "all other things" will rarely be equal in a realistic user scenario within the area of interface design, but this is the general thought that describes the objective of efficiency for a visulization tool. Combining perspectives from cognitive studies to considerations of best practices from the visual communication design field will establish a more solid ground for the success and efficacy of the visualization tool.

*48. Bertin, 2011, page 9*

*Scope of perception and memory*

Limitations of human perception and memory affect how visualizations would best be constructed, and therefore some key aspects will be discussed here.

Successful information design should call for economy of perception[49]. Economy of perception refers to design that aims at utilising natural tendencies of cognitive processes and perception thus narrowing the bandwidth of cognition needed to process the visualization. As the annual review of biomedical data science stated in 2018, "A successful tailored visualization arranges all relevant data into a compact, immediately accessible two-dimensional (2D) view. This facilitates spatial reaso-

*49. Tufte, 1991, page 29*

ning, which, in turn, reduces the cognitive load needed to read a visualization and gain insight from data. By using familiar or intuitive visual conventions, successful visualization strategies also reduce the cognitive load needed when first learning to how to read them"[50]. To underline these considerations, designing compact, information-dense visualizations that allow effective readings and comparisons while minimizing ambiguity through clear guidance regarding the visual encoding methods will help reduce their cognitive load.

*50. O'Donoghue et al., 2018*

Compact and data-dense does not have to mean cluttered and distressing; Allocating visual breathing space to the visualization helps in highlighting things that are important and supporting effective analysis. Visual breathing space can be achieved, among pre-determined scoping of the rendered data and its attributes, as a result of *progressive disclosure*. Progressive disclosure referes to allowing the information to be revealed gradually, for example in the case of data-driven tools by user demand. Reducing the amount of items in a display through item or attribute filtering through user interactions would be another solution. These can work as a strategy to reduce the amount of items on a display without compromising the amount of data in an exploration space. More objects rendered in a data display will cause an increase in the timeframe needed for the user to make decisions, as outlined by Hick's law where a time required to make a decision is increased together with the number of variables on display[51]; Thus, opting to enabling comparisons between the most relevant objects and lightening the visual display will be of interest when considering congnitive-conscious visualizing strategies.

*51. Lima, 2011, page 92*



The number of variables to be decoded can also be discussed in the context of the short-term visual working memory; It has been found that the short-term memory is able to hold up to around seven items at a time[52]. It has also been suggested that these items disappear from the reach of cognitive functions at a relatively fast rate[53]. This does not mean, however, that the amount of items on a display in a visualization should be constrained according to these numbers specifically, especially if the users can easily return back to these items and not hold them in their short-term memory to be able to perform tasks on them. The tasks related to comparing or contrasting elements through visual inquiry are some of the most relevant to consider in terms of data visualization, so these aspects of memory should not be left without consideration when designing how to enable comparisons within the data display. A study conducted by Alvarez and Cavanagh[54] suggests that both the visual information load of each particular item up to inspection and the number of those objects impose capacity limits on visual short-term memory. In other words, they suggested that the visual information load increases with the increase of information content of the objects themselves; Thus the greater the information load of each item, the fewer items from that class one can hold in memory. They found that while the visual short-term memory capacity varies between different objects, it is still expected to reach a gap at four or five items at a time even if little to no information is being presented by each item.

*52. Miller, 1956*

*53. Ritter et al., 2014, page 149*

*54. Alvarez & Cavanagh, 2004*

The nature of the working memory as a temporary storage space for items to work with also imposes another immediate area

of consideration relating to attributes of cognition called *primacy*; Meaning that the items that appear for the users first, for example on top of lists, are easier to remember[55]. The reasoning for arrangement of items in lists or groups should therefore be evaluated keeping primacy in mind, and aim at showing the most relevant items that respond to the arrangement criteria best first. As well as paying attention to the primacy principle, reordering the data, also called *sorting*, according to user inputs utilizes the most efficient visual encoding channel of spatial position and as such can help with different kinds of tasks related to the reading of the data[56].

*55. Ritter et al., 2014, page 149*

*56. Discussed for example in Munzner, 2014, chapter 11*

*Visual searches through cognition*

There are some factors that affect how easily humans find objects through visual searches. Our visual system is not capable of perceiving the entirety of the information in our field of vision all at once, but only a fraction of the full scope of vision is allocated our conscious attention. This narrow area of focus is also referred to as *foveal vision*[57]. This attention is constantly shifting; by understanding this and the limitation it imposes upon visual design, better guidance and visualizations can be produced. Some of the factors affecting the design of data-driven visualizations in the context of visual searches are pre-attentive features, directed top-down searches and the effects of familiarity on search behaviours[58].

*57. Discussed for example in Cairo, 2012, chapter 5*

*58. Ware 2004 /2013, p. 143 as cited in Koponen & Hilden 2019, p. 48*

First of these factors is that there are some features that our visual system is naturally specialised in recognising pre-attentively, meaning that the processing of the information happens faster than what we are able to direct our conscious attention at. These pre-attentive features guide visual searches; "Because it is not possible to match all of the input to all of the possible objects at the same time, selection mechanisms, guided by a set of preattentive features, govern the transfer of a limited amount of information from one massively parallel stage to the next"[59]. In other words, selective focusing needs to happen in order to direct and focus attention. Selective focusing in this context is not mindful or conscious, but is not random either; As declared before, pre-attentive processes guide this focus that in turn facilitates visual searches. A comprehensive listing of all the visual features that prompt pre-attentive attention does not exist, but among those relevant for data-driven visualization are colour, size, motion, shape, depth cues and opacity[60]. Relating this information to the information on short-term working memory, it is best practice to restrict the amount of visual features that are up for processing to a minimum. There are empirical study results that suggest that the amount of different colour hues that can be used to support effective visual queries would go up to five different hues, after which their identification becomes more difficult[61]. These results work in paraller to the maximum amount of separable colour luminance bins discussed in the context of sequential colour scales.

*59, 60. Itti et al., 2005, chapter 17*

*61. Healey, 1996, pages 263-270*

Then there are behaviours called top-down searches that affect our ability to spot specific information regarding what we are searching for; In other words, we might say that we are more focused on finding specific information and not just browsing

openly. Top-down search behaviours are contrasted by bottom-up processes where something that appears on our visual field grabs our attention as opposed to searching for it[62]; This is the process occurring with the tunable features discussed above, or when we are noticing something familiar for us in the mass on unfamiliar things. Returning to the study by Alvarez & Cavanagh, the unfamiliarity of objects to be searched through a top-down search would then have to be a factor that affects the search times since the more visual information per item needs to be decoded, the longer search rates they would induce. Adding to the discussion, Miller notes that increasing the attributes of the objects that need to be distinguished is likely to help us do so[63]; In conclusion, multivariate objects might be easier to recognize and distinguish from one another dispite the suggested addition in response time following the added information load.

62. *Koponen & Hildén, 2019, page 50*

63. *Miller, 1956*

Within the context of data-driven tools that enable discovery rather than a narrative, it is worth keeping in mind to rather enable searches through user-selected features in the data rather than pre-processing the visible information based on a criteria not visible to the user, when that information might support discoveries within the scope of the data. As well as enabling the users to facilitate visual search queries, a legend to support guided visual top-down searches needs to be provided when the visualization contains information that is not intuitive to read such as textual information. Legend is necessary also for the fact that in order to enable visual top-down searches, the visual encoding methods need to be known first in order for them to guide the user's search in a directed manner. On top of providing legends, building visual hierarchies that support perceiving information at different levels supports a cognitive tendency to perceive coarse visual structures and ideas of the general shapes of things before their details relating to top-down searches[64].

64. *Itti et al., 2005, chapter 25*

The third aspect related to visual searches discussed here is called a *scene gist*, meaning the fact that familiar visual structures are easier to navigate through; This also relates directly to the goal of reducing the cognitive load by considering convention and familiar structures in data visualization practice. People have a tendency to not only navigate with ease, but also in general to rate familiar things higher in terms of their preference in relation to unfamiliar things; This phenomena is often called a *mere exposure effect* referring to the fact that mere exposure to stimuli has influence on how they are perceived.

Relating to visual searches, there is also an additional framework of though suggested in order to guide the design of a system called a *visual scent*[65]; Visual scent meaning that the user is looking for certain information, and by doing so inspects elements that, metaphorically speaking, smell like the information that they are looking for. This metaphor suggests a thought exercise for the designer; How to make sure that the information scent of each target option is not diffused while making sure that guidance towards any certain element is not emphasised over another element of the same priority. Too general of a guidance or too lengthy of a search diffuses the metaphorical visual scent.

65. *Ritter et al., 2014, pages 238-239*

*Perceiving structures*

To encourage truthful representation of data, the visual form of data visualization needs to take into account how readers induct information from visualization's structure that might not be directly implied in the visual encoding methods or textual information available. Meaning emerges from comparison; Therefore, the designer of an exploratory graphic must ask what comparisons will be done and how they should be enabled. The designer must be careful to not visually encourage or enable comparisons or groupings that do not reflect any actual relationship or meaningful comparison in the data simply because of their placement in the layout[66]. Where there are comparisons or categorisations to be made, the design should encourage local comparisons where possible meaning that the objects that are to be compared should be grouped together within the reach of an active eye, also referred to as uninterrupted visual reasoning[67]. This results in supporting the previously discussed short-term memory while making comparisons; Tufte discusses this principle of visually enforcing comparisons of changes[68] specifically in the context of small multiples, but the same principles could be applied in a wider context as well.

66. *Ritter et al., 2014, page 131*

67, 68. *Tufte, 1991, page 67*

# *Methodology*

Following chapters describe the methodological choices of planning and conducting the user study, as well as some background information to support understanding and scoping the production component and designing for expert users.

### *User-centered design practice and contextual approach*

The user study adopts a variation of a contextual inquiry framework in order to access qualitative information about the processes of the expert users in the context of their research-related task performing, and to deploy that information in the process of developing a data-driven visualization tool. More specifically, contextual design approach in this study aids in detecting usability issues of an existing system (Risteys phenotype browsing environment) and determining methods to support user's task by inspecting their behaviours and suggesting solutions based on key findings from the study. Perspectives adopted from contextual inquiry as a method allow the user study to aim at gaining a holistic understanding about the issues that the users might be having and how a novel technology, or in this case a visualization tool, could be of aid to resolve these issues. The goal is for the process to navigate itself towards a meaningful scope and targeting a direct point in users workflows rather than defining the scope first and after that studying users inside that particular scope. A wider space of inspection allows for a comprehensive view of the processes and promotes flexibility in the process, where observations and discussions can guide the user study and the following design process.



User-centered design is used to describe an iterative process where the focus stays on the users and their needs throughout the design process. Methods adopted from contextual inquiry relate to the goal of establishing a user-centered design practice in the scope of this study. To provide further clarity on the terminology of user-centered design standpoint in this thesis, a separation between user experience and usability design is to be made. "...Usability and usability engineering focus on task related aspects (getting the job done); User experience and experience design focus on and foreground the users' feelings, emotions, values, and their immediate and delayed responses"[69]. This study will mostly focus on aspects of usability research as the basis for scoping the production and understanding the questions that can be asked from the visualization, but does recognize that user experience will also have effect on the experience of usability. This study will recognise both terminology and discussions from user experience design combined with the main focus on usability to gain a more holistic understanding and criteria for evaluating the resulting production. Usability of a system can be evaluated by asking questions relating to its efficiency and effectiveness as well as how easy it is to learn to use and remember how to use it[70].

*69. Ritter et al,2014, page 70*

*70. Sharp et al., 2019, page 19*

### *The scope and background for the production*

The brief for the visualization project is to explore how users interact with phenotype-level data exploration environment Risteys by utilizing user-centered design research methodolo-

gies described above. Based on key findings of the user study, a prototype of a tool that might streamline the user's browsing behaviours in the system will be designed and developed. The production is referred to as a phenotype browser or a browsing component in relation to its function for the users.

*What is phenotypic information?*

As well as understanding the users, establishing a base for understanding the phenotypic data itself supports the visualization process. *A phenotype* can refer to any observable trait of an organism, and different phenotype classification and organisation models are used in disease definitions, for example . In this case, the phenotype information that is used as a material for the case study's production are FinnGen's clinical endpoints[71] that are currently being displayed and explored in the public online Risteys phenotype browsing environment. Clinical endpoints are indicators of health events or outcomes defined using different data from health registries. The definitions take into account the International Classification of Diseases (ICD) codes as well as codes from registries describing medicine purchases and reimbursements, cancer registries and registries containing information on operations among others. An endpoint is not therefore to be seen as something that simply stems into existence from an attribute inherent to any particular data, but instead is a work of definitions. In FinnGen, the terms endpoint and phenotype are used interchangeably.

*71. Clinical endpoint definitions are, at the time of writing this, an ongoing work by Tuomo Kiiskinen, Elisa Lahtela, Aki S. Havulinna and an extended group of experts from FIMM and THL (Finnish Institute for Health and Welfare).*



Despite this case study focusing on phenotype browsing in the context of FinnGen, it is to be noted that there are other phenotype browsing and classification systems that are potentially familiar for the same users looking to access the endpoint information on Risteys. These include but are not limited to, the ICD-10 tree[72], GWAS catalogue[73] and Open Targets Genetics[74] just to name a few. One of the primary research questions during the initial state of the user research was to find out what kind of phenotype taxonomies different user groups are familiar with, and what kind of new or existing categorisations would assist them in their searching and browsing behaviours. Understanding the user's knowledge in terms of familiar categories is essential in understanding how the system could assist them in locating their target information and establishing a common discourse when the information to be searched is a result of multiple different taxonomies.

*72. https://icd.who.int/browse10/2019/en#/*

*73. https://www.ebi.ac.uk/gwas/*

*74. https://genetics.opentargets.org/*

Searching the endpoints is a particular topic because of their definitions, and for the fact that they don't simply translate to any known hierarchies such as the ICD-10 tree categorisation. Acknowledging this, the user research focuses on finding out whether searches could still be aided by allowing alternative categorisations of the endpoints based on some of their defining information without directly attempting to assign them to classifications that wouldn't necessarily reflect the underlying structure of endpoint definitions.

## *User study*

*Designing for expert users*

Attempting to understand the users is a defining feature of a user-centered design process; Understanding their characteristics makes the design process of a system much more likely to be successful and result in a system that supports the users in completing their tasks. Among some of the most important factors to consider in order to achieve this is the awareness and avoidance of a bias called fundamental attribution error[75]; Designers assuming that users act and are like them and fail to address who the users are and how they actually use systems. Particularly in the case of expert users such as researchers working with FinnGen, it is crucial that a designer avoids assumptions and instead works towards understanding the user and the effects of existing knowledge and patterns that impose considerations on the design process.

*75. Ritter et al., 2014, page 37*

For example, the existing knowledge that expert users have might effect their search behaviours. Researchers can base their searches of target information to this knowledge without a need to browse through similar information; This would only impose an additional step in their process that does not necessarily provide value for the research opposing a situation of a person who does not know immediately what they are looking for, and instead recognises the target only after locating it from a list of items, for example. This usage of searching information based on prior knowledge is also referred to as *recalling* information, and based on arguments presented above is described to have potential for greater efficiency in the context of expert users compared to recognising the target information through browsing[76].

*76. Ritter et al., 2014, page 156*

Another implication to consider based on existing knowledge is respecting the language and textual references that are used by experts to retrieve information. This relates to knowledge specific to the field, but also to common interface elements that might appear across the systems and tools in the respective organization's software architecture; Finding common, familiar language can result in a more intuitive interface for expert users[77]. Language considerations have another direct implication for this particular case study in the context of the phenotypic data, that is often described in natural language. Language-based descriptions are rich in meaning potential, but as such are prone to spelling mistakes as well as deviations and ambiguities in both wording and spelling[78], both of which can have direct consequences in terms of success of the search queries. Thus, the defining criteria of the endpoints in the form of definition codes of each phenotype description might be of interest for the users, if the names and descriptions describes in natural language are not sufficient in forming a comprehensive and accurate view of the phenotype definition, or if they have a potential to disguise the phenotype from search results.

*77. Similar discussions also in Ritter et al.,2014, page 227*

*78. Gkoutos et al., 2018, pages 1008–1021*

*User profiles*

An initial definition of the user architypes was made with the help of an expert researcher from the FinnGen community.

These prominent user groups in wider definition were defined as clinician, data analyst and a research geneticist; The clinicians group was to include pharmaceutical clinicians as well (*see figure 2*).

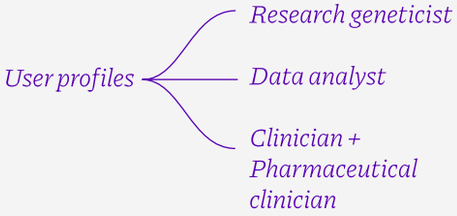



*Figure 2. Main user profiles in the beginning of the user study.*

These groups might have different ways of accessing phenotypic data as well as varying background expertise on different phenotypic browsing systems outside of FinnGen. Along with their background knowledge and habits they might be asking different questions from the data in order to advance their research.

User research and gathering information to support scoping of the project was started by creating an online document where insights regarding accessing phenotypic information through different user profiles could be gathered from the community. During this initial state of the research it was noted by the users that especially the roles of the expert users with a research background overlap with what could be identified as a role of a data analyst, so this division was merged while the process was ongoing. The main user profiles were then re-identified as clinician and a research geneticist, and understanding how these profiles might differ in their behaviours concerning the search patterns and usage of phenotypic information then became one




*Figure 3. Screenshot of Risteys landing page.*

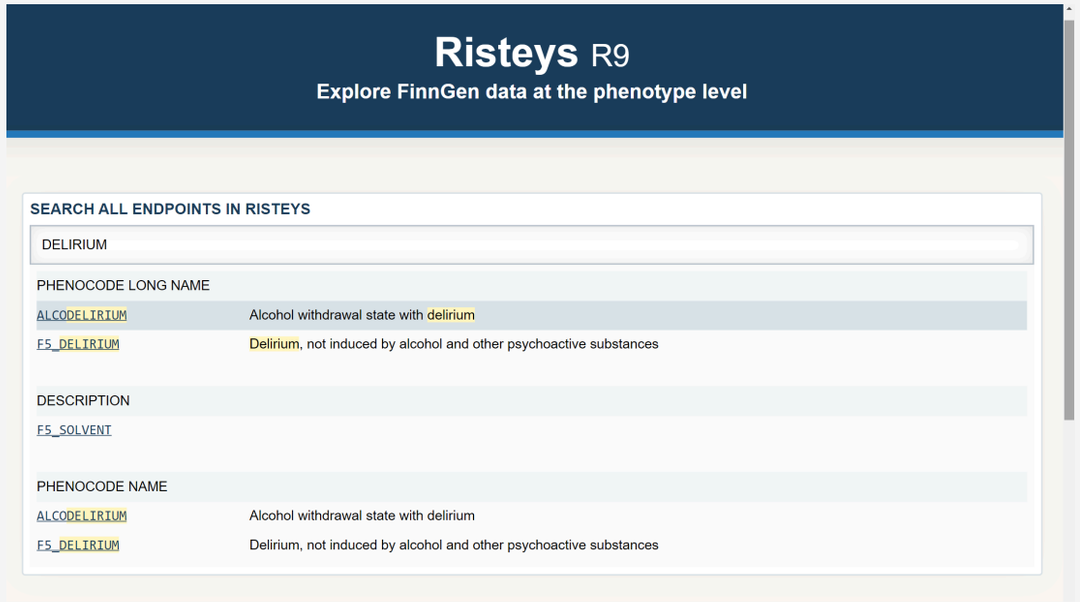



*Figure 4. Screenshot demonstrating the display of search results in Risteys.*

of the target questions to be asked from the data gathered in user interviews.

*Mapping accessing phenotypic information in FinnGen*

Forming a comprehensive understanding of the search and use of phenotypic information was supported by creating flow charts of related user processes based on the data gathered from the cloud document as well as the FinnGen Analyst Handbook[79] *(see figure 4)*. The first flow diagram *(figure 5)* demonstrates the fact that users who are looking to access specific phenotypic information might want to access either individual-level data that requires access to FinnGen's sandbox environment or non-individual level data that is accessible without restrictions to the public. Users might acquire phenotypic information from multiple sources, but to inspect phenotypic non-individual level information defined by FinnGen the users need to access Risteys either directly with a previously acquired URL or through the FinnGen homepage or a search engine query.

*79. Introduction - FinnGen Analyst Handbook. (n.d.). FinnGen Analyst Handbook. Retrieved April 12, 2022, from https://finngen.gitbook.io/finngen-analyst-handbook/*

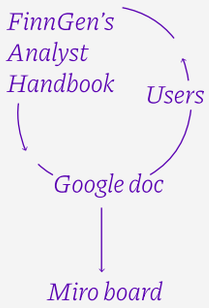



*Figure 4. Gathering information about user's phenotype data searching workflows.*

After locating the Risteys environment, users are looking to find the phenotype information that they are after. Risteys allows users to locate information by clicking through lists of different pre-categorised endpoints as well as to access the information directly through a search bar (*see figure 3*). The search bar returns results by matching the input string to either the endpoints long name, their description or their name *(see figure 4)*. Both the search bar and the categories aim at leading the user to specific endpoint's landing page that contains information such as the definition for that specific endpoint, summary of

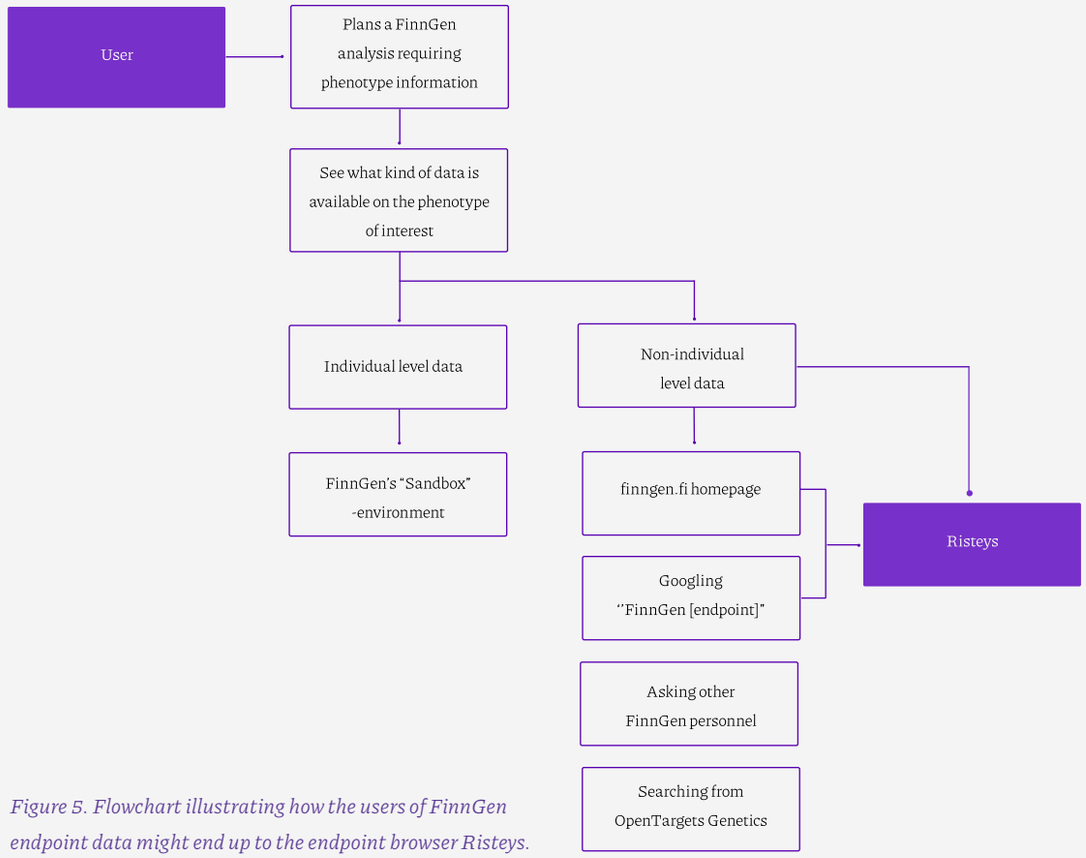



*Figure 5. Flowchart illustrating how the users of FinnGen endpoint data might end up to the endpoint browser Risteys.*

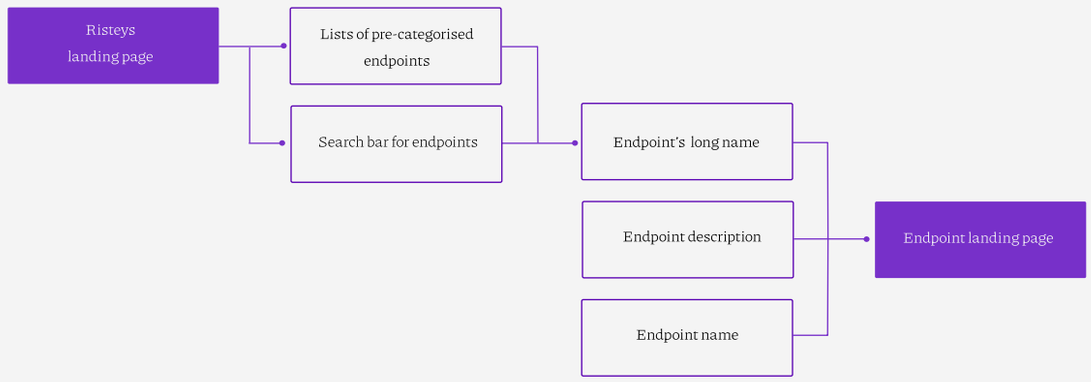



*Figure 6. Flowchart illustrating the Risteys search model; Risteys refers to the endpoints as phenocodes in their search results, but they are reffered to as endpoints in this figure for the sake of clarity and continuity.*

the statistics and information on related endpoints. Another flow diagram *(figure 6)* was constructed based on the simplified user search models inside Risteys to get to the target information. The scope of search behaviours with a motivation to locate a specific endpoint became the point of interest for this case study.

*Background of the user interviews*

After idenitifying the main user profiles and acquiring initial understanding about their phenotype -related workflows in order to scope the production space, the user study was continued by user sampling, profiling and interviews (*see figure 7*). The goal for arranging the interviews was to find out about task behaviours and general usability of Risteys from the perspectives of the main user profiles. This was to be achieved by looking at how the users currently complete tasks inside the system, and in the process develop a more comprehensive understanding of the system as well as possibilities for its improvement.

*Poll to map interest in taking part in the user testing*
|
*Users who answered 'yes' to intial poll*
|
*Profiling questionnaire*
|
*Scheduling*
|
*Zoom invitations*
|
*Interviews*

*Figure 7. Generalized workflow of the user study before the interviews.*

To reach users from the FinnGen community, an expert researcher within FinnGen helped with identifying available users from the target user groups. Preliminary polls were sent out in order to find out which users would be interested in taking part in the user testing.  The poll resulted in a pool of 18 potential users to be interviewed. Then, in preparation of the sampling, a background information questionnaire (*see annex A*) was sent to the users in order to gain knowledge about their backgrounds to support the most balanced selection of users from different user groups. After that, a Doodle poll was created and sent to the users who answered the questionnaire in order to find the most suitable times for conducting the interviews. This process resulted in a poll of six users, that were then interviewed. The users that we successfully interviewed were a group of experts working for or in collaboration with FinnGen, and consisted of three women and three men, between which two of them identified their main position in FinnGen to be a research geneticist and four came from clinical backgrounds with varying research areas.

The sampling method was thus a combination of convenience sampling and theoretical sampling; The user group was mostly defined by which users were able to be reached, while at the same time making sure that there would be sufficient representation between different user perspectives.

*Data gathering methodology*

Since the goal of the user study is to understand the flow of each participant's process and their strategies in retrieving information from the phenotype browsing system, the data gathering methodology for the user testing was decided as a combination of a think-aloud method for task demonstration and a semi-structured interview with a standardised set of questions for the discussion. In a think-aloud method, the user is asked to talk aloud while executing a given task or solving a problem, and the request is then repeated if necessary during the process to encourage the user to tell the observer what they are doing. The purpose of the think-aloud method is to produce relatively objec-

tive qualitative data about the user's actions and processes that would not be as affected by different expectations and mental models than the ways people would speak about their actions in retrospective. Complementing the data gathered from the think-aloud, inspecting user behaviors occurring simultaneously will assist in forming a sharper image of the user's realistic task behaviors. As a task, the users were asked to *find any phenotype from the browsing system that is near or inside their main area of research*. The preface is that the task should require some problem-solving, so that the answer to the task would not be immediately recoverable from memory leading it to being automatic. Balancing the difficulty of the task is important for retrieving the most accurate data about the user flows through the programme.

Other considerations that might affect the data to be gathered relate to the instructions given for the user for the think-aloud session itself; The think-aloud should only concern task-related navigation in the system to be able to understand those procedures by giving neutral instructions rather than requesting specific information. This is also aptly noted by McDonald, McGarry and Willis: "Procedures that ensure that verbalizations are constrained to task based cognitive processes (working memory) will produce valid data. By contrast procedures that require users to verbalize information that goes beyond working memory, such as requests for explanations via explicit instructions or evaluator probes, will produce invalid data because participants must engage in additional processing linking the content of working memory to information in long-term memory"[80]. It has also been found that explicit instructions do not result in better data gathered from a think-aloud session, but might only result in burdening the user[81]. This notion aligns with the one made by Ericsson and Simon[82] explaining that it is best to create a space for the experiment where the user can focus on verbalising and executing the task itself rather than explaining about the task to the examiner at the same time as this might create social influence to both the thought and verbalising patterns of the user being examined.

*80. McDonald et al., 2013*

*81. Zhao et al., 2012*

*82. Ericsson & Simon, 1998*

After initiating an appropriate methodology for data gathering, the standard set of questions was defined for the interviews. They were divided in two parts based on whether they concerned the searching behaviour of the user or how they handle the in-formation *after* locating the endpoint of their interest from the system. The questions were asked in a semi-structured way within the interview flow; The contextual methodology allowed some deviations from the core questions in order to form a more comprehensive overview of the user's experience and perspectives.

***The questions concerning the search behaviours were:***

*Was or is there anything confusing about the search results resulting from this task?*

*Is there anything that you hope would appear on the Risteys start page that might aid your search?*

*Are you more often using the search bar or the lists of endpoints on Risteys homepage?*

*Are you more often searching for a particular phenotype or an endpoint*

*or rather browsing a wider group of phenotypes or endpoints?*

*How often do you search or browse for phenotypes or endpoints outside of your main area of interest?*

*When using the search bar, Risteys displays results based on phenocode, phenotype name or their description. Which ones of these do you most often find useful?*

*Do you find yourself wanting to find specific phenotype or endpoint information using other criteria than the ones Risteys already displays?*

*Do you find that FinnGen endpoints might not always encompass the exact phenotype information that you are looking for?*

***The questions concerning found information were:***

*After you locate an endpoint that you are interested in, what kind of information are you looking to get out of it?*

*Do you need to save some of that information for further research purposes, such as copy-past text or images?*

*Do you often compare different phenotypes or endpoints?*

*Do you often compare different groups of phenotypes or endpoints?*

*Have you been unable to find a phenotype or an endpoint that you have been looking for?*

*How do you use the information that you get out of Risteys?*



*Key findings from the user interviews*

User interviews were recorded, and notes were taken both during the interviews and during watching the recordings after the interviews to interpret the data. Investigator triangulation[83] was performed by having two observers, another the writer of this thesis and another an expert researcher from FinnGen interpreting the raw data gathered. These notes were then cross-examined in order to retrieve patterns of information that occurred during the interviews. This coding resulted in the following key findings regarding user's experience and the usability challenges of phenotype browsing in the Risteys environment.

*83. Sharp et al., 2019, page 264*

Users found the text-based search by the phenotype's name more challenging compared to search by ICD-10 classification codes used in the endpoint's definitions. Challenges arise both from the spelling of the names being more prone to errors compared to the number-letter codes of the ICD-10, as well as the fact that FinnGen endpoint definitions might end up using a different name for a phenotype compared to the known version that the users might be familiar with and search with. It was also noted that in the case of user not being able to locate a specific phenotype that they were looking for, Risteys did not indicate to the user whether that phenotype simply does not exist or if the search term was not the one to return the wanted result.

In short, known conventions were trusted over the new definitions, and as such ICD-10 codes were trusted over the pheno-

type names in Risteys. Users often relied on their memory and existing knowledge in order to find phenotypic information. In other words, users are "browsing through what they know"[84] based on their research interests. Some users were also looking to compare the phenotype definitions made within FinnGen with other phenotype browsing systems such as the UK or Estonian biobanks, and then ICD-10 codes were used to enable this comparison and locating information from multiple sources.

*84. Quote by an expert researcher from FinnGen*

Most users were not looking to compare similar endpoint's information other than with the intention of finding the most interesting endpoint for their research purposes. Information supporting this kind of comparison is, at the time of the user study, almost completely missing from Risteys. Many users encountered that a single search query returns many instances of similar results, and the users have trouble distinguishing between them in the absence of any further information about each result. Users also found that the occuring search result categories of phenocode, phenotype's "longname" and description sometimes confusing. Some criteria that would help the users were mentioned, among those the case numbers assigned to each endpoint on definition. Therefore, the main challenge with the phenotype browsing in Risteys is not so much about not finding the endpoints of interest, but finding a few too many and having to compare between them.

In other words, Risteys search queries result in a list of endpoints but does not allow any mid-point comparisons between them leading the users to a narrow exploration space. This results in them feeling unsure about whether their selection of focus was indeed the correct one and having then to go back and forth in the search results (*see figure 8*). On top of this, many users felt like Risteys interface might have acted as a curtain between them and all the phenotype information that there is to offer in Risteys; One user even demonstrated using a spreadsheet of all the phenotype information in FinnGen to locate an endpoint and only after that copy-pasting the information into Risteys to access further phenotypic information that is not available in the spreadsheet. Sometimes users also mentioned that relating their focus endpoints to a wider framework was challenging, but some of them were also stating the current means of browsing to be sufficient despite maybe not ideal.

The users recognise that phenotype taxonomies have their own sets of difficulties and unstabilities not related to phenotype browsing systems such as Risteys. These understandings were discussed in the context of ICD-10 codes and how they were included in the endpoint definitions; The users recognise that the definition process is a set of compromises, and they have a motivation to find out more about which codes have been included in the definition to understand them better.

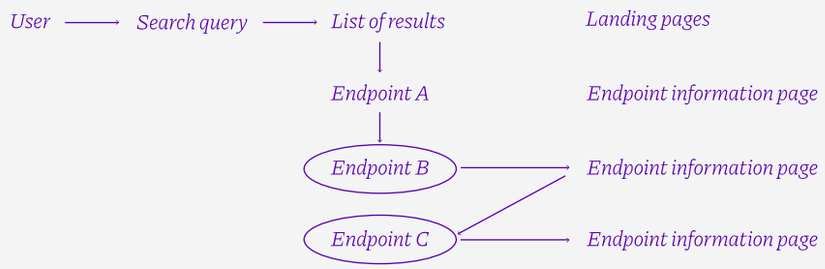



*Figure 8. Diagram illustrating users toggling between endpoints and their landing pages to determine their focus.*

*Practice*

## *Landscape of tools and technologies in scientific data visualization from a visual communication design point of view*

Because of its cross-disciplinary nature, the field of data visualization is defined through the tools and the expertise available within several fields of study. Some accessible solutions exist for producing static visualizations and data sketches such as RAWGraphs[85] and Flourish[86] just to name a couple, but special care and collaboration is still required especially considering the production of interactive data visualization tools that would answer the requirements of use in scientific research.

The design, development and deployment included in the production process add specific challenges to the workflow from a designer's point of view. These include gaining a level of understanding regarding the visualization topics and the data that is used to produce the visualization; Especially considering the field of scientific visualizations, this task should not be underestimated. On top of comprehending the wider context, customized visualization solutions might not be easily produced; In particular, programming languages and workflows within scientific visualization might present a significant deviation from a designer's standard practice. The specifics mentioned above could be some of the contributing attributes in why many of these visualizations are made by scientists for scientists without the involvement of a user experience designer or visual designer especially considering highly specialized fields such as research in genetics.

Producing scientific data visualizations requires methods that can answer these versatile challenges. Executing and combining flexibility, user interactions, system compatibility and clear presentation of data is a constant effort that especially designers working with online data visualization are battling. Despite the workflow having the potential to stem into multiple directions and steps, producing visualizations by programmatic methods can act as an answer to set the design process relatively free of the constraints of software that would inevitably have their own set of rules and act as authors in the process. Tufte has also recognised this aspect of practicalities in the life of an information designer by asking "Why should the intellectual architecture of our reports and our evidence reflect the chaps of software bureaucracies producing those reports?"[87], and in asking so steps into the realm of discussing authorship and accessibity of the data design processes. Of course, this study recognises that programmatic methods don't simply exist without biases, but are made and used by humans, and therefore inevitably present their own set of rules and unstabilities regarding their use and the attributes they impose to the work they aid producing. The level of flexibility and authorship that they provide from the perspective of the design process is therefore relative, but they are still a preferable solution for working in an agile framework of custom interactive visualizations.

In order to establish the landscape of tools and technologies specific for this case study, they will be briefly described here. This case study includes working with multiple data sources

*85. DensityDesign Research Lab, Politecnico di Milano. (n.d.). Home | RAWGraphs. RAWGraphs. Retrieved April 7, 2022, from https://rawgraphs.io/*

*86. Flourish | Data Visualization & Storytelling. (n.d.). Flourish. Retrieved April 7, 2022, from https://flourish.studio/*

*87. Tufte, 2006, page 61*

and file types including spreadsheets as well as comma- and tab-separated files that require the initial handling of the data with tools such as Excel and a text editor Visual Studio Code[88]. Processing the data further and being able to merge, nest and format it in other ways requires processing using Python together with a Pandas[89] library. Visualising the data is then done in D3.js[90] together with vanilla JavaScript, CSS, HTML and React.js[91] library.

Selection of the tools and technologies was done considering the level of familiarity and accessibility without compromising any crucial aspects of the process such as data wrangling that could be done with Python with a reasonable learning curve despite the lack of existing knowledge. By gaining authorship of the data file design the process was able to reach a better coverage in authorship and flexibility. In terms of the visual aspects, D3.js was used as a visualization library for its data-driven approach that natively takes advantage of scalable vector graphics (SVG) and Document Object Model (DOM) manipulation, both excellent features and crucial aspects of web-based data-driven visualizations. Deploying D3.js as the main visualization tool poses some implications to the visual framework of the production as the library has its own predefined set of visual strategies, many of which can, however, be modified to provide further flexibility.

88. *Microsoft. (n.d.). Visual Studio Code - Code Editing. Redefined. Visual Studio Code. Retrieved April 6, 2022, from https://code.visualstudio.com/*

89. *pandas - Python Data Analysis Library. (n.d.). Pandas. Retrieved April 6, 2022, from https://pandas.pydata.org/. Available from https://zenodo.org/record/3715232#.Yk1QKp-U-QR and https://github.com/pandas-dev/pandas/tree/v1.0.3*

90. *Bostock, M. (n.d.). D3.js - Data-Driven Documents. D3.Js. Retrieved April 6, 2022, from https://d3js.org/. Available also from https://github.com/d3/d3*

91. *Meta Platforms. (n.d.). React – A JavaScript library for building user interfaces. React JS. Retrieved April 6, 2022, from https://reactjs.org/. Available also from https://github.com/facebook/react/*



## *Architecture of the production based on user interviews*

*Positioning the production component in relation to Risteys*

Among the first considerations when structuring the productions visual architecture was scoping the visualization both in relation to the user flows and the existing system Risteys. This required gaining understanding of both and then determining the aspects where changes in the system would promote a more streamlined user flow. As declared in the user study, the scope of this production was to address user's needs to compare certain information about the endpoints and relate them to wider contexts before transitioning into any certain endpoint's landing page.

User's familiarity with the existing system is a factor to consider especially within the interface aspects of the project, since "the way people use a system will be greatly influenced by how well they can retrieve commands and locations of objects from memory. Similarly, their feelings of success with a system will be influenced by their biases in retrieving information about past successes and failures with the system"[92]. These remarks align with the previously introduced effects of familiarity. Considering this, it is important to consider the layout of Risteys especially insofar as it was considered supportive of the user's research efforts and orient towards crafting a solution to fit that system. This perspective would call for respecting the locations of elements that users find pleasant to use and that are integral

92. *Ritter et al., 2014, page 148*

to user flows; In this particular case, the new visualization component can replace the pre-determined category listings from Risteys frontpage without disturbing the user flows, since no users identified as taking advantage of them for their endpoint browsing purposes. As well as replacing a suboptimal section of the interface, the visualization component has the potential to compliment the existing search bar as a method to facilitate search queries, providing additional context and a possibility to explore endpoints relating to the search results. Users are actively using this search bar to conduct direct searches in order to locate endpoint information, so respecting that will be a consideration when deploying the visualization component to Risteys outside the scope of this study.



*Meaningful scope of comparison*

In light of the key findings from the user study, it was clear that most users are looking to locate an area of interest and then locate phenotype(s) of their interests or directly locate a certain phenotype through a direct search. This implies that they are more rarely exploring phenotypic categories that are different or unrelated to one another meaning that after finding relevant context the core task does not require the user to go back and forth between these larger contexts. This renders the need to introduce a global visual context including all the endpoints and categories negotiable, but instead suggests an exploration of techniques to arrange information so that the visual displays would provide context within a meaningful scope for each user; This frameworks appears more aligned with the search, show context, expand on demand -framework[93] compared to Shneiderman's visual information seeking mantra of overview first, zoom and filter, details on demand[94]. Lima introduces the concepts of macro-analysis (global pattern recognition), relationship analysis (connections within a scope) and a micro analysis (single entities) in regards to network graphics[95]; Within this terminology, the findings from the user study promote most emphasis to be built upon the relationship analysis stage, where users wish to explore connectivity and compare datapoints within a scope rather than discover inter-category patterns.

*93. Van Ham & Adam, 2009*

*94. Shneiderman, 2003*

*95. Lima, 2011, page 91*

The structure of the endpoint data suggests common attributes that can be used as methods of arrangement in order to create categories forming these meaningful scopes. These attributes can then enable visualizing the interconnectedness and taxonomy of the data; Thus, the visual architecture and data sketching of the production is started with a tree structure than can have the potential to stem into a direction of a network-like structure later on.

In the seemingly unlikely case where users would want to compare different categories to each other, choices in layout and organisation of the spatial space become of even higher importance. Visual proximity of two elements in different categories could suggest a comparison without a conscious motivation to do so if arranged in a non-deterministic manner[96]; therefore, it should be paid close attention to what kind of proximities are meaningful and how the organisation of comparable categories might encourage those meaningful comparisons without lead-

*96. Also discussed by Munzner, 2014.*

ing the user to believe that there are connections or semantic proximities where there are none.

The benefits of introducing meaningful scope of comparison rather than a global view might also extend to considering computational costs and the effects of latency for the user experience; It has been found that delays of 500ms decrease user activity and rates of observation[97]. Rendering a global view of a big enough dataset can cause said latency.

97. *Liu & Heer, 2014*

*Search and information architecture to guide the visualization*

Based on the findings of the user study, the user's search behaviours are generally best described with either a *lookup* behaviour, where a location and the name of the target endpoint is already known, or a *locate* behaviour, where the user still knows what they are looking for but they do not have adequate information to locate that target immediately by a direct search[98]. Considering this, the navigation solutions should aid the location of the target endpoint rather than aim at maximizing the explorative aspect of browsing through the visualization. To get to a relevant starting scope, like previously mentioned, the currently implemented search bar in Risteys would likely continue to be a successful solution for the lookup behaviours. When the users need to locate an endpoint of interest through other than a direct lookup search or when they are interested in comparing between endpoints, a novel visualization would aid current processes. Looking up a relevant category to get to a meaningful scope could be also executed through a search bar, since expert users will have the knowledge of their categories of interest, at least in the case where categorizing is done through common conventions, and will likely prefer to get directly to that area without additional browsing.

98. *Munzner, 2014*

After locating an area of interest, the goal for the visualization is to clearly present all of the endpoints in that category, resulting in a tree structure. The focus tree(s) resulting from an initial search themselves already reveal categorical information relevant to the user acting as an overview of the context rather than a global view of all the data available, but visual encoding and textual references are still needed to help the user identify their focus between endpoints in the same category. Relevance of the datapoints for the user is determined by their task; in this case, being able to locate the endpoint of interest from a category of already related endpoints. Criteria that directly supports this goal are to be considered as primary information and implemented to the direct view of the visualization and not to be hidden under an interaction layer. Primary supporting information in the case of the endpoint browsing are the endpoint name, case numbers, gender balance of the cases, meta-analysis information and information on whether a certain endpoint can be considered a core endpoint or not.

### *Initial choices in visual language*

Visual encoding methods and language of the visualization are mainly determined based on aspects of known best practices

established in the literature review as well as the goal of establishing a coherent and aesthetically pleasing visual hierarchy and encoding methods. Chosen visualization methods respect the stakeholders associated with the production, these being defined as FinnGen and their existing phenotype browsing system Risteys, where the visualization tool would be implemented. At the time of conducting this study, FinnGen's visual brand has not been found to extend to or include any restrictions in the visual language of their data visualization tools, but their established visual language will be taken into account nevertheless.

Knowing the framework and scope for the visualization project, some grounds can be laid to guide the work in visual design. Following the notion of consistency is the first and foremost principle to follow since it is needed for successful reading of the visualization. Thus, same visual markers and methods are used to encode information of the same category. Regarding visual markers, ample separability of the forms of different markers used is to be ensured for the sake of legibility. Encoding quantitative information through the markers by expanding their area is avoided for it does not provide the most accurate readings and has the potential to cause occlusion in the display as well as confusion in terms of visual hierarchies. Instead, codings through colour and spatial positioning could be applied.

Visual continuity is to be ensured extending from the readings inside of the visualization into the wider context of the system, meaning that existing visual encodings will be respected to ensure the continuity of reading and meaningful comparisons.



Risteys currently deploys some colour strategies in summary statistics charts (*see figure 9 and 10*), so similar direction is to be taken with a new browsing component at least to the extent of ensuring that no complete or major deviations from the color scheme are made. Beyond the summary statistic charts, Risteys does not deploy visual encoding strategies to encode the information that will be available in the new browsing component, so their encoding choices beyond colour are not guided by any existing solutions.

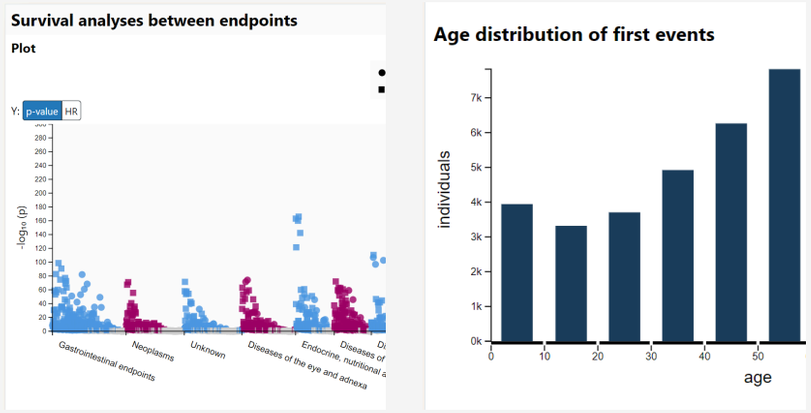



*Figure 9 and figure 10. Screenshots demostrating examples of the charts used in Risteys.*

Existing colour scheme in Risteys uses cool tones of #1c3d5a, #70ace6 and #b13283. FinnGen's visual identity is found to use at least two different hues from the blue-violet scale, these being #3500d3 and #4b1dd9.

Considering the knowledge of existing colour strategies, the following guidelines and decisions were made for the visualization. A violet hue #5900b3 would be set up as a main hue to

highlight visual markers or text. Knowing that the visualization will code binary gender balance[99] data, a complimentary hue of magenta #FF33F2 is set to compliment the violet so that the contrast between them will be sufficient. As well as establishing visual ties to stakeholders colour palettes, magenta and dark violet have the potential to encourage semantic reading of the gender balance data. On top of combining at least two hues for a categorical scale, the visualization needs to code quantitative information in the form of growing GWS hit numbers; These could be addressed through a quantitative luminance scale. The heatmap resulting from utilizing the primary violet hue in the luminance scale was checked to be safe for use considering color vision anomalities with a chroma.js[100] palette tool.

*99. The data expresses the gender balance as a balance between male and female case numbers, and does not recognize data beyond this binary partition.*

*100. Aisch, G. (n.d.). chroma.js palette helper. Chroma.Js. Retrieved April 8, 2022, from https://gka.github.io/palettes*

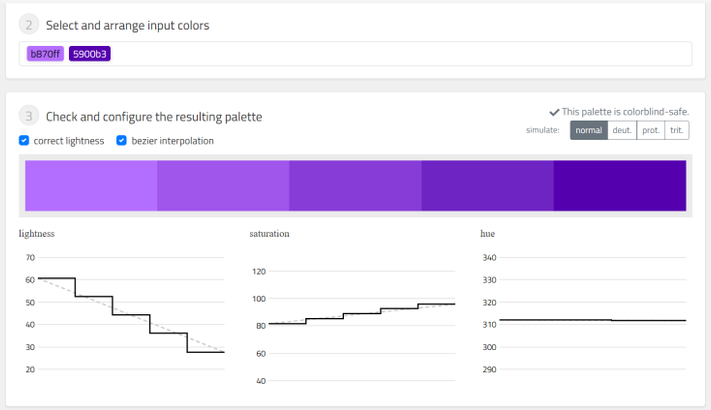



*Figure 11. Screenshot demonstrating the colour palette check via chroma.js.*

The visual weight of the glyphs and the text is also to be considered through colour; Like stated in the literature review, colour is an inseparable encoder from shape and size. The bigger the area gets, the visual weight and boldness of the colour should be considered resulting in dimming the saturation or opacity of the hue, for example. On the other hand, smaller marker's visibility needs to be ensured through using a bolder hue with better visual salience. Attention is also to be paid for the colour contrast between the background and the typography or other glyphs, as it should result in a contrast that aligns with the minimum contrast success criterion for the web[101].

*101. W3C, World Wide Web Consortium. (n.d.). Understanding Success Criterion 1.4.3: Contrast (Minimum). W3.Org. Retrieved April 6, 2022, from https://www.w3.org/WAI/WCAG22/Understanding/contrast-minimum.html*

Encoding methods themselves will be determined by the attribute of the phenotype data that the particular encoding will visualise; Deciding the visual encoding in this production takes into consideration whether that particular data is of primary use or of secondary use for the researchers in order to facilitate comparisons, and that will affect the establishment of visual hierarchies and interactions. Spatial position is to be used to encode endpoint information relating to their categories and internal arrangements, and after that encoders such as shapes or luminance can be deployed to further enrich the coding of the endpoint data by visual means. Visual markers are only to be used to depict data or aspects of the data, and therefore non-data related visual attributes should not be highlighted by encoding strategies.

Typography is to be visually light to not distract from the overall reading of the visualization; Considering that the endpoint data has the potential to be textually heavy, this principle becomes even more urgent with this production. Typeface is set in a sans-serif typeface to establish visual ties to Risteys that also deploys a sans-serif typeface. Especially considering the choice of a sans-serif typeface and the fact that the data contains combinations of upper- and lowercase text as well as different numbers

and number-letter code combinations, the typeface of choice should be able to be legible even with visually challenging glyph combinations (*see figure 12*). Considering these criteria, the type is to be set in Source Sans Pro, which is a Google font designed especially for user interfaces.

*Figure 12. Examples of glyph combinations that can be easily transposed set in Source Sans Pro.*

Il1i  Oo0  m rn  cl d  vv w

Minimizing the cognitive load regarding the reading of the visualization can be worked towards by paying consideration to whether data attributes can be encoded in an intuitive way as to avoid cross-reading between an external legend and the visua-lization. Dual coding of signs is to be implemented by including both the textual information and its visual reference wherever the textual information present primary information for the user that cannot be coded by only visual means achieving preciseness of information.

Overall, the goal is to lay down a neutral zone of affect for the users meaning that the goal is not to unnecessarily and unknowingly emphasise any data in a way that might result in a distracted reading of said data. This implies that the nodes and visual markers related to them should remain of equal or balanced visual weight in the initial state of using the tool, where interactions or user selections do not yet guide the visuality of the display. Such balanced representation would support the explorative interaction framework of the visualization. Information architecture in terms of primary and secondary information is to be respected, and not rearranged in favor of the visual form. The guiding principle for the visual design is that the visualization should act as a framework against which discoveries can be made, but not to direct or force the user for any particular directions, unless the direction can be defined useful based on the findings of the user study.

### *Iterative data design and visualization process*

The practicalities of the design process within the scope of this production included both data design and wrangling as well as data sketching and programming through an iterative process. In this case study, handling the data formed a substantial part of project's workload, and therefore is worth discussing briefly.

First state of the data handling moving towards being able to sketch with the data was to acquire an initial set of data from public sources that would be quickly available. First set of data was a spreadsheet downloaded from FinnGen's public results directory[102]; This particular spreadsheet that was downloaded describes all the endpoints available in FinnGen and some of their defining information such as registry codes used in their definitions. After acquiring the data it was re-formulated into a JavaScript object notation (JSON) file format using Python and a related Pandas library. The reason for choosing the JSON format was influenced by the fact that when working with D3.js to visualise the dataset, the JSON format provides convenient key-value pairs with which to access different attributes of data.

*102. Clinical endpoints. FinnGen Clinical Endpoints. https://www.finngen.fi/en/researchers/clinical-endpoints*

Looking for a way to establish initial hierarchical structure to sketch with, the data was grouped according to a column that describe the FinnGen disease categories where each endpoint in the spreadsheet belongs to. These categories are described as particular tags in the data file, and the same column could contain one or more. These tags were transformed into a python list that is equivalent to an array format in JavaScript. Then, all of the data was grouped according to the first tag in each list. This resulted in a nested JSON file, that could then be sketched with: Data sketching was done by initiating a React.js -project and then rendering sketches made with D3.js as components within the application.

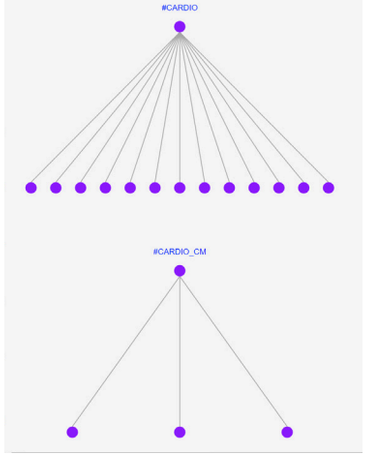



*Figure 13. First examples of trees describing the amount of related endpoints in each tag category.The tags can be seen on top of the parent node.*

The tags described by the spreadsheet, however, would not respond well to what could be described as a familiar convention or a useful way of looking up information for the users since their language format does not directly respond to any familiar conventions outside of FinnGen's definitions. Reflecting on the results of the user interviews, most users preferred to search with familiar naming conventions instead of any FinnGen -specific categorisations; ICD-10 hierarchy was mentioned in particular, but issues lie in the fact that FinnGen's clinical endpoints do not align directly within the ICD10 -tree since endpoints are essentially compound information with a likeliness of including several ICD -codes and other registry information.

*ICD-10 tree ⟶ Chapters ⟶ Blocks ⟶ Phenotypes*

*Figure 14. Terminology that describes the hierachy of the ICD-10 tree.*

Reflecting on this, the data analysts inside FinnGen community were asked to produce a data file with alternative categorical information that would indicate where in the ICD-10 blocks the endpoints settle in. ICD-10 blocks indicate a neighbourhood of separate categories of ICD-10 codes (*see figure 14*), so this separation method would likely produce a better result for enabling comparisons of related endpoints that would include similar ICD-10 codes in their definitions. As established previously, this will support the users better in relation to the category information available in the public spreadsheet since ICD-10 categorisations are a part of a well-established taxonomy of phenotypic information and adopted by many users.

This newly acquired data was then grouped in Python according to the now available ICD-10 block information, and available block descriptions were also added to both the data file and the data sketch; The issue of not all FinnGen endpoints responding to any ICD-10 blocks was not addressed at this

point of the process. The data wrangling actions described above resulted in a data file where the ICD-10 blocks formed the parent elements in the data, and the individual endpoints formed the child elements. The structure was then visualised (*see figure 15*). The layout of the tree was adjusted together with the contents; Horizontal trees provide better visual support for datasets with long names, and support a native linear reading direction. Occlusion resulting from big variations in child node array sizes was also tackled in the sketch layout by calculating the amount of child nodes and dynamically assigning pixel space after each one. On top of providing visual breathing space, the approach of having all the endpoints visible in the categories might respond to the user's uncertainties of being able to find all the endpoint information from the current search architecture of Risteys.

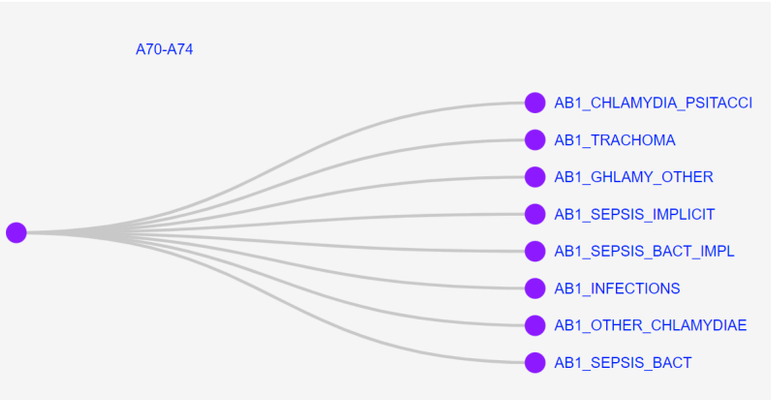



*Figure 15. Screenshot demostrating the new, horizontal approach with data grouped based on ICD-10 blocks. Here, the tree describes endpoints falling within the A70-A74 block.*

Next, some simple user interface was added in the form of a search bar. The search bar enables the testing of browsing behaviours by taking the input string and matching that to

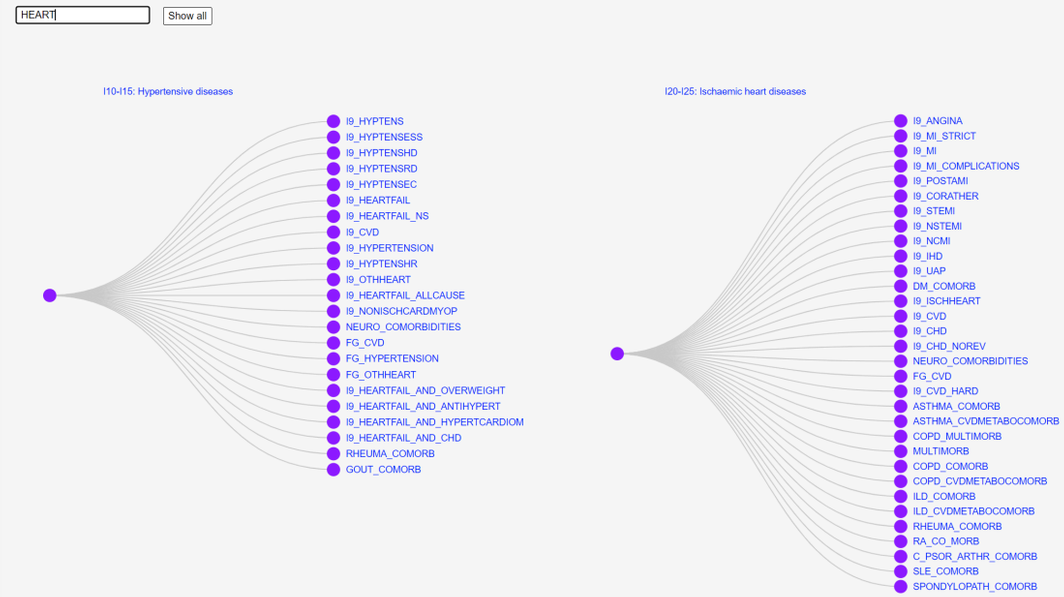



*Figure 16. Screenshot demonstrating the working of the search bar; the search function matches the string 'HEART' by returning trees that contain a matching string of text in any of the nodes or their block titles.*

the text content of the trees in either the block name of any of the nodes. The search then returns all of the trees that contain information matching the search string (*see figure 16*).

As a next step in the process, the data file was enriched with additional information from a JSON file containing data on case numbers of each endpoint as well as a breakdown of female to male distribution of these cases assigned to each endpoint. The original data file was enriched by matching the name of the endpoint in both files and then adding the case number information as separate key-value pairs to the original, nested JSON file's child entries. In the visualization, the overall case numbers of each endpoint were appended inside brackets after each endpoint's name as a number, since their amount is significant for the users; Contrasting this, the gender balance data is more of an indication of the balance between the ratio of female to male case numbers, so a horizontal bar chart devided according to each endpoint's binary gender distribution was added next to the node to code that information (*see figures 17 and 18*). Magenta hue is coding the amount of cases classified as female, and contrasting that the cases classified as males are encoded with a violet hue. The length of the bar and it's aspect ratio was decided as to make it compact, but still allowing ample visual space for perceiving the coloured in areas and thus grasping the ratio between the gender balace data. Compactness of the bars allows them to appear in line with the visual hierarchy of the visualization as a whole, and not to grasp more attention than other attributes of the visual encoding despite it having to take up more space.



Considering that the overall case numbers are among the most defining information for users in order to determine their focus endpoints, a custom default ordering of the nodes could be considered based on the case number data. The purpose of reordering is to present the endpoints with most case numbers assigned to them on top of the tree, descending down to the endpoints with the least number of cases assigned. Some of the node lists might get long due to large variations in category sizes, so the concerns of search time expansion and cognitive load while going over large lists and taking into account the primacy effect in reading lists could be partially addressed with this type of deterministic arrangement of nodes in the tree.

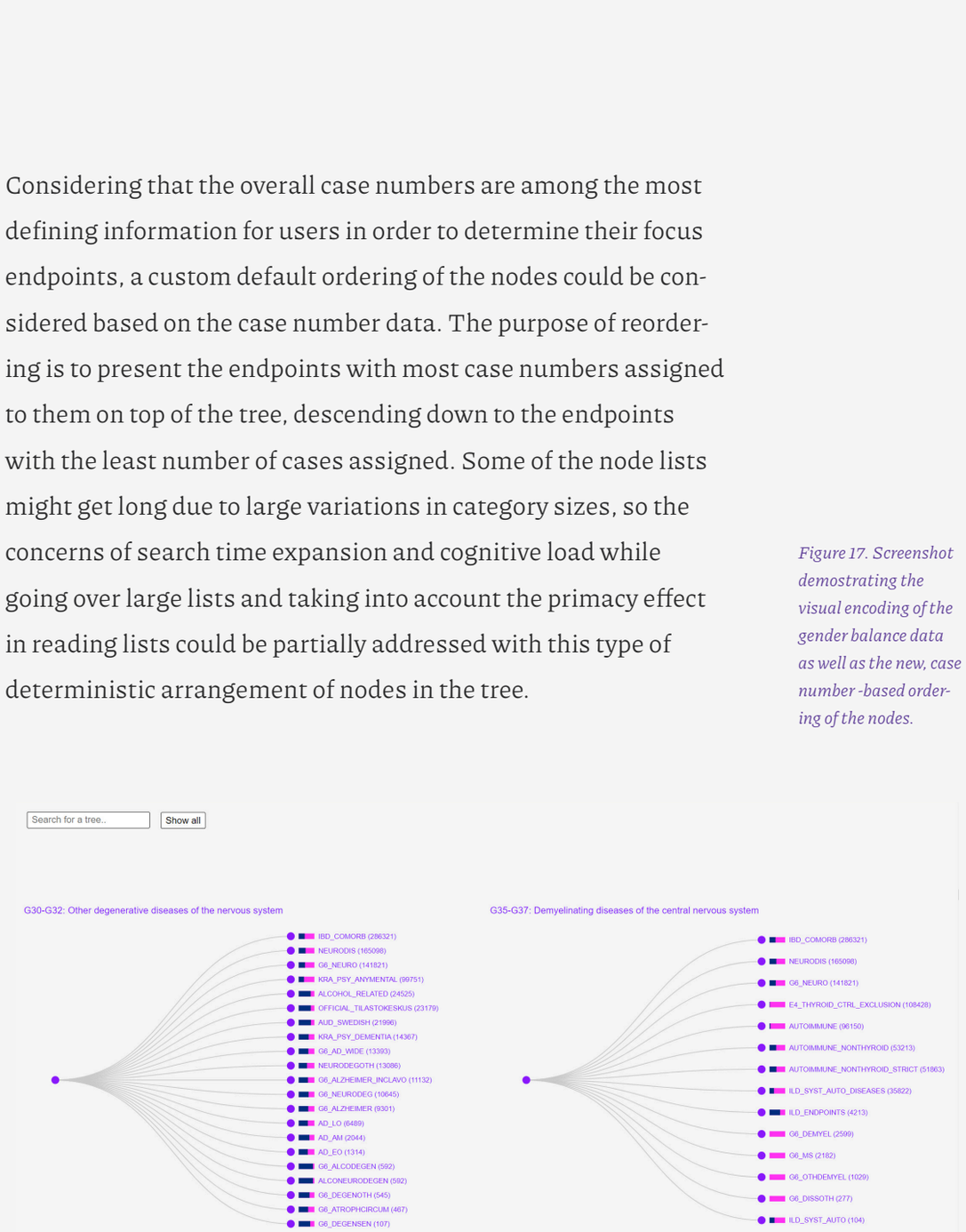



*Figure 17. Screenshot demostrating the visual encoding of the gender balance data as well as the new, case number -based ordering of the nodes.*

After these iterative rounds, more relevant data was to be added. On top of the data already added, primary information for the users also includes whether an endpoint is considered

a core endpoint or not and whether it has been meta-analysed with Estonian or UK biobanks or not, as well as information on the amount of genome-wide significance loci (GWS hits). This data was acquired through two different data files, another one being a tab-separated file and another one being a spreadsheet. Both of these were transformed into a JSON file in Python, after which relevant information could be matched with the data objects corresponding to each endpoint in question in the original file and added there as key-value pairs.

The information on whether an endpoint is a core endpoint or not is coded with an additional signifier dot that follows the visual conventions of its parent node for the sake of quick sketching; This would be changed later on to comply with the principle of allowing ample separability between visual markers coding divergent categories. The genome-wide significance (GWS) data, on the other hand, is numeric information that tells the amount of genome-wide significance loci related to each endpoint in question. In order to avoid visual clutter, this information was coded with a heatmap utilizing the visual space available in the circle of the node; The more GWS hits allocated to an endpoint, the darker the luminance of the node's fill colour. Heatmap through darkening or lightening the hue works as a natural quantifier for GWS information that does not have internal categories, but only increasing number of assigned hits. Bolder hue that highlights the endpoints with most hits guides the user to a direction of an endpoint of interest without necessarily having to filter based on GWS hit information. GWS hits were then double coded into the nodes by adding a number to



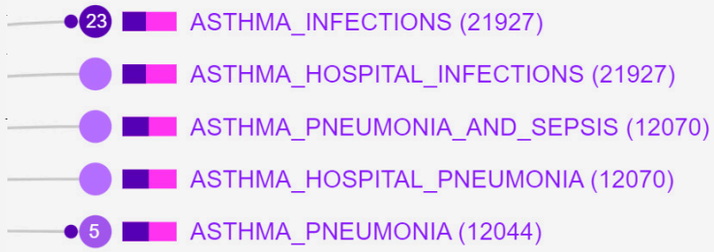



*Figure 18. Screenshot showing a closer look at the double-coding of the GWS hit information and the bars indicating the binary gemder balance.*

At this point the possible issue of missing case number data was addressed; Some endpoints do not have case number information at all, and some of them might have only female or male case information. This was solved by assigning the hue encoding to a same one that the node itself has by default (*see figure 19*), so that the missing information is not filled in the bar if there is none. The practice of recognising incomplete data and showing those gaps is in direct alignment with the agile design framework needed for custom scientific visualizations.

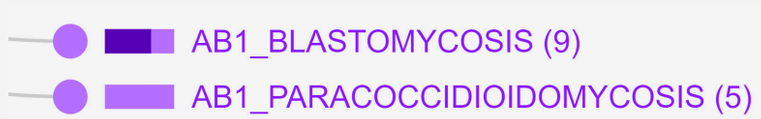



*Figure 19. An example where blastomycosis has a totality of 9 cases of which 6 are male, but no female cases are reported. Paracoccidiomycosis, on the contrary, returns null for gender balance data dispite having five cases assigned to it.*

As the next step of the process, the interface layout was adjusted to a composition that provides a more meaningful viewing of a single tree and also provides a list of all the ICD-10 blocks that are fetched dynamically from the data (*see figure 20*). The list of ICD-10 blocks is a scrollable list of clickable elements that render the tree matching the name of the ICD-10 block that is clicked by the user. The use of a list element is not directly supported by the results of the user interviews, but is rather

an additional component to support the search criteria of the blocks that has not been available before. A visible list available at all times might also support exploration among the trees, despite that not being one of the main goals of the browsing tool. The position of the search-supporting interface elements on the left and the resulting tree on the right support a left-to right reading direction.

The indicator for the core endpoint information was also adjusted in the process. A dot was switched to a diamond shape available natively in D3.js to provide more visually distinguishable features from the circular node itself. Case number formatting was also adjusted so that case numbers reaching thousands would be easier to comprehend at a glance (*see figure 21*).

On top of advancing the interface and existing visual elements, the visualization was enriched by adding the meta-analysis information for the endpoints; this information indicates whether a meta-analysis in the Estonian or UK biobanks has been performed on an endpoint. Meta-analysis information was visualised through the usage of country codes to support the most intuitive reading of the visualization while avoiding visual clutter in the form of adding a longer text string or additional graphic elements (*see figure 21*).

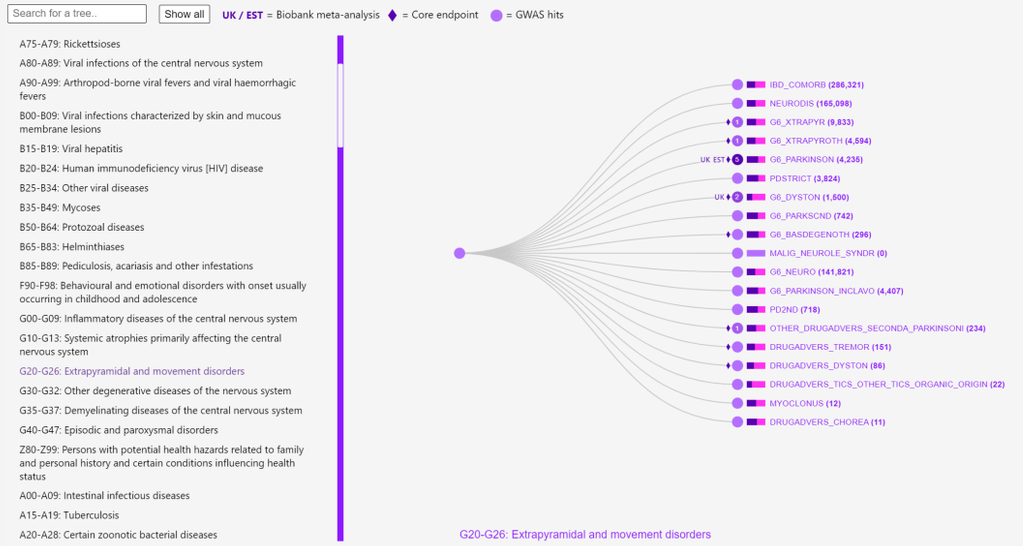



*Figure 20. Screenshot showing the block list browsing as well as the totality of the visual encoding methods added.*

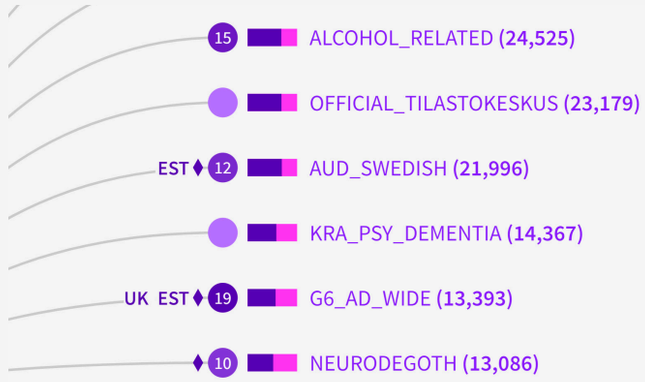



*Figure 21. A closer view of the added and improved visual encoding methods used.*

All of the information added to the core data file and the visualization at this point have been of a similar level of priority, and therefore represented at an equal level in the visualization system. Following steps, however, included the implementation of registry code information relevant for each endpoint's definition; The registry code information has the potential to be vast containing multiple different codes from different registries and as such will be quite a heavy element visually. Considering the above, their representation will be examined through a layer of interaction in order to make it available for the users but not to completely clutter and occlude the initial display of all the endpoints in a category.

First step to consider in implementing new, nested information to the existing data was the data file design and wrangling in a manner that would best assist the following visualization flow. FinnGen analysts were asked to construct a JSON data file containing each endpoint as a parent element equipped with child elements containing individual code's information as separate objects. Codes would not be constructed as simple lists inside the JSON but rather as key-value pairs inside an object for the fact that some information regarding the registry codes might be best to access through their keys without interrupting other information concerning the same code. In other words, information separation provides more flexibility and precise targeting in terms of the visualization workflow. Figure 22 demonstrates the related data file design, and figure 23 shows a snippet of the actual dataset after the data design and wrangling iteration. After this round of data enrichment, the information was added

```
[{
    ENDPOINT: "Name of endpoint"
    CODES:  [
                {
                    REGISTRYCODE_LONGNAME: "Atrial fibrillation and flutter",
                    REGISTRYCODE_CASES: 35847,
                    REGISTRYCODE_CODES:
                    [{
                        CODE_KEY: "I48",
                        CODE_NAME: "OUTPAT_ICD10"
                    }]
                },
                {
                    REGISTRYCODE_LONGNAME: "Chronic arrhythmias",
                    REGISTRYCODE_CASES: 25354,
                    REGISTRYCODE_CODES:
                    [{
                        CODE_KEY: "207",
                        CODE_NAME: "REIMB_KELA"
                    },
                    {
                        CODE_KEY: "I48",
                        CODE_NAME: "ICD10"
                    }]
                }
            ]
}]
```

```
"block":"A15-A19",
"decription":"Tuberculosis",
"children":[
    {
        "endpoint":"AB1_RESP_TUBERCU_CONF",
        "n_cases_all":543.0,
        "n_cases_female":162.0,
        "n_cases_male":381.0,
        "core_endpoint":"yes",
        "est_meta_analysed":"No",
        "uk_meta_analysed":"No",
        "gwas_hits":0,
        "definition_codes":[
            {
                "CODE_KEY":"A150",
                "CODE_SYSTEM":"INPAT_ICD10",
                "CODE_LONGNAME":"Tuberculosis of lung, confirmed by sputum
                microscopy with or without culture",
                "CODE_CASES":201
            },
            {
                "CODE_KEY":"A150",
                "CODE_SYSTEM":"OUTPAT_ICD10",
                "CODE_LONGNAME":"Tuberculosis of lung, confirmed by sputum
                microscopy with or without culture",
                "CODE_CASES":181
            },
```

*Figure 22 and figure 23.Screenshots describing the data file design plan (above), and the core data file after enriching with the data acquired from FinnGen (below).*

to the visualization in a robust manner to get an evaluation of the usefulness of the information from the user's point of view. Code information was implemented as a list of top ten definition codes that appears while hovering over each endpoint. The list of codes includes the code itself, the code system that it belongs in, the more descriptive name for the code and the cases assigned to the endpoint including the code in question (*see figure 24*); The instances where definition codes were not available or included in the data were accounted for with a message recognizing the missing definition. The list was separated from the general tree area with a dashed line in order to lighten the separation visually instead of creating a barrier with a heavy line.

Change blindness resulting from jump cuts while changing from one tree to another was also combatted with a slight transition animation lasting for half a second targeting the opacity property. The starting state was also refined to prompt the user to search for or select a block that interests them or search for a relevant endpoint to see that in the context with other related endpoints. Adjustments were also made for the legend to include all of the visual encoders as well as implementing a sticky positioning for it in order to keep the legend available while scrolling down vertically wide tree structures.



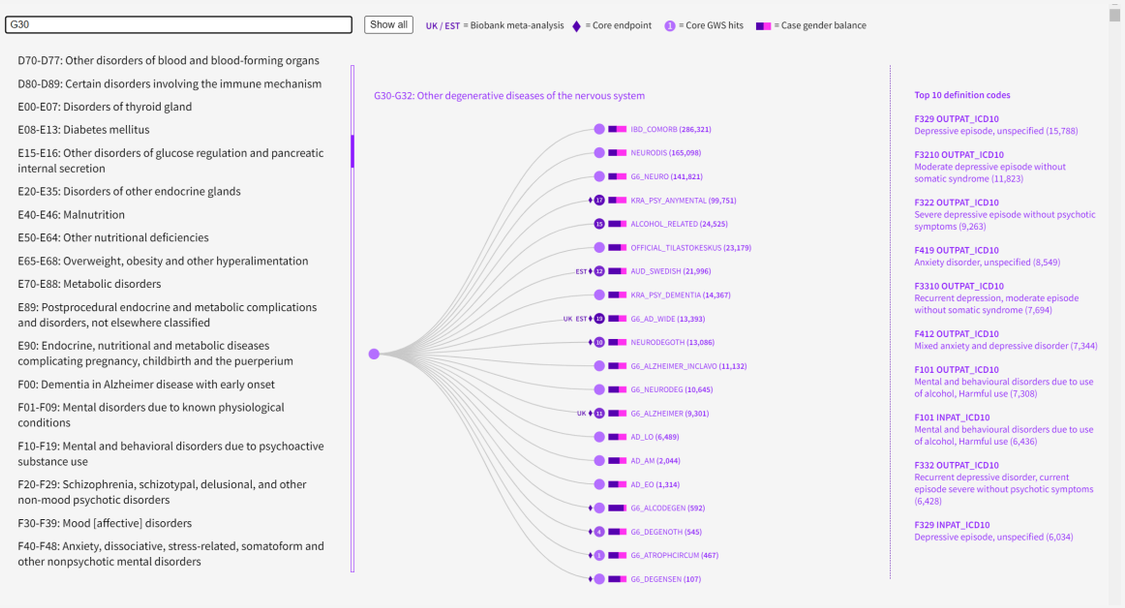



*Figure 24. View of the state of the visualization as a whole showing the top ten definition codes for the endpoint G6_NEURO.*

### *Midpoint evaluation discussion with the prototype*

Iterative evaluations and communication with the users are crucial factors when following a user-centered, contextual design process. Following this principle, a small-scale midpoint evaluation was performed on the prototype; After the iterative data handling, sketching and design process described above, the production reached a state where the prototype was relatively stable and could be tested by the users. The goal for the midpoint evaluation was to get the expert user's reactions concerning the prototype at its current state and test the interaction methods as well as the visual encoding techniques used.

A convenience sampling on expert users familiar with the production was performed resulting in three users, two from a research geneticist background and one from a clinical background. The interview was planned as an unstructured group interview, where questions are open-ended and directed towards enabling discussion between the interviewer and the participants; This was determined appropriate since scoping of the project and understanding of user's behaviours regardig their tasks was mapped in the initial state of the user research. Being prepared to follow unanticipated lines of discussion and listening to the participants during the development is vital considering the nature of the application and the user-centered design methodology.

In the interviews, the scope and purpose of the visualization was re-iterated for the participants concluding that it is meant to aid the comparison of endpoints prior to inspecting the focus endpoint's landing page in Risteys, as well as providing context of their relationship to other endpoints in the same ICD-10 block. A link was then shared containing the interactive prototype[103].

*103. First version of the interactive prototype. https://geneviz.aalto.fi/endpoint_browser/v1/*

Key topics from the discussion included both the attributes and the structure of the data itself and the visualization methods used. In terms of the data and the related architecture, users raised questions about decisions to include some of some endpoints that could have been omitted in an earlier stage of the data handling; Possible future exclusion of these was mentioned, as well as adding complimentary categories such as medication endpoints as their own, novel blocks that did not exist in the data structure yet. Addressing these concerns and aiming for refining the structuring of the data will make the reading of the visualization more streamlined in the next version of the prototype. As well as fining down the data, including the longer, more descriptive names of the endpoints on top of the currently displayed endpoint's code name was mentioned as a sure way to help users identify their relevant endpoints.

The amount of information encoded in the visualization was found sufficient considering the goal of making comparisons and choices based on these attributes. Further evaluation would be preferable, but these remarks imply an indication of successful scoping of the primary information in terms of locating focus endpoint(s).

In terms of interaction methods, some improvements are to be made in terms of locating the endpoint that is being searched through the search bar; Highlighting those in the structure or redefining the view in other ways would be a definite part of concurrent steps in the project. Currently, the search bar only renders a tree that includes the string from the search query, but does not guide users to the endpoint specifically.

Filtering methods were also suggested to be tied into the legend as users might benefit from looking into meta-analysed core endpoints that have GWS hits, for example. The knowledge gained through the literature review of the working memory and other cognitive abilities concerning visual searches also support this remark. Currently the attributes are only flagged with appropriate glyphs, the reading of which is enabled through the legend. Resulting from flagging instead of filtering, all the nodes remain of relatively equal visual weight throughout the workflow. Implementing more salient user-initiated highlighting strategies might better assist the users in determining their focus endpoints and not having compare between too many instances in the same hierarchical level. Refinements in the data itself mentioned earlier will also contribute to concentrating the number of comperable nodes.

No major confusion was detected in terms of interpreting the legend and reading the visual encoding, although some additions and points of further attention were discussed. A legend symbol coding the node showing the GWS hit numbers should be redefined to include guidance that the number and the heatmap inside the node signal the number of hits, as the legend currently only states an empty node and does not include a visual example of an endpoint containing GWS hits. Clarification on the fact that the GWS hit numbers refer to the core endpoint would further reduce the potential for ambiguity.

A more intuitive legend symbol coding the missing gender balance data would also streamline the reading of the visualization; At the time of midpoint evaluation, the prototype features a base color identical for the base colour of the nodes for the bar to indicate missing data. Solution could relate to showing the area of the bar where the information is missing as empty with an outline rather than using any base color that can signify content when there, in fact, is none. Considering the fidings above, a more comprehensive, interactive legend explaining each visual symbol in their spatial context in the visualization could be a more sustainable solution for the future development of the application.

As for the signifiers used in the visualization itself, a point was made regarding the balance of the visual hierarchy of the meta-analysis and core endpoint information; The country code tags encoding meta-analysis information were found to be visually overpowering compared to the diamond symbol coding the core endpoint information. This can be adjusted without changing the encoding, but further evaluation could prove useful in determining whether that should be necessary.

## *Prototype version 2: Collapsible tree*

An alternative version of the prototype was also produced as a part of this study[104], but it's user evaluation and visual refinement will be the subject of future developments of the visualization project. The second version features slight improvement for the section of the legend indicating GWS number information based on the findings of the midpoint evaluation; Now both the information that the GWS hits refer to the core endpoint as well as the coding of those hits was made clearer. The visualization structure itself features an interactive, collapsible, dynamically adjusting tree structure that visualises the definition codes used in each endpoint's definition as subtrees to the original parent tree encoding the ICD-10 block information thus enabling comparisons between definition codes of the endpoints. Implementation of the top definition codes as collapsible subtrees to individual endpoint nodes more clearly communicates them as internal parts of the endpoints integrating them to the same parent structure comparing to the method deployed in the first version of the prototype. Expanding on a simple tree structure in this way has a lot of potential for further versions, since it allows the visualization to span into a direction of a semi-tree or a network structure visualizing wider contexts allowing further explorations.

*104. Second version of the interactive prototype. https://geneviz.aalto.fi/endpoint_browser/v2/*

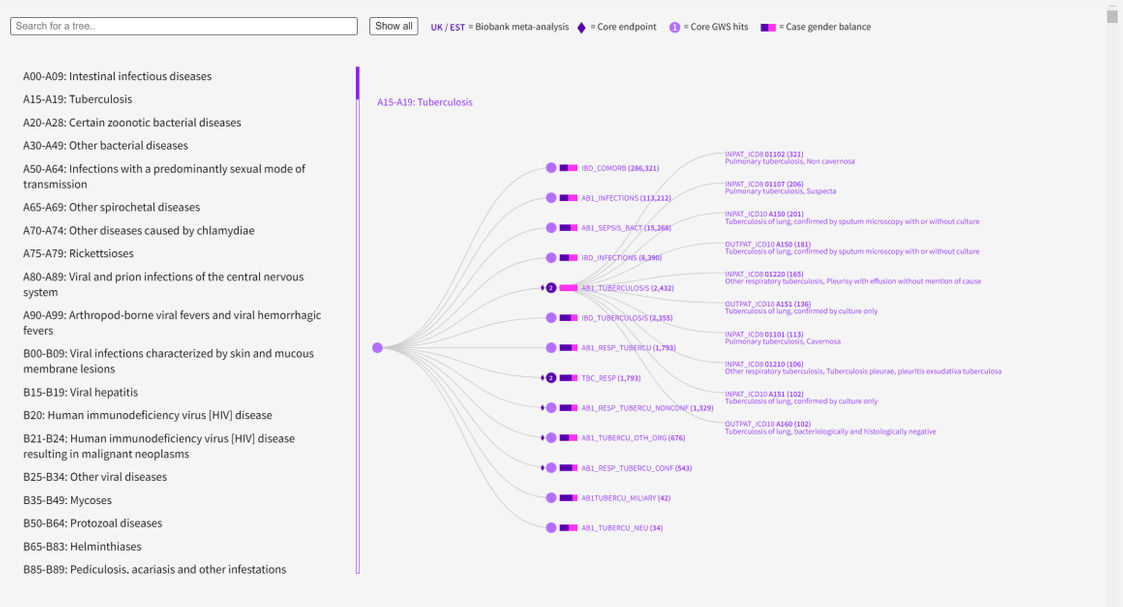
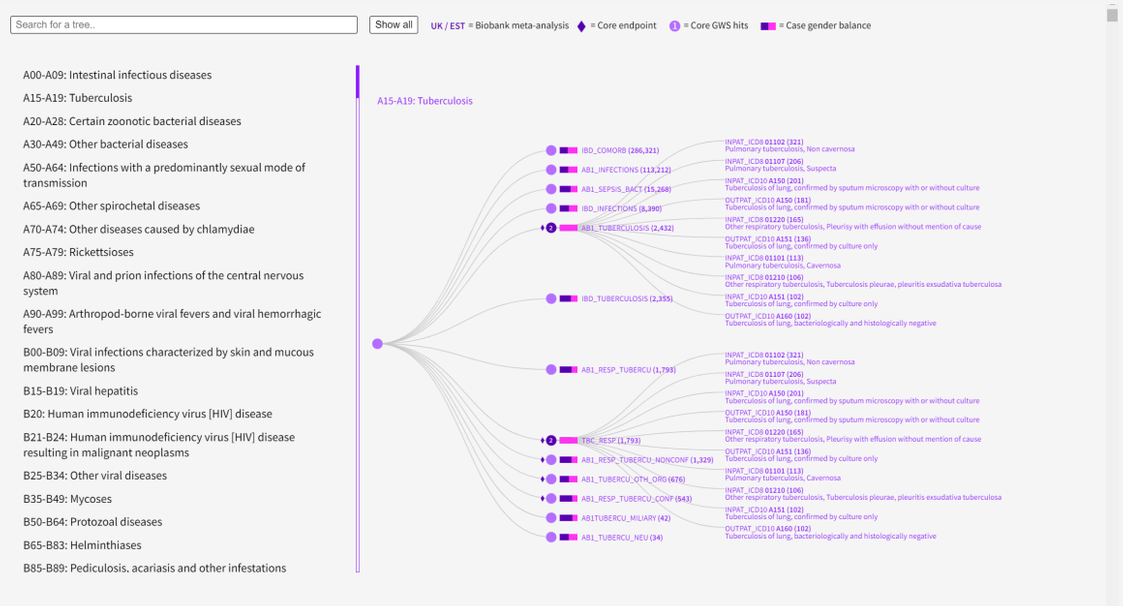



*Figure 22 (above) and figure 23 (below). Screenshots demostrating the dynamic scaling and the workings of the collapsible tree model from version 2.*



# *Conclusions*

To conclude this study, an overview of the answers relating to the research questions is made based on the results of the literature review, user study and the production parts of the study. After these, limitations are discussed followed with a conclusive chapter on learning outcomes and future developments.

### *RQ1. Are users experiencing challenges with FinnGen's phenotype browsing system Risteys?*

Key findings of the user study indicate that expert users were generally content with the service and the workflow opportunities that Risteys provides, but nevertheless some opportunities for improvement were detected. Observing the user's task behaviours through a think-aloud method as well as engaging in discussions through a semi-structured interview revealed that the users share a motivation to do comparisons between differend endpoints that fit their search criteria in order to find a focus endpoint that best responds to their research needs. However, these kinds of comparisons are not, at the time of conducting the user study, supported in the Risteys search model which leads users to inspecting different endpoints separately in order to determine the best one, going back and forth in the search results. Risteys does return different endpoints fitting to each search query as a part of the search model, but users were still eager to connect these results into a wider context in order to interpret the results. Browsing of the phenotypes was only enabled through a limited scope of pre-determined categories, even though the users are expressing motivations for grasping the connections and the bigger picture of a larger group of endpoints beyond their exact search results.

Scoping and determining primary information concerning the comparison of possible focus endpoints was also done within the scope of the user study. These features were the information on whether the endpoint is a core endpoint or not, whether it has been analyzed within Estonian or UK biobanks or not, the genome-wide significance (GWS) hits of each endpoint and the case numbers assigned to each endpoint as well as the gender balance data of those cases. Complimenting these, the users often rely on the definition codes of each endpoint in being able to form an informed decision on whether or not that particular endpoint fits their research interests and comparing them with other phenotypic information.

To conclude, browsing and providing wider context for each endpoint is missing from Risteys resulting in the users feeling unsure about their position of orientation in relation to other similar endpoints. The users are eager to gain a more open vantage point over the endpoints in Risteys; This motivation was shared among the users regardless of their backgrounds, and so this became the focal point to address through the visualization tool.

### *RQ2. How can a user-centered data viz design practice answer to challenges related to phenotype browsing?*

User-centered data visualization practice provides a way to respond to usability challenges in data-driven systems by deploying techniques that consider the cognitive attributes of the users both by structuring the information as well as by scoping and encoding it. Visualizing structures, hierarchies and attributes of the data with a customized visualization solution provides an agile framework through which usability challenges and visualization problems can be addressed; Through this framework, the objectives for the design process of this study were directly derived from the user study. The goal was to first initiate a meaningful scope of comparison for the users that would provide a wide enough context for the users in order to make comparisons between endpoints as well as understanding their relationships to other endpoints, and secondly, gather all the necessary data and visualize it in a way that supports the users goal of making quick and effective comparisons between endpoints. Visual encoding methods are effective in supporting these comparisons, since the users can simply detect information compared to searching for it; This is the primary benefit of deploying data visualizations in general.

Determining relevant categorisation criteria is the key to establishing meaningful scopes for browsing through phenotypic information. Categories need to be well understood in order to take full advantage of them and in being able to form meaningful connections from the reading of the visualization; Data visualization methods provide a quick way to sketch with data and form an understanding on what kinds of categories will be formed with each criteria. Based on the results of the user study and intial data sketches, the scope for meaningful comparisons was set to the ICD-10 blocks that would act as neighbourhoods to provide familiar, contextual information about the endpoints for the users.

Besides visualizing structures and connections, data visualization practice is particularly helpful in addressing challenges in phenotype browsing since it provides a flexible toolkit for encoding vast amounts of information regarding the phenotypes in a manner that will be quick and intuitive to read and benefit from. Interactive visualizations encourage explorations with the data, that in turn have potential to lead into new discoveries. A comprehensive understanding of the visualization methods available and best practices relating to their use are helpful in addressing the complexities and particularities of scientific data beyond phenotype browsing as well.

### *Limitations and critical reflections*

In order to gain a more widespread understanding of how the group of expert users use the phenotype browsing system Risteys and to be able to fully take advantage of the contextual inquiry as a method of the user study, inspecting behaviours in users realistic working life situations would have likely proved more useful than task observation and the think-aloud method. This approach would not necessarily replace, but compliment the information gathered from the interviews and show a more

realistic view on different multi-tasking and environmental factors that the users might face in their everyday life. Lack of proper task and voice transcription should also be considered a limitation relating to the user study.

In terms of technical execution, the subjective limits in prior knowledge-scapes always present limitations, as was also the case for this study. The limitations were not debilitating by their nature, but posed delays for the development of the prototype, as was to be expected especially in the case of data wrangling where prior knowledge base was limited.

In terms of endpoint taxonomies and their truthful representations, the need to address endpoints not fitting into the established categories will remain a task for the future developments of the production. Not all endpoints can be described by an ICD-10 block, so they have not been included in the data file enrichment round where endpoints were assigned to particular blocks. Additional categories differentiated from the ICD-10 blocks are one possible solution, but how to properly represent and acknowledge the connectedness of these endpoints and their benefits for the users remain open questions to be addressed outside of the thesis scope.

Future possibilities exist for expressing a wider range of interconnectivity between the data outside of the current hierarchical structure. Showing related endpoints in other trees or the positioning of the focus endpoint in an endpoint hierarchy separate from the ICD-10 -related hierarchy might result the future visual architecture of the tool to lean towards a network structure; However, the scope, design and value of such reframing should be evaluated through the users before further considerations.

Some practicalities concerning the production will continue outside of the scope of this case study; Among those is the deployment of the application meaning its embedding to the system where it is to be used, in this case FinnGen's Risteys environment. Risteys is a public platform, so after the possible deployment the application will be available to use for anyone, although the most likely user groups will remain the expert users working in different collaborative manners in FinnGen. Risteys uses Vue.js as their framework, so migrating the current application to correspond to the structure will also be an expected part of the upcoming workflow in case of deployment.

The dynamic nature of the endpoint tree poses some limitations relating to its visual design that were not yet addressed within the scope of this study. Among those are aspects of typographic hierarchies and the sizing of the subtrees in the second version, where the flexible sizing of the subtrees and their nodes can result in imperfect displays were the overflows and typographic adjustments have not been harmonized within the display. Considering the visual design process, it should also be noted that no individual designer can completely avoid their personal biases, and so those also pose an inevitable limitation on what could be considered best solutions in the production despite taking into account the conventions and best practices of the field. This study recognises that the final product is always a compromising act between the different goals, design options and technical limitations.

## *Learning outcomes and future developments*

This thesis has presented how theories and techniques emphasised in user-centered data visualization design research and practice can contribute to the process of designing and developing a data visualization tool for browsing of phenotypic data. Establishing common ground and reaching towards concrete collaborations between scientists and designers in the context of data visualization tools will without doubt be continued, and further research will hopefully strengthen and expand the knowledge space that these shared practices create. Further transparency, documentation and resources allow both scientists and designers to engage in these processes and see the benefits of cross-disciplinary collaborations that have the potential to create value for both fields of study.

A successful result in a visualization project within the realm of sciences calls for close collaboration with the designer of the tool and the scientists themselves. The design process relies heavily on the information that the scientists, the users, are providing in defining the project scheme. The same applies in clarifying what are some of the most interesting and important features of the data in order to define the visual hierarchy together with the suitable visual encoding methods. These decisions will determine which attributes of the data will be emphasized over others; This is information that the designer can not, and in most cases should not, attempt to determine without certainty from the expert users. On top of prioritizing certain aspects of the data it is important to make sure that no relevant information is hidden or removed in the design process; As important as the discussion on what should be shown and in which manner is the discussion of what should be left out of the visualization in order for the end result to be well scoped in term of visual representation. Underlining these considerations, a user-centered approach is critical for a successful practice within these interdisciplinary collaborations.

This study has also shed light on what to expect as a designer that is engaging in the process of creating scientific visualizations. The possible instabilities and complexities of scientific data can steer the design and production process towards agility and multiple iterations, where an open line of communication, knowledge of the subject as well as the ability to consider visual handling of partially or completely unavailable instances of the data all become a factor that shape the design process. Particularly in the case of the design production of this study, the iterative process and shaping the source data and dealing with the implications of it had great impact on the process. From a visual communication designer's perspective, embarking on a journey to learn about these visualizations is an enlightening one providing an opportunity to expand existing practices to new directions. Creating customised solutions that have tight-knit roots on the needs of scientific community inevitably pose challenges and requires good communication skills as well as the ability to tackle and take on new areas of knowledge both in the sense of tools, techniques and technologies as well as on the sciences themselves. On top of gaining this novel practical knowledge, learning about collaborative processes including

open cross-disciplinary communications and co-creating solutions as a team is valuable knowledge for any designer. As well as engaging in open communication within the team, having a tight connection with the users and accounting for their needs is an invaluable asset in being able to shape a successful data visualization design practice and an end product that adds value to its users.

# *References*

Aisch, G. (n.d.). chroma.js palette helper. Chroma.Js. Retrieved April 8, 2022, from https://gka.github.io/palettes/#/9|s|00429d,96ffea,ffffe0|ffffe0,ff005e,93003a|1|1

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological science, 15(2), 106-111.

Bertin, J., & Berg, W. J. (2011). Semiology of Graphics. Amsterdam University Press.

Bostock, M. (n.d.). D3.js - Data-Driven Documents. D3.Js. Retrieved April 6, 2022, from https://d3js.org/. Available also from https://github.com/d3/d3

Cairo, A. (2012). The Functional Art: An introduction to information graphics and visualization (1st edition). New Riders.

Clinical endpoints. FinnGen Clinical Endpoints. Retrieved April 5, 2022, from https://www.finngen.fi/en/researchers/clinical-endpoints

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. Mind, Culture, and Activity, 5(3), 178-186.

Furnas, G. W. (1986). Generalized fisheye views. Acm Sigchi Bulletin, 17(4), 16-23.



Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2017). The anatomy of phenotype ontologies: principles, properties and applications. Briefings in Bioinformatics, 19(5), 1008–1021. https://doi.org/10.1093/bib/bbx035

Healey, C. G. (1996, October). Choosing effective colours for data visualization. In Proceedings of Seventh Annual IEEE Visualization'96 (pp. 263-270). IEEE.

Heer, J., & Robertson, G. (2007). Animated transitions in statistical data graphics. IEEE transactions on visualization and computer graphics, 13(6), 1240-1247.

Introduction - FinnGen Analyst Handbook. (n.d.). FinnGen Analyst Handbook. Retrieved April 12, 2022, from https://finngen.gitbook.io/finngen-analyst-handbook/

Islam, M. N., & Bouwman, H. (2016). Towards user–intuitive web interface sign design and evaluation: A semiotic framework. International Journal of Human-Computer Studies, 86, 121–137. https://doi.org/10.1016/j.ijhcs.2015.10.003

Itti, L., Rees, G., & Tsotsos, J. K. (Eds.). (2005). Neurobiology of attention. Elsevier

Koponen, & Hildén, J. (2019). Data visualization handbook (First edition.). Aalto University School of Arts, Design and Architecture.

Lima, M. (2011). Visual Complexity: Mapping Patterns of Complexity. Princeton Architectural Press

Liu, Z., & Heer, J. (2014). The effects of interactive latency on exploratory visual analysis. IEEE transactions on visualization and computer graphics, 20(12), 2122-2131.

Maciejewski, R. (2011). Data representations, transformations, and statistics for visual reasoning. Synthesis Lectures on Visualization, 2(1), 1-85.

McDonald, S., McGarry, K., & Willis, L. M. (2013, September). Thinking-aloud about web navigation: the relationship between think-aloud instructions, task difficulty and performance. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 2037-2041). Sage CA: Los Angeles, CA: SAGE Publications.

Meta Platforms. (n.d.). React – A JavaScript library for building user interfaces. React JS. Retrieved April 6, 2022, from https://reactjs.org/. Available also from https://github.com/facebook/react/

Microsoft. (n.d.). Visual Studio Code - Code Editing. Redefined. Visual Studio Code. Retrieved April 6, 2022, from https://code.visualstudio.com/

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Ferster, B. (2012). Interactive visualization: Insight through inquiry. MIT Press.

Munzner, T. (2014). Visualization Analysis and Design (1st edition). A K Peters/CRC Press.

Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., & Zhou, Y. (2003). Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. In ACM SIGGRAPH 2003 Papers (pp. 453-462).

Norman, D. A. (2010). Living with Complexity. MIT Press.

Norman, D. A. (2004). Emotional design: Why we love (or hate) everyday things. Basic Civitas Books.

O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., ... & Procter, J. B. (2018). Visualization of biomedical data. Annual Review of Biomedical Data Science, 1, 275-304.

pandas - Python Data Analysis Library. (n.d.). Pandas. Retrieved April 6, 2022, from https://pandas.pydata.org/. Available from https://zenodo.org/record/3715232#.Yk1QKp-U-QR and https://github.com/pandas-dev/pandas/tree/v1.0.3

Peute[a], L. W., Spithoven[a], R., & WM, P. J. B. M. (2008).

Usability studies on interactive health information systems; where do we stand?. In EHealth Beyond the Horizon: Get IT There: Proceedings of MIE2008, the XXIst International Congress of the European Federation for Medical Informatics (p. 327). IOS Press.

Ritter, F. E., Baxter, G. D., & Churchill, E. F. (2014). Foundations for designing user-centered systems. Springer-Verlag London, DOI, 10, 978-1.

Sadoski, M., & Paivio, A. (2012). Imagery and text : A dual coding theory of reading and writing. Taylor & Francis Group.

Samsel, F., Bartram, L., & Bares, A. (2018, October). Art, affect and color: Creating engaging expressive scientific visualization. In 2018 IEEE VIS Arts Program (VISAP) (pp. 1-9). IEEE.

Sharp, H., Preece, J., & Rogers, Y. (2019). Interaction Design: Beyond Human-Computer Interaction (5th ed.). Wiley.

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In The craft of information visualization (pp. 364-371). Morgan Kaufmann.

Shneiderman, B., Plaisant, C., & Hesse, B. W. (2013). Improving healthcare with interactive visualization. Computer, 46(5), 58-66.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. Interacting with computers, 13(2), 127-145.

Tufte, E. R. (2006). Beautiful evidence. Graphics Press.

Tufte, E. R. (1991). Envisioning information (2nd pr., with rev.). Graphics Press.

Tufte. (1984). The visual display of quantitative information. Graphics Press.

Van Ham, F., & Perer, A. (2009). "Search, show context, expand on demand": supporting large graph exploration with degree-of-interest. IEEE Transactions on Visualization and Computer Graphics, 15(6), 953-960.

W3C, World Wide Web Consortium. (n.d.). Understanding Success Criterion 1.4.3: Contrast (Minimum). W3.Org. Retrieved April 6, 2022, from https://www.w3.org/WAI/WCAG22/Understanding/contrast-minimum.html

# *Annex*

# *Annex A*

# *Profiling questionnaire*

*A Google form that was sent to all the users who answered 'yes' to a poll mapping the interest of taking part in the user study. The purpose of the form was to gain background knowledge of the users and keep track of relevant variables such as the roles and expertise within the user group, as well as to map their knowledge of existing phenotype browsing systems other than Risteys.*



## Background information for phenotype browsing interview

This background information helps us getting to know you and structuring the interviews. Thank you!

keppostella@gmail.com (not shared) Switch accounts

*Required

Your name: *

Your answer

Select the title that best describes your main occupation in your work with FinnGen: *

- Clinician
- Data analyst
- Research geneticist
- Other:

If needed, select additional roles that describe your work with FinnGen:

- Clinician
- Data analyst
- Research geneticist
- Other:

Your area of research:

Your answer

How would you describe your familiarity with the current phenotype browser Risteys?(https://risteys.finngen.fi/) *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| I'm new to it | ○ | ○ | ○ | ○ | ○ | I'm very familiar with it |

Select any other phenotype browsing systems that you are familiar with:

- ICD10 tree
- SNOMED
- OHDSI/OMOP
- Finnish registry data
- UKBB phecodes and/or self-report data
- Japan BB
- Estonian BB
- GWAS Catalog
- OpenTargets
- HPO (Human Phenotype Ontology)
- MPO (Mouse Phenotype Ontology)
- BCPlatforms
- Medisapiens
- EFO from EMBL-EBI
- Other:


Aalto University
School of Arts, Design
and Architecture