Grant Agreement no. 643529

Empowering patients and strengthening self-management in cancer diseases

Research and Innovation Action

PHC-26-2014: Self management of health and disease: citizen engagement and mHealth

D8.2: Implemented visualization techniques

Contractual Due Date: 31st July 2017
Actual Submission Date: 11th August 2017

Lead partner for deliverable: University of Bedfordshire

Dissemination Level: Public

Revision: v1.0

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	***iManageCancer***
Project Full Name:	Empowering patients and strengthening self-management in cancer diseases
Project Duration	1 February 2015 - 31 July 2018
Deliverable No.:	D8.2
Deliverable Name:	Implemented visualization techniques
Nature (R, DEM)[1]	DEM
Dissemination Level (PU, CO)[2]	PU
Version:	1.0
Actual Submission Date:	11 August 2017
Editor: Institution: E-Mail:	Youbing Zhao University of Bedfordshire, UK youbing.zhao@beds.ac.uk
Contributors (Institution)	
Reviewers (Institution)	Lefteris Koumakis (FORTH)

Copyright

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643529.

[1] **R** = Document, report (excluding the periodic and final reports), **DEM** = Demonstrator, pilot, prototype, plan designs

[2] **PU** = Public, fully open, e.g. web, **CO** = Confidential, restricted under conditions set out in Model Grant Agreement

Document History

Issue Date	Version	Changes Made / Reason for this Issue
1st August 2017	0.9	Draft version for internal review
11th August 2017	1.0	Final version for submission

Table of Contents

1. Executive Summary

iManageCancer data analysis services require presentation, analysis and comparison of a large volume of heterogeneous data from cohorts of a large number of patients with covering areas from the medical, the environmental and the lifestyle domains. Without proper tools it is impossible to achieve this goal.

Work Package 8 "Smart analytical data services" is proposed to meet this challenge. In Task 8.1 "Data analysis and data mining services", data are extracted by the data mining services. In Task 8.2 "Visualisation", an interactive visual analysis interface is suggested be designed to empower the ender users to view, utilise, analyse and understand the mined data.

This document is a deliverable report of D8.2 "Implemented visualization techniques" of the iManageCancer project. In particular it covers the advanced visualisation tasks designated in Task 8.2. This deliverable report focuses on the design and implementation of visual analysis components in order to provide views and analysis of the iManageCancer patient and cohort data. The data is retrieved based on the work presented in D8.1 "Implemented data analysis and data mining services". This report is organised by the introduction, related work, data, the related visualisation techniques, the visual analysis system, the design and implementation of the single patient visualisation and the cohort visualisation application.

2. Introduction

iManageCancer patient data analysis services mine data from a variety of data sources covering the medical, the environmental and the lifestyle domains. Without proper tools it is almost impossible to present and analyse these large, heterogeneous, time-varying data.

The longitudinal health and medical data imply that a huge amount of data from a large number of sources needs to be collected, stored, processed and presented. Visualisation provides technologies to present the data via mature visual paradigms and well-designed user interactions. Without proper design of visualisation and user interaction, it is not possible for the user to select, view, understand and gain knowledge from a large collection of health data. "Visualization and visual analytics researchers can contribute substantial technological advances to support the reliable, effective, safe, and validated systems required for personal health, clinical healthcare, and public health policymaking" [Shneiderman 2013].

In iManageCancer, WP8 "Smart analytical data services" is proposed to address this challenge. In Task 8.2 "Visualisation", an interactive visual analysis approach is suggested be designed to empower both the patients and the medical experts to view, utilise, analyse and understand the mined data from Task 8.1 "Data analysis and data mining services".

Visual analytics is an integral approach which combines visualisation, human factors, and data analysis. This process incorporates automatic and visual analysis methods with a tight coupling through human interaction in order to view, analyse and understand the data. As a science of analytical reasoning facilitated by interactive visual interfaces, the interface of visual analytics is critical in presentation and analysis of the data mined from the data mining services. Visual analytics interfaces allow the analysts to interact directly with the representation of the data or to modify visualisation parameters. In the visual analytics process knowledge can be gained from visualisation, automatic analysis, as well as interactions between visualisations, models, and the human analysts.

In WP8 the visual analytics techniques deliver different perspectives on specific data types to provide heath data visualisation and cohort analysis. The interactive visual interface will directly support complex cohort study tasks. Interaction techniques are incorporated with the visualisation to facilitate data exploration and knowledge discovery.

The report of D8.2 is the deliverable report of Task 8.2 "Visualisation" which aims to provide a design of visual analytics interfaces and components for iManageCancer patients and medical experts to facilitate understanding and analysis of patient and cohort data. It is based on the work presented in D8.1 "Implemented data analysis and data mining services".

The visual interface will be a common access point for iManageCancer data analysis users, from which users can perform the typical data query and analysis. We define the visual analytics tasks provided by the iManageCancer visual interface as the follows:

- Visualising individual medical data from patients
- Visualising medical data of patient cohorts
- Visualising patient cohorts for cohort comparison and cohort study.

The system architecture is shown in Figure 1.

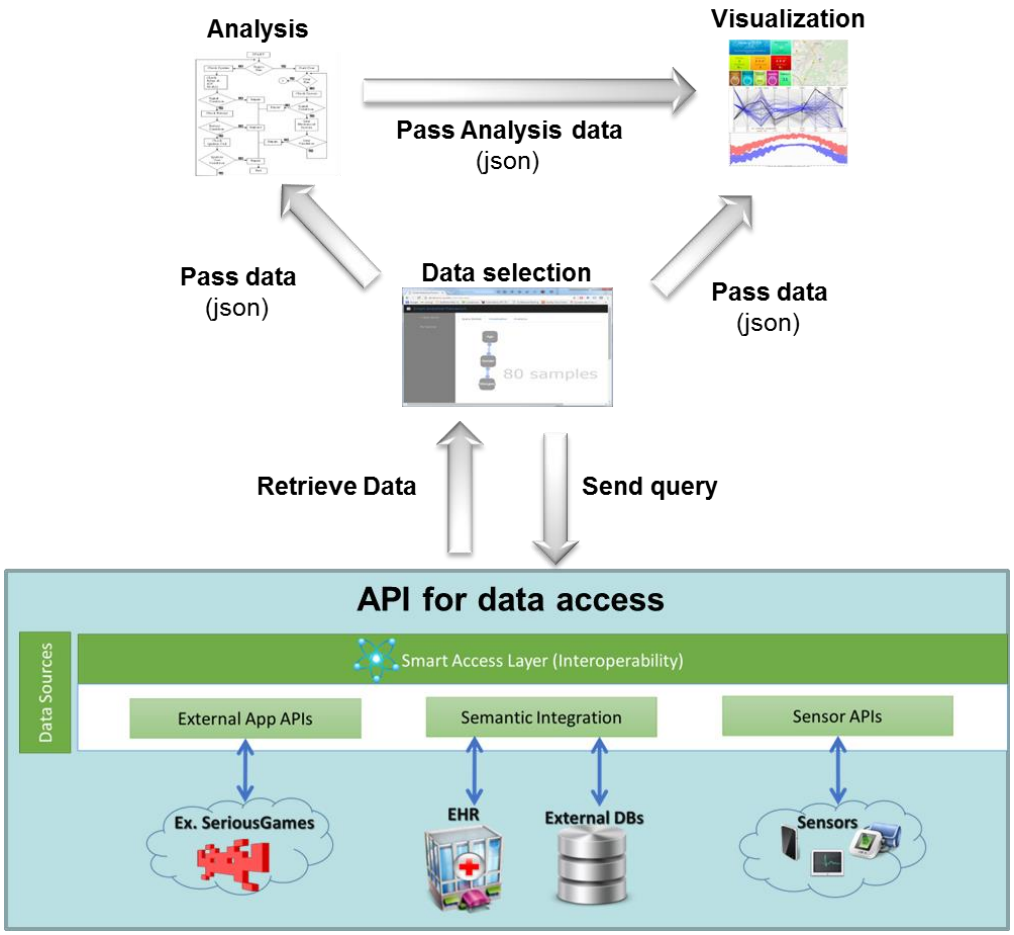


Figure 1. **Architecture of the analytical services over personal health infrastructure**

3. Background

3.1. Visual Analytics

Though there are several definitions of Visual Analytics listed as the follows:

- "outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces" [Wong 2004];
- "the science of analytical reasoning supported by interactive visual interfaces." [Thomas 2005];
- "combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets" [Keim 2010],

all of them point out that visual analytics is an integration of data analysis and interactive visualisation which introduces human intelligence at an early stage in the data analysis process. Visual Analytics methods allow decision makers to combine their human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems. Using advanced visual interfaces, humans may

directly interact with the data analysis capabilities of today's computer, allowing them to make well-informed decisions in complex situations.

The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. The visual analytics loop is in Figure 2 shows an abstract overview of the different stages in the visual analytics process.

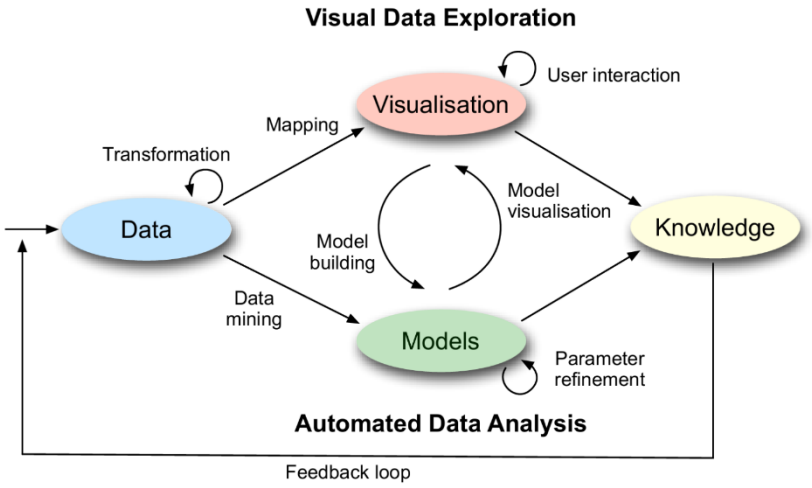


Figure 2 The visual analtyics loop

Visual Analytics can be seen as an integral approach combining visualization, cognition, human computer interaction and data mining, as shown in Figure 3, which illustrates the related research areas of Visual Analytics. The most closely related areas of Visual Analytics are information visualization and data mining.

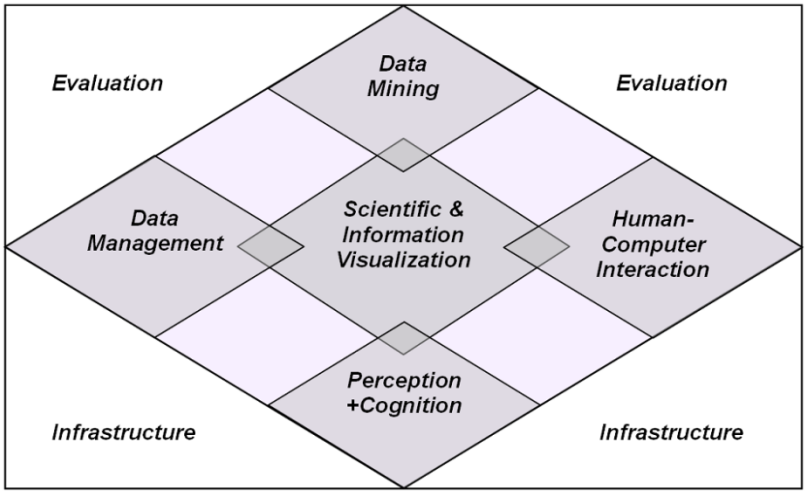


Figure 3 Related research areas of visual analytics [Keim 2006]

In iManageCancer, Visual analytics makes use of data from the iManageCancer data mining services and bring together valuable information in visual form to support health data visualisation and cohort study.

3.2. Time-varying Data Visualisation

In the real world, time-oriented data are ubiquitous in many application domains as, for example, in business, medicine, history, planning, or project management. Providing appropriate methods to facilitate the visualisation and analysis of time-varying data is critical for many applications. Very detailed reviews on time-varying data visualisation can be found in [Aigner 2011] (book) and [Aigner 2008] (survey paper).

In iManageCancer, the medical and health data of patients are time dependent. Moreover, most of cohort studies are based on longitudinal data. Visual analysis of time-varying data is an indispensable requirement in iManageCancer.

3.2.1. Time Data Modelling

The properties of time-varying data can be described by scale, scope, viewpoint, granularity, etc., which will be introduced as follows:

Scale:

Ordinal vs. Discrete vs. Continuous

For the ordinal time data, only relative order relations are present. Discrete time values can be mapped to a set of integers or labels which enable quantitative modelling of time values. Continuous time data are time data which can be mapped to a real domain

Scope

Point-based vs Interval-based

Point-based time data are discrete point data on the time axis with the length of the temporal extent equals to zero. Different from the point time data, interval-based time data represent continuous time spans on the time axis with temporal extents greater than zero.

Arrangement

Linear vs. Cyclic

The time data can either be linear or cyclic based on the need to process and present them. The basic time data is essentially linear and can be represented by points and spans on a time axis. However, periodicity is very common in time data due to the periodicity in the nature and people's life. The cyclic time data can often be represented by a radial layout design.

Viewpoint

Ordered vs. Branching vs. Multiple Perspectives

Ordered time data describe things that happen one after the other. Branching time data are multiple strands of time branch to allow description and comparison of alternative possibilities. While in branching time there is only one path through time will actually happen, multiple perspective data describe simultaneous views of time.

Granularity

None vs. Single vs. Multiple

Granularity refers to the human abstraction level of the time data. It describes mappings from time values to larger or smaller conceptual units. There may be a single granularity only or even none granularities are supported. Multiple granularities can also be used to support multi-scale visualisation.

In iManageCancer, most of the medical and health data are point based, linear, ordered data of numeric values or enumerate values.

3.2.2. Time-varying Data Visualisation

Linear and cyclic time data visualisation

The standard technique for visualising time-varying data is 2d plots where the x axis represents time and the y axis represents time-varying variables, as shown in Figure 4(a). This technique works well to show the variations of the data variable with the time but can hardly visualise the periodicity of time-varying data. The radial layout as shown in Figure 4 (b) and (c) can be used to uncover the periodic patterns in the data.

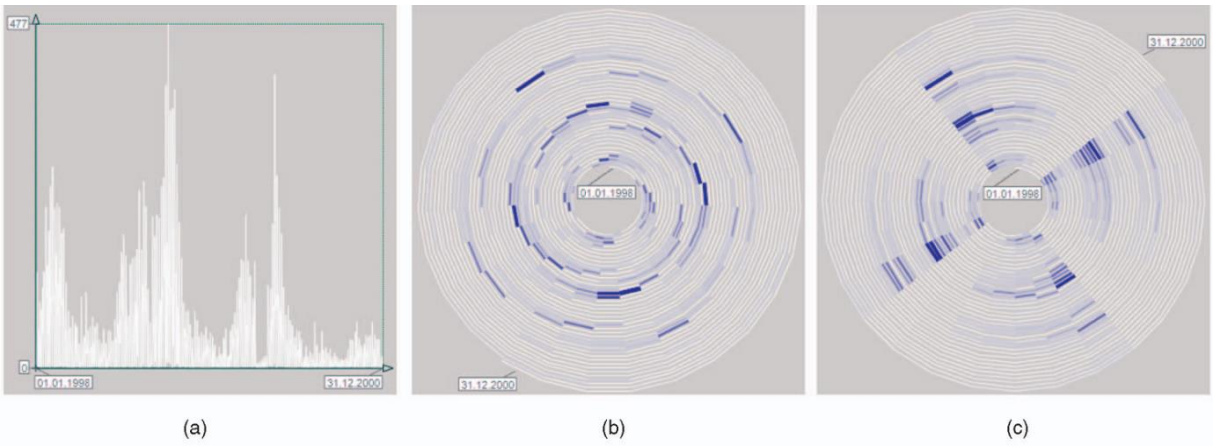


Figure 4 Different visual representations of a time-oriented data set

(a) Time series plot (periodic pattern is difficult to discern). (b) SpiralGraph encoding 27 days per cycle (improperly parameterized—periodic pattern is hard to see). (c) SpiralGraph encoding 28 days per cycle (properly parameterized—periodic pattern stands out),

ThemeRiver

ThemeRiver [Havre 2000] uses the metaphor of a river that flows through time. Currents within the river represent changes, such as climate changes, as shown in Figure 5.

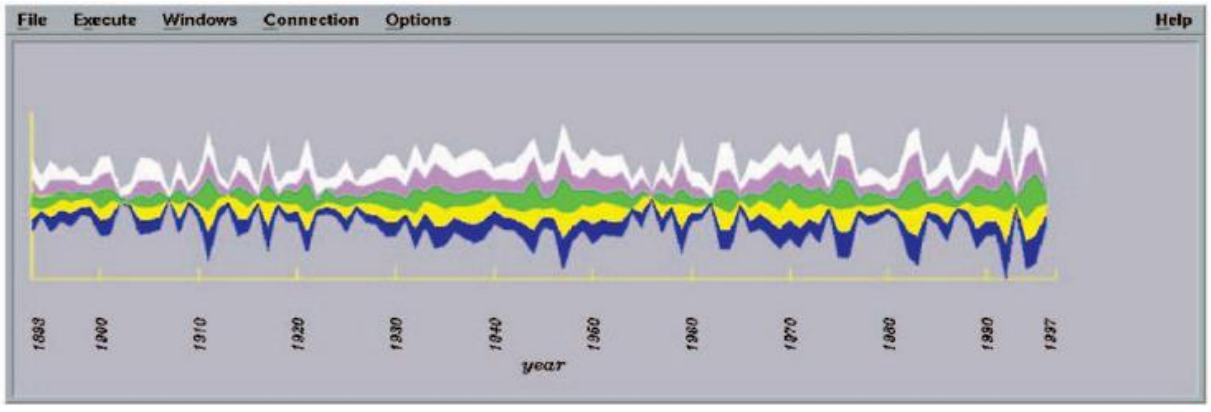


Figure 5 Visualisation of a climate data set using a ThemeRiver approach

The graph depicts five time-dependent variables: summer warmth (blue), summer days (violet), hot days (green), summer mean temperature (yellow), and mean of extreme (white) for a period of more than 100 years. Source [Aigner 2008]

Calendar visualisation

Calendar is the traditional way used by people to visualise the time long before the advent of computer graphics. Figure 6 shows a calendar heatmap [Calendar Heatmap] which visualises the attendance data in two years.

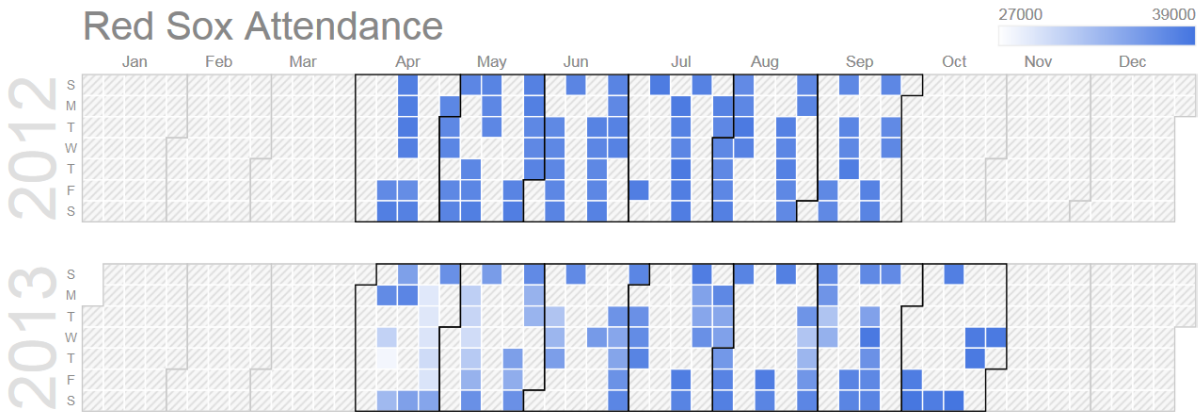


Figure 6 Visualisation of attendance for a sports team with a calendar heatmap

The brightness indicates the values. Source : https://developers.google.com/chart/interactive/docs/gallery/calendar

[Wijk 1999] is a well-known example which uses 3D and 2D calendar views to show the energy consumption and employee number data. It is a combined representation of daily patterns and clusters. Patterns are shown as 3D calendar graphs, clusters are shown on a 2D calendar, as shown in Figure 7. Colours indicate corresponding clusters and patterns.

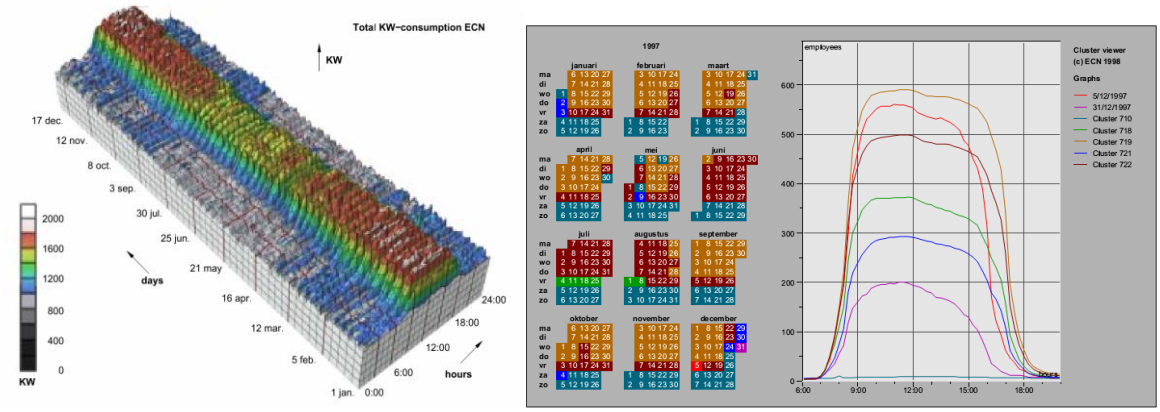


Figure 7 Calendar visualisation for pattern discovery

Left : Power demand by ECN, displayed as a function of hours and days, right: Calendar view of the number of employees.

Multi-perspective and branching data visualisation

LifeLines

LifeLines [Plaisant 1996] provides a general visualisation environment for personal history records, as shown Figure 8. A one screen overview of the record using timelines provides direct access to the data. For a patient record, medical problems, hospitalisation and medications can be represented as horizontal lines, while icons represent discrete events such as physician consultations, progress notes or tests. Line colour and thickness can illustrate relationships or significance. Rescaling tools and filters allow users to focus on part of the information, revealing

more details. LifeLines reduces the chances of missing information, facilitates the spotting of anomalies and streamlines the access to details.

Figure 8 Lifeline screenshot

Storylines

The river metaphor can represent the variation, merge, branching and ending of events, this effect is similar to the storylines used in fictions and movies to depict the temporal dynamics of social interactions. This visualisation technique was first introduced as a hand-drawn illustration in XKCD's "Movie Narrative Charts", as shown in Figure 9. If properly constructed, the visualisation can convey both global trends and local interactions in the data.

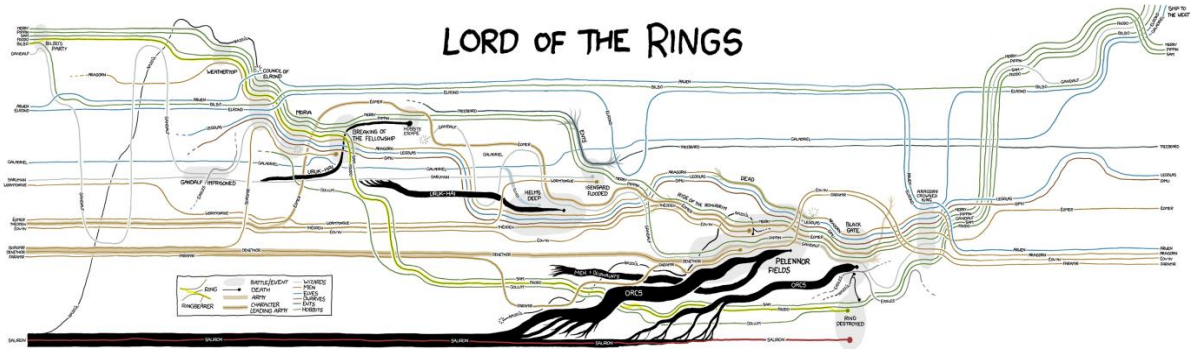


Figure 9 Handdrawn storyline visualisation of the main characters in Lord of the Rings

Source: http://xkcd.com/657/.

Dynamic Visualisation

Though animation has certain limitations and is not the mainstream technique for time-varying data visualisation, if used properly, it can still help the users to understand the whole process of the events more easily.

GapMinder

GapMinder [GapMinder] uses dynamic animation to visualise and compare the socioeconomical development of the countries over the world in last two centuries, as shown in Figure 10.

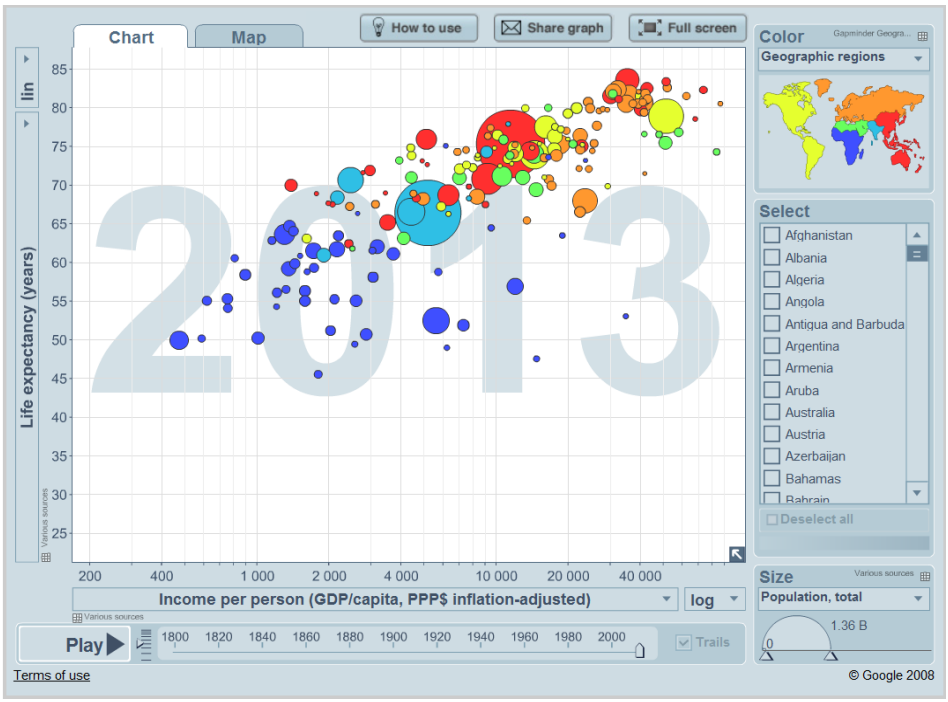


Figure 10 GapMinder's animation visualisation of GDP per capita of countries in the world from year 1800.

Source : http://www.gapminder.org

3.3. Health Data Visualisation

Visualisation and visual analytics combines automated analysis techniques with interactive visualisations for effective understanding, reasoning and decision-making on the basis of very large and complex datasets [Keim 2010]. It is designed to promote knowledge discovery and utilisation of large datasets via effective visual paradigms and well-designed user interactions. Visualisation becomes the medium of an interactive analytical process, where humans and computers cooperate using their respective distinct capabilities for data processing and visual recognition for the most effective results and is an indispensable technology for healthcare information representation and analysis.

Healthcare has been a crucial research and application field of data analysis and visualisation for several decades [Reddy 2015]. In the previous work, much of the focus have been on the visualisation of electronic health records (EHRs). Rind et al. gives a detailed review of the related work [Rind 2011], categorising by individual patients or a group of patients. In each category, the work is further divided by visual analytics of time series data or status at a certain time point. West et al. also presents a systematic review of visual analytics approaches that have been proposed to illustrate EHR data [West 2015].

As mentioned in the last subsection, Lifelines [Plaisant 1996], a pioneer work in visualisation of individual patient records, provides a general visualisation environment for problems, diagnoses, test results or medications using timelines. Lifelines2 [Wang 2009] moves further by providing visualisation of temporal categorical data across multiple records, which is better for a doctor to

view to discover and explore patterns across these records to support hypothesis generation, and find causal relationships in a population.

VISITORS [Klimov 2010], which is based on KNAVE [Shahar 1999] and KNAVE II [Shahar 2006], uses aggregation to extract meaningful interpretations from multiple patients' raw time-oriented data. PatternFinder [Fails 2006] provides tools for the user to query patterns by specifying the attributes of events and time spans.

LifeFlow [Wongsuphasawat 2011] and EventFlow [Monroe 2013] are tools for event sequence analytics for a group of patients. They extract and highlight the common event sequence from patient records. Outflow [Wongsuphasawat 2012] and DecisionFlow [Gotz 2014] uses Sankey diagram [Riehmann 2005] style visualization to help visualise and analyse the causal relationships of events in complex event sequences.

The existing work has made detailed visualisation research mostly pertaining to EHRs and event analysis. Other types of work are still relatively limited due to their complexity and dependence on medical expertise. In the next section, the existing work on cohort visualisation is introduced.

3.4. Cohort Visualisation

While visualisation of a single cohort can be based on existing visualisation techniques, especially time-varying data visualisation, the challenge exists in visualisation of cohort study or visual comparison of different cohorts.

Gleicher et al. [Gleicher 2011] provide an extensive survey of visual comparison techniques classified into three categories: juxtaposition, superposition, and explicit encoding.

Several works pay attention to event sequence comparison. MatrixWave [Zhao 2015] is a visualization tool designed to compare the flow of users in click-stream datasets. It focuses on differences in the occurrence of pairwise steps in the event stream. Vrotsou et al. [Vrotsou 2014] introduce a set of event sequence similarity measures and cluster similar groups of event sequences.

On cohort study, there are tools proposed to combine visualization and statistics for medical cohort selection. CAVA [Zhang 2014], as shown in Figure 11, is a visualization tool for interactively refining cohorts and performing statistics, but limited to a single group. It focus on combining visualization with automated statistics and providing an interactive interface for selecting cohorts.

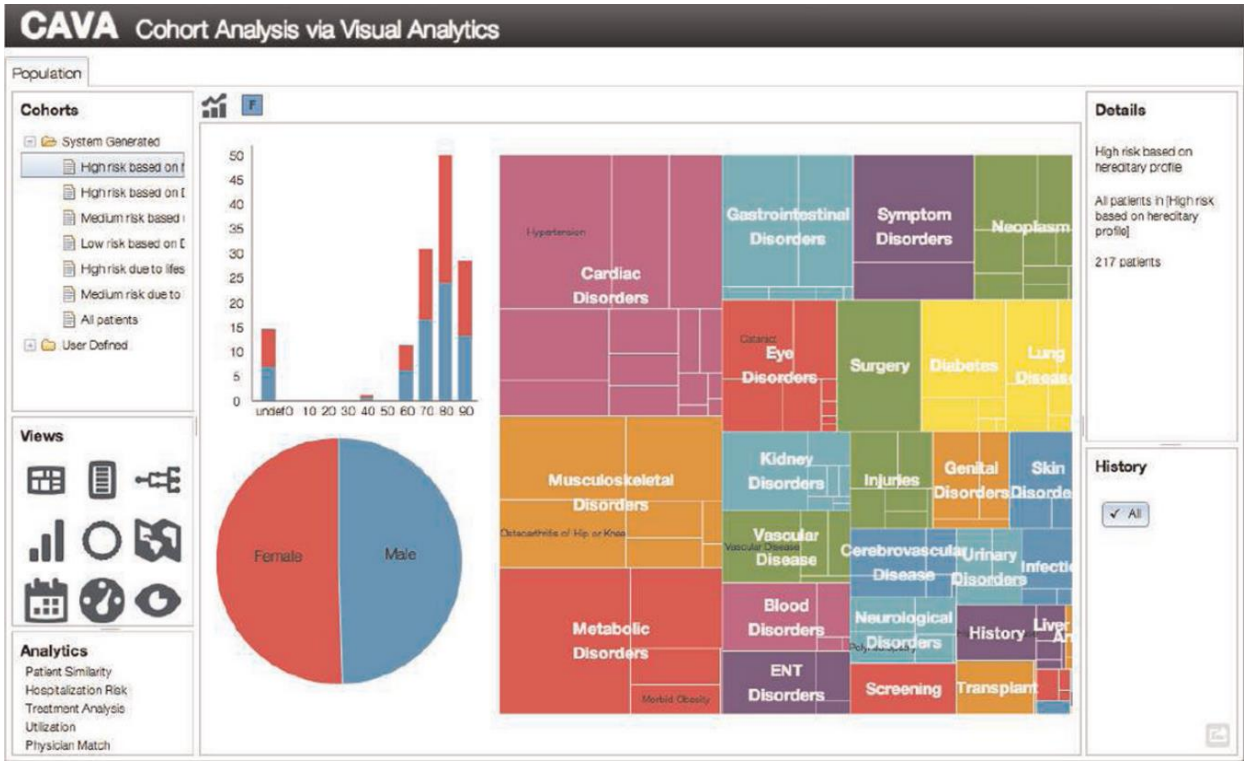


Figure 11 CAVA Cohort Visualisation

CoCo [Malik 2016b] focuses on cohort event sequence comparison. It generalizes the comparison to differences in single events and sequences of any length, as well as differences dealing with time, as shown in Figure 12.

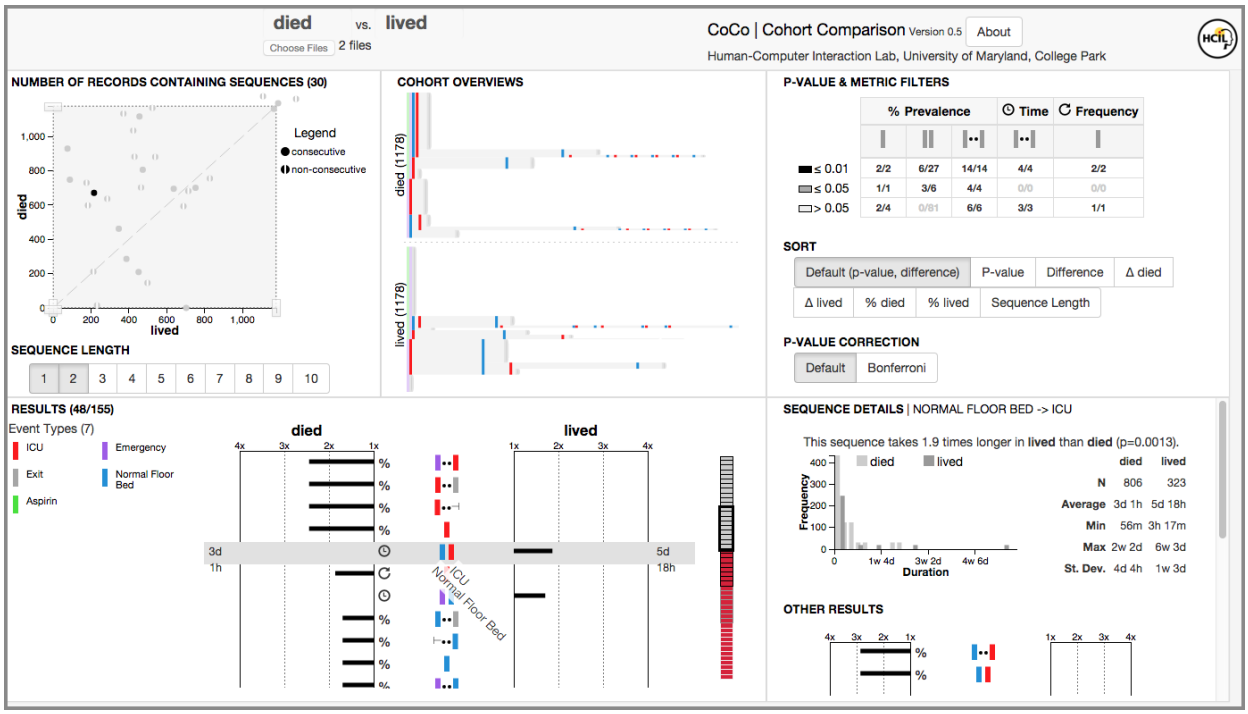


Figure 12 CoCo : A Visual Analytics Tool for Comparing Cohorts of Event Sequences

In medical cohort studies, the Kaplan-Meier method [Kaplan 1958] is often used to analyze the survival time of patients on different treatments and to compare their risks of death [Bewick 2004]. In the analysis of the cohort comparison, hypothesis tests are often used to test the validity of a claim that is made about a cohort. In a hypothesis test in statistics, the p value which is the probability of observing a difference between the exposed and unexposed groups is used to determine the significance of the results. In risk analysis cohort study, the confidence interval provides the range of risk ratios at a given probability, such as 95%, it is viewed as a measure of the precision of the estimated risk ratio. [Jamsen16] presents a detailed review of statistical methods and models for cohort study.

Existing work on visual cohort study and cohort comparison are still limited. Only a few works have been done to propose visual methods for cohort selection with statistics visualisation. The existing work on cohort comparison focus more on event comparison, which is not the focus in iManageCancer.

3.5. Scalable Visualisation

3.5.1. Introduction

Dealing with a huge amount of patient data in iManageCancer, the visualisation needs to be scalable. The focus of this section is about scalable visual analysis. The term "scalable visualisation" generally implies good scalability of visualisation. Scalable visual analysis supports visualisation, analysis of and interactions with large scale data and big data. Domains such as medical visualisation, architecture and urban design, geospatial scanning, astrophysics, biochemistry and abstract data analysis are regularly producing massive datasets containing features that are many orders of magnitude apart in scale. Big data is a term in the internet era for large and complex data sets, especially at internet scale, with the properties of increasing volume, velocity and variety, that it becomes difficult to process them using traditional data processing applications. Without special technologies, the visualisation and analysis will not work on such kind of large datasets.

3.5.2. Supporting Techniques for Scalable Visualisation

A wide variety of techniques have been researched, proposed and developed for scalable visualisation. We propose that almost all of these techniques be grouped under the following categories, according to the function which they perform in the visualisation pipeline:

• Scalable data management

• Scalable user interaction techniques

• Scalable data representation

•Temporal scalability.

Scalable Data Management

Large datasets might neither fit into the memory of a typical PC nor a typical GPU. In this section, we shall focus on describing approaches for visualising datasets larger than the main memory available. A review of massive data sets and their visualisation can be found in [Joy 2009].

Data Partitioning

While multidimensional arrays are usually stored in a file using linear storage, a common way of improving access to them is to reorganise the file using partitioned storage following the philosophy of divide and conquer. That happens when data is split in chunks (cubes or bricks) of equal size and the same dimensionality as the original volume and each partitioned chunk is stored contiguously in the file using linear storage. This increases the likelihood that data that are physically close in the n-dimensional volume will be stored at locations that are close together in the file.

Dynamic chunking views a dataset stored using linear storage as if it were chunked into blocks of configurable size and shape. As soon as an item in any block is accessed, the entire block is read. Dynamic partitioning was described in the context of a slicer visualisation application [Lipsa 2011] and it was shown that it provides some of the benefits of file chunking without having to reorganise or maintain multiple copies of the file.

Data Prefetching

Prefetching has also been an active area of research for handling large data. Common ways to mediate the effect of slow I/O are to use prefetching or to use a separate thread to overlap rendering with data I/O.

Scalable database

Dynamic scalability as one of the core concerns of big data visualisation has proven to be a particularly essential problem for databases [Pokorny 2013]. Scalable databases provide better scalability on handling a large amount of data. Traditional database technology relies on vertical scaling by installing new big servers. Unfortunately, this approach requires higher level of skills and the redistribution data on the fly can cause decreasing system performance. NoSQL (Not Only SQL) database is an important move to scalable data management with much better horizontal scalability. New capacity to these systems is simply expanded by adding new machines. The database system itself takes care of rebalancing the data and ensuring that it is sufficiently replicated across the cluster. They can almost provide linear horizontal scalability without compromising performance. Relatively low-cost NoSQL databases also have other features in addition to horizontal scalability, such as support for weaker consistency models, more flexible data models, and support for simple low-level query interfaces.

Commonly used NoSQL databases include Apache Cassandra [Cassandra], MongoDB [MongoDB], Apache CouchDB [CouchDB], etc.

Scalable Data Analysis

Filtering

Data sources typically contain large amount of data which might be too big for visualisation processing. However, visualisation usually needs only a specific subset of data that meets certain criteria. Users can select specific data by using filters. For example, rather than retrieving information about the full time range, the user can create filters to select data in a given date range. This is almost indispensable for interactive scalable visualisation of large amount of data as it is usually not possible to load the whole dataset into the visualisation application. The filter can also be a spatial one to select events in a given space area, such as search of objects near a geographical location.

Dimensionality reduction

In machine learning and statistics, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration and can be divided into feature selection and feature extraction [Pudil 1998]. Dimensionality reduction can also be seen as the process of deriving a set of degrees of freedom which can be used to reproduce most of the variability of a data set. Commonly used dimensionality reduction techniques are Principal Component Analysis (PCA). Multidimensional scaling (MDS), Isomap, Maximum Variance Unfolding (MVU), Kernel PCA, etc.

Clustering

Clustering is a fundamental technique in data mining and visual analysis. Clustering uses unsupervised learning to find structures in a collection of unlabelled data. It organises objects into groups based on similarity of objects [Rokach 2010]. In this way, the number of objects to be visualised is greatly reduced and part of the underlying knowledge in the objects is shown to the user. Hierarchies of objects can be built and used for interactive exploration and analysis. Common techniques of data clustering include k-means, hierarchical clustering, Fuzzy C-means, Gaussian Mixture Model (GMM), etc.

Scalable Data Presentation

Level-of-detail

Scalable data representation needs to consider which Level of Details (LoD) is to be rendered at a particular scale. In its simplest form, LoD might simply consist of modifying the resolution of an object that are too small to be resolved. More complex examples can be found in fields such as cartography, in which a large body of algorithms exists describing how features should be added, deleted and redrawn at different scales in order to preserve properties such as connectivity and the semantic requirements.

Call-outs

In visualisation and illustration, a call-out is an annotation that is associated with a point. This could be a text label connected to an object, a flag, or even a speech balloon. Call-outs can also be used interactively to view the sub-scale detail and global context at the same time, which is the "Focus + context" visual analysis paradigm desires.

Bundling

Long thin objects in the form of collections of fibres are a special case in scalable visualisation. They occur in InfoVis applications such as diagrams of interconnected items and densely connected edge graphs as well as vector field visualisations. The problem of visualising a large number of fibrous connections is so important that it has its own technique, called bundling.

Edge bundles render large graphs via edge clustering, by collecting together long edges analogous to the way electric wires are merged into bundles along a shared mutual path segment, fanning out at ends to connect distinct endpoints. Hierarchical edge bundling [Holten 2006] was introduced as a means to view a compound graph while reducing visual clutter. Edges are modelled as B-splines, with those following a similar path to one another within the hierarchy being grouped together for relevant subsections of their paths. Those bundles containing a greater number of edges are rendered as being brighter. [Balzer 2007] used edge bundles to simplify edges in a clustered level-of-detail graph visualisation that filtered the layout of the original graph. In [Holten 2009] a force-directed technique is proposed that uses a self-organising approach to bundling, in which edges are modelled as flexible springs that can attract each other. Note that

edge bundling determines both edge grouping and the paths of the edges. Streamline bundles only determine grouping. Meanwhile, streamline bundles are organised hierarchically and can capture flow features of varying scales at different levels of detail.

User Interactions for Scalable Visualisation

Fly-through

Two common modes in which a user interacts with the graphical display of a spatial scene are scene-in-hand and fly-through. Scene-in-hand is widely used when interacting with graphical objects on a desktop display: the user visualises the object from the outside and manipulates it using the mouse. Fly-through interaction is used in the street level visualisation of Google maps [GoogleMap]: the metaphor is that of the user being immersed in the scene and walking or flying through it. The user might move about between pre-defined targets or have unconstrained movement in the form of flight controls. Fly-through interaction is well-suited for large, complex scenes, such as cities, where the scene is so much larger than the scale of the user that the scene-in-hand metaphor cannot be sustained.

Zooming

The most common approach to magnifying data is with a zoom interaction, usually initiated by a lick on the target. Sub-scale data which is too small to be resolved on the display screen is marked by a placeholder token, and clicking the mouse on the token invokes a zoom to the target data. Ideally, the zoom should be slow, allowing users to see what is happening and where they are heading. Thus the familiar click-and-zoom interaction can be regarded as a composite of two techniques: placeholder tokens to indicate the targets and a zoom to magnify them. A token can be treated as any glyph, landmark or label which acts as a placeholder for data in the scene at that point.

Lensing

A lens is a magnified region of an image or scene, located in the scene at the point being magnified. Lenses are normally used to inspect image data and are a useful tool for analysing small-scale phenomena within an enlarged visualisation. Lenses provide magnification of detail which is in-place and therefore retains the position and context in the global view.

A major advantage of lenses is that they can be employed to show different information in the lens regions, such as a parameter or scalar other than that currently being displayed in the background image and they are thus very useful in multi-parameter visualisations. The concept of the movable lens for exploring multi-parameter data was termed “see through interfaces” and the lenses referred to as “magic lenses” to distinguish them from conventional magnifying lenses. The magic lens might also be semi-transparent, acting as a small moveable overlay.

Focus + Context Visualisation

Focus+Context [Cockburn 2009] is a principle in Information Visualisation to display the most important data at the focal point at full size and detail, and display the area around the focal point (the context) to help make sense of how the important information relates to the entire data structure.

It starts from three premises: First, the user needs both overview (context) and detail information (focus) simultaneously. Second, information needed in the overview may be different from that needed in detail. Third, these two types of information can be combined within a single (dynamic) display.

Focus + context allows the viewer to inspect an interesting portion of the data in detail (the focus) without losing global context—the global view is preserved at reduced detail, highlighting the focused region.

Temporal Scalability

Data scale can be large in time as well as space, in that the data contains features of interest at a range of timescales. Time-varying data is very common in everyday life and in the field of visualisation, such as scientific simulation, health/medical data, social media data, etc. A temporal zoom expands the time-axis of the graph to show the activity on that scale, while the corresponding spatial animation was slowed down in order to be visible to the user.

Lenses are a popular technique for time-series data, since they retain the global context of the magnified region. Because time is one-dimensional, selecting the time of interest and the timescale is relatively simple compared with the spatial case. However, unlike spatial data, the sampling frequency of time data can be very high, with the scales of interest somewhere between the highest and lowest frequencies. Another important factor is that time data is often periodic, in which case it may be necessary to select and track small periodic or recurring features over longer timescales.

3.5.3. Conclusions

Scalable visualisation is widely used in cartography, astrophysics, information visualisation, etc. Scalable techniques are not visualisation styles in their own right, they provide additional navigation features which are integrated into a view after the visualisation style has been decided by the user.

There are many scalable techniques, most of which can be categorised into scalable data management, scalable data representation, scalable user interaction and temporal scalability. These functional groups form both a classification and a design menu for developers. Visualisation design consists of choosing a technique for each functional component, and depends on the data type, visualisation style, interaction style and various properties inherent in the data.

4. Data

4.1. Data Fields

The iManageCancer patient health data are collected, stored and retrieved by the data mining and analysis service reported in D8.1 “Implemented data analysis and data mining services”. The data format is customised to accommodate the requirements of iManageCancer. The data may contain the following fields:

```
{
	"Age":[
		{"name": "Age", "type" : "number"}
	],
	"Gender":[
		{"name": "Gender", "type" : "list", "values": ["Male", "Female"]}
	],
	"Race":[
		{"name": "Race", "type" : "list", "values": ["African",
"American", "Asian", "White"]}
```

```
	],
	"Language":[
		{"name": "Language", "type" : "list", "values": ["English",
"Icelandic", "Spanish"]}
	],

	"Problems":[
		{"name": "Title", "type" : "text"},
		{"name": "Onset", "type" : "datetime"},
		{"name": "Resolution", "type" : "datetime"},
		{"name": "Category" , "type" : "list", "values": ["Primary
Desease", "Co-morbidities"]}
	],

	"Procedures":[
		{"name": "Title", "type" : "text"},
		{"name": "Start Date", "type" : "datetime"},
		{"name": "End Date", "type" : "datetime"},
		{"name": "Institution" , "type" : "text"},
		{"name": "Location" , "type" : "text"}
	],

	"Allergies":[
		{"name": "Title", "type" : "text"},
		{"name": "Allergen Type", "type" : "list", "values": ["Food
Allergy", "Environmental Allergy", "Drug Allergy", "Drug intolerance", "Food
intolerance"]},
		{"name": "Drug Class Allergen", "type" : "text"},
		{"name": "Severity", "type" : "list", "values": ["Mild",
"Moderate", "Severe", "Life Threatening", "Fatal"]}
	],

	"Measurements":[
		{"name": "Title", "type" : "text"},
		{"name": "Value", "type" : "number"},
		{"name": "Unit", "type" : "text"},
		{"name": "Type", "type" : "list", "values": ["Weight", "Systol",
"Diastole", "Pulse", "Body Temperature"]}
	],

	"Laboratory Results":[
		{"name": "Lab Name", "type" : "list", "values": ["CBC: WHITE
BLOOD CELL COUNT", "CBC: RED BLOOD CELL COUNT", "CBC: HEMOGLOBIN", "CBC:
HEMATOCRIT", "CBC: MEAN CORPUSCULAR VOLUME", "CBC: MCH", "CBC: MCHC", "CBC:
RDW", "CBC: PLATELET COUNT", "CBC: ABSOLUTE NEUTROPHILS", "CBC: ABSOLUTE
LYMPHOCYTES", "CBC: NEUTROPHILS", "CBC: LYMPHOCYTES", "CBC: MONOCYTES", "CBC:
EOSINOPHILS", "CBC: BASOPHILS", "METABOLIC: SODIUM", "METABOLIC: POTASSIUM",
"METABOLIC: CHLORIDE", "METABOLIC: CARBON DIOXIDE", "METABOLIC: ANION GAP",
"METABOLIC: GLUCOSE", "METABOLIC: BUN", "METABOLIC: CREATININE", "METABOLIC:
TOTAL PROTEIN", "METABOLIC: ALBUMIN", "METABOLIC: CALCIUM", "METABOLIC: BILI
TOTAL", "METABOLIC: AST/SGOT", "METABOLIC: ALT/SGPT", "METABOLIC: ALK PHOS",
"URINALYSIS: SPECIFIC GRAVITY", "URINALYSIS: PH", "URINALYSIS: RED BLOOD
CELLS", "URINALYSIS: WHITE BLOOD CELLS"]},
		{"name": "Date", "type" : "datetime"},
		{"name": "Value", "type" : "number"},
		{"name": "Unit", "type" : "list", "values": ["no unit",
"rbc/hpf", "mg/dL", "pg", "m/cumm", "k/cumm", "gm/dL", "mmol/L", "U/L",
"g/dl", "fl", "%", "wbc/hpf"]}
	],

	"Psycho-Emotional":[
```

```
            {"name": "Date", "type" : "datetime"},
            {"name": "Received Health State", "type" : "number"},
            {"name": "Psychological Aspects", "type" : "number"},
            {"name": "Psycho-social Aspects", "type" : "number"},
            {"name": "Cognitive Aspects", "type" : "number"}
      ],

      "Family Resilience":[
            {"name": "Date", "type" : "datetime"},
            {"name": "Family comm. &amp;amp; problem solving ", "type" :
"number"},
            {"name": "Social &amp;amp; economic resources", "type" : "number"},
            {"name": "Maintaining a positive outlook", "type" : "number"},
            {"name": "Family connectedness", "type" : "number"},
            {"name": "Family spirituality", "type" : "number"},
            {"name": "Ability to make meaning from adversity", "type" :
"number"}
      ]
}
```

4.2. Data Query

The patient data can be queried by calling the iManageCancer data query API with a POST AJAX [AJAX] call.

```
http://139.91.210.63:8084/DataManagementAPI/s1/webservice/querybuilder/
```

The query is in JSON format and can be used to set query criteria and select user interested data columns. A sample Query is as follows:

```
{
    where:[
        {
            id:   "Age_1",
            name: "Age",
            featureData:{
                  "age": "45",
                  "condition" : "gt"
            }
        }
    ],
    select:{
        columns:  "Age,Laboratory Results,Gender"
    }
}
```

It tries to retrieve patients with age greater than 45. The returned data contains columns of age, gender and laboratory results.

For details of the data format and data queries, please refer to D8.1 “Implemented data analysis and data mining services”.

4.3. Sample Data

The following is a sample of the returned patient data.

```
{
```

```
"head": {
  "labdate": {
    "datatype": "http://www.w3.org/2001/XMLSchema#dateTime",
    "type": "typed-literal"
  },
  "gender": {
    "datatype": "nominal",
    "range": [
      "Female",
      "Male"
    ],
    "type": "literal"
  },
  "labvalue": {
    "datatype": "http://www.w3.org/2001/XMLSchema#integer",
    "type": "typed-literal"
  },
  "labname": {
    "datatype": "nominal",
    "range": [
      "CBC: WHITE BLOOD CELL COUNT",
      "CBC: RED BLOOD CELL COUNT",
      "CBC: HEMOGLOBIN",
      "CBC: HEMATOCRIT",
      "CBC: MEAN CORPUSCULAR VOLUME",
      "CBC: MCH",
      "CBC: MCHC",
      "CBC: RDW",
      "CBC: PLATELET COUNT",
      "CBC: ABSOLUTE NEUTROPHILS",
      "CBC: ABSOLUTE LYMPHOCYTES",
      "CBC: NEUTROPHILS",
      "CBC: LYMPHOCYTES",
      "CBC: MONOCYTES",
      "CBC: EOSINOPHILS",
      "CBC: BASOPHILS",
      "METABOLIC: SODIUM",
      "METABOLIC: POTASSIUM",
      "METABOLIC: CHLORIDE",
      "METABOLIC: CARBON DIOXIDE",
      "METABOLIC: ANION GAP",
      "METABOLIC: GLUCOSE",
      "METABOLIC: BUN",
      "METABOLIC: CREATININE",
      "METABOLIC: TOTAL PROTEIN",
      "METABOLIC: ALBUMIN",
      "METABOLIC: CALCIUM",
      "METABOLIC: BILI TOTAL",
      "METABOLIC: AST/SGOT",
      "METABOLIC: ALT/SGPT",
      "METABOLIC: ALK PHOS",
      "URINALYSIS: SPECIFIC GRAVITY",
      "URINALYSIS: PH",
      "URINALYSIS: RED BLOOD CELLS",
      "URINALYSIS: WHITE BLOOD CELLS"
    ],
    "type": "literal"
  },
  "labunits": {
    "datatype": "nominal",
    "range": [
```

```
        "rbc/hpf",
        "mg/dL",
        "pg",
        "m/cumm",
        "k/cumm",
        "no unit",
        "gm/dL",
        "mmol/L",
        "U/L",
        "g/dl",
        "fl",
        "%",
        "wbc/hpf"
      ],
      "type": "literal"
    },
    "age": {
      "datatype": "http://www.w3.org/2001/XMLSchema#integer",
      "type": "typed-literal"
    }
  },
  "results": {
    "bindings": [
      {
        "labdate": "1953-11-28T14:09:29.363",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "10",
        "labname": "CBC: HEMOGLOBIN",
        "labunits": "gm/dl",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T21:02:51.897",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "29",
        "labname": "METABOLIC: AST/SGOT",
        "labunits": "U/L",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T00:21:31.683",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "132",
        "labname": "METABOLIC: GLUCOSE",
        "labunits": "mg/dL",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T11:19:18.283",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "2",
        "labname": "URINALYSIS: RED BLOOD CELLS",
        "labunits": "rbc/hpf",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T00:20:01.843",
```

```
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "74",
        "labname": "METABOLIC: ALK PHOS",
        "labunits": "U/L",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T01:27:39.687",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "33",
        "labname": "CBC: MCHC",
        "labunits": "g/dl",
        "age": "87"
      },
      {
        "labdate": "1953-11-28T08:17:31.660",
        "gender": "Female",
        "patientid": "81C5B13B-F6B2-4E57-9593-6E7E4C13B2CE",
        "labvalue": "21",
        "labname": "METABOLIC: CARBON DIOXIDE",
        "labunits": "mmol/L",
        "age": "87"
      }

    ]
  }
}
```


5. Visualisation Techniques

In this section, visualisation techniques that are useful in cohort visualisation and analysis are introduced. These techniques are candidates for design and implementation of iManageCancer cohort visualisation and analysis.

5.1. Data Charts

Data charts are widely used in daily life, business and scientific research, etc. Common data charts include the line chart, dot chart, bar chart, pie chart and the scatter plot. The following is a list of simple introductions of these charts.

5.1.1. Bar Chart

A bar chart or bar graph is a chart or graph that presents grouped data with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one.

5.1.2. Line Chart

A line chart or line graph is a type of chart which displays information as a series of data points connected by straight line segments. It is a basic type of chart common in many fields. A line chart is often used to visualize a trend in data over intervals of time. In iManageCancer, both of the bar chart and the line chart can be used to visualise the time-dependant health data of patients.

5.1.3. Pie Chart

A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

Pie charts are very widely used in business. However, it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts. Pie charts can be replaced in most cases by other plots such as the bar chart.

5.1.4. Scatterplot

A scatterplot encodes two quantitative value variables using both the vertical and horizontal spatial position channels, and the mark type is necessarily a point. They are effective for the abstract tasks of providing overviews of datasets and are very effective for studying the correlation between two attributes. A scatter plot is also very useful when we wish to see nonlinear relationships between variables.

5.2. Calendar

Most of the medical and health data in iManageCancer come with timestamps, which means they have the date information. A natural form of date-based data organisation, display and editing is a calendar, which is a traditional way to visualise information that is associated with dates. A calendar can be used for health data visualisation.

5.3. Timeline

Time-dependant data are ubiquitous in many application domains, for example, in business, medicine, history, planning, or project management. In iManageCancer, most of data are time dependant. Providing appropriate methods to facilitate the visualisation and analysis of time-varying data is a key issue in iManageCancer. A timeline is a traditional method to visualise time-varying data and events in a linear layout which is suitable for visualisation of the trend of longitudinal data which may cover a relatively long period, such as health indicators and medical measurements.

The timeline can be used for visualisation and analysis of medical indicators. Data trends can be observed from the variable curves and data correlations may be discovered by comparison of the data curves of the multi-variables. As the data records may cover a long time range, interactive techniques such as zooming and overview+details are integrated with the visualisation.

5.4. Parallel Coordinates

The technique of parallel coordinates is an approach for visualising multiple quantitative variables using multiple axis which are placed parallel to each other in the most common case [Inselberg 1990]. The advantage of parallel coordinates is that it supports visualisation of multiple variables and correlation between attributes can be discovered by certain visualisation patterns. It is a common technique of visualising high-dimensional data and analysing multivariate data.

The parallel coordinates can be employed for multi-variable correlation analysis of the biomarkers. An example of the parallel coordinate view is shown in Figure 13 where negative correlations can be found between walking minutes and blood pressures as well as BMI (Body Mass Index).

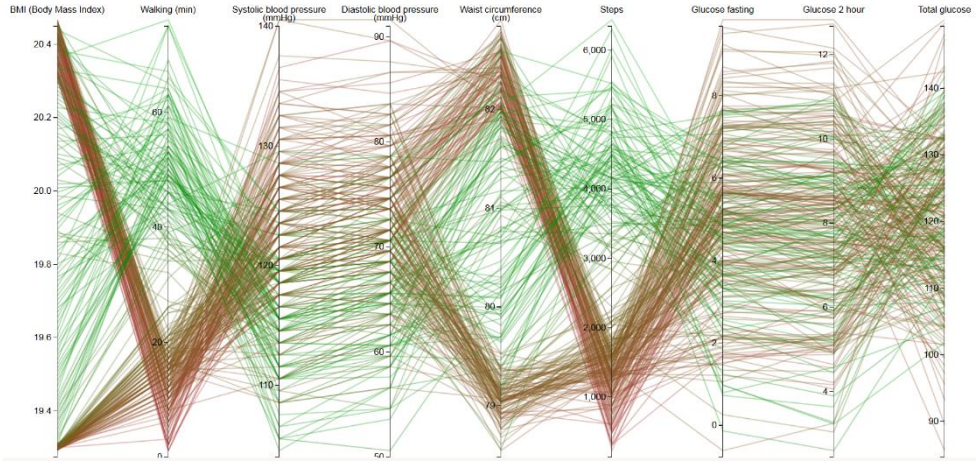


Figure 13 An example of Parallel Coordinates for health data visual analysis

User Interactions of parallel coordinates include axis reordering and brushing. The user can dragging an axis to a new position. In this way the user can put closely related axis in a close neighbourhood to better observe the data correlations. The user can also select a subrange on an axis to filter the data lines. The brush can be moved along the axis to dynamically select the data lines.

5.5. Sankey Diagrams

In medical cohort analysis, risks factors of diseases are commonly studied. A Sankey diagram [Riehmann 2005] can visualise the causal relationships of different risk transitions. As introduced in the related work, OutFlow [Wongsuphasawat 2012] and DecisionFlow [Gotz 2014] use Sankey diagram style visualisation to visually analyse the causal relationships of events. The advantage of the Sankey diagram is that it shows the multi-layer causal relationships of the elements in a clear and understandable way. Figure 14 shows a Sankey diagram visualisation of risk factors.

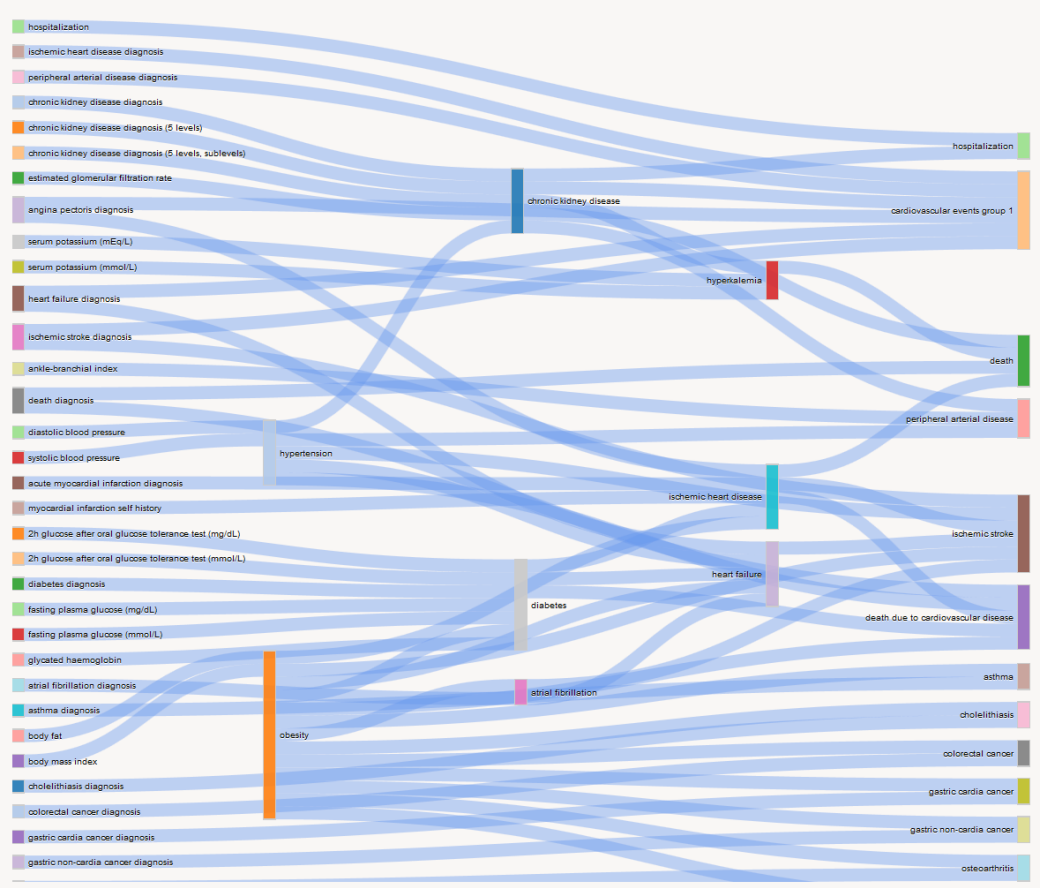


Figure 14 An example Sankey diagram of risk factors

6. iManageCancer Visual Analysis System

The iManageCancer Smart Analysis Framework (SAF) reported in D8.1 is composed of different web-based functional modules that communicate with each other. The visual analysis system is a web-based sub-system that provides supporting functions for data exchange, visualisation and analysis of patient data and cohort data from the data mining and analysis service. In this section, the design requirements and the general user interface of the visual analysis system are introduced.

6.1. iManageCancer Smart Analysis Framework Interface

The user interface of iManageCancer Smart Analysis Framework (SAF) is shown in Figure 15. The Smart Analysis Framework is available online at http://dimitra.no-ip.info/saf/index.php . It provides the query builder, the data analysis functions and the container for the visual analysis components. The details of the data analysis system is introduced in D8.1 "Implemented data analysis and data mining services". The iManageCancer Smart Analysis Framework provides the container for interactive health data visualization of individuals and cohorts.

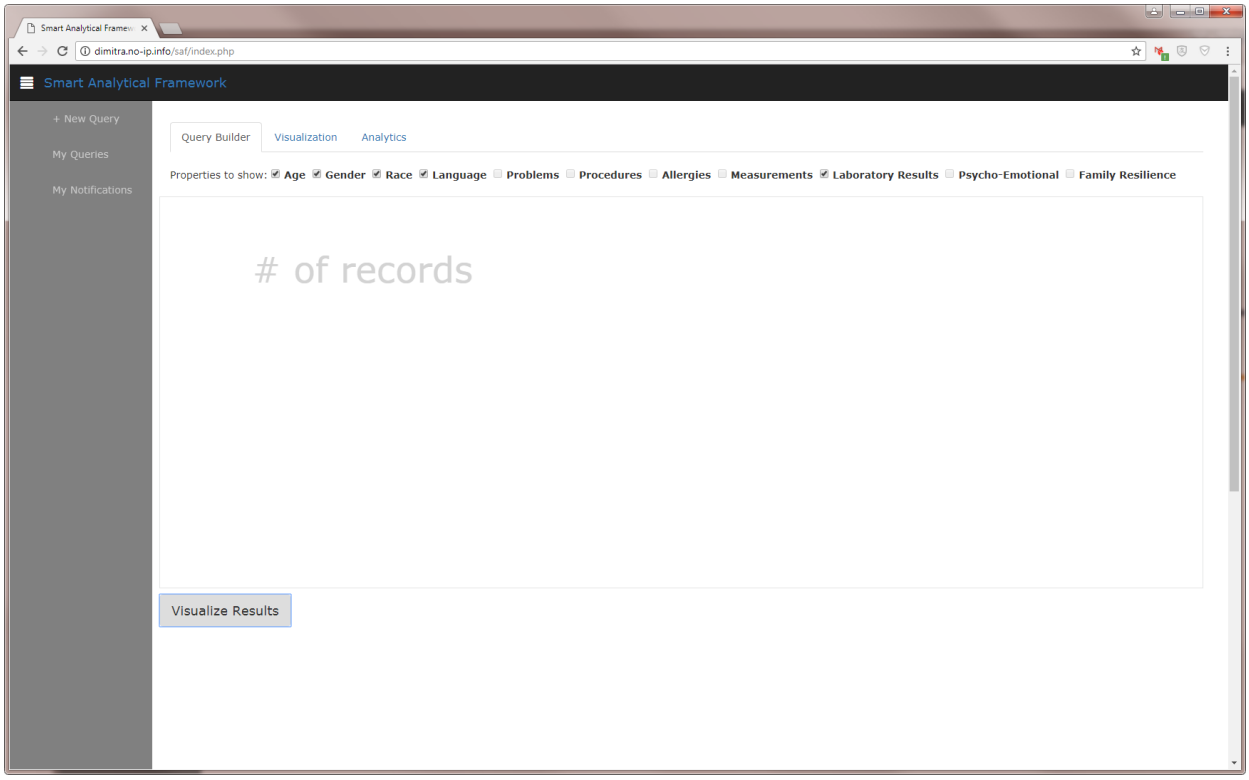


Figure 15 iManageCancer smart data analysis interface

6.2. Visualisation Design Requirements

In iManageCancer the data can be generally categorised as personal electronic health records (PHR) data and cohort data extracted from user defined cohorts. The role of visual analysis is to visualise and analyse health data as well as cohort data.

To gain intuitive knowledge of the health data and cohort data, visualisation in iManageCancer needs to provide users with the ability to view, understand, compare and interact with the health data and cohort data. The visualization design requirements of iManageCancer include:

- Visualisation of individual's medical and health data, including PHR data and other types of patient data, to help users to understand the data;
- Visualisation of cohort data, including PHR data and other types of patient data, to help users to understand the cohort data;
- Visualisation of data clusters from the data analysis service;
- Visual comparison of cohort data allowing for analytical analysis of cohort differences to help understanding of the relations between the biometric markers and the health outcomes.

6.3. Single Patient Visualisation

6.3.1. Visual Interface

Components

Variable Selector

In health and medical applications, a patient is naturally associated with multiple biometric markers and health indicators such as weight, blood pressure, glucose, etc. However, different patients may have different biometric markers and health indicators in concern. The visualisation needs to provide the end users abilities to select and visualise variables from the patient's available variables.

A variable selector provides a column of available variables from the patient. The variables are displayed in groups if group information is available. The users can use the variable selector to select desired variables to visualise.

Timeline

Medical measurements of patients are normally time dependent. In addition, longitudinal studies are highly important in medical cohort studies. All these implies that visualisation of time-varying data is indispensable in patient visualisation as well in cohort visual analysis. Temporal data can be visualised in a linear form or a radial form. A linear timeline visualisation is more intuitive and has been used by many of the previous works. Moreover, linear timelines make it easier to compare different data variables. In addition, it is also better to achieve a compact layout.

In the timeline for each variable or variable group, data trends can be observed from the variable curves and data correlations may be discovered by comparison of the data curves of the multi-variables.

Layout

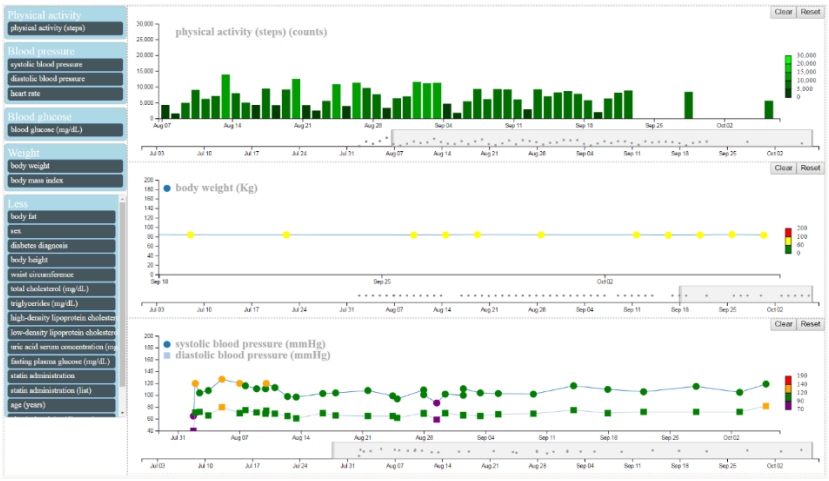


Figure 16 iManageCancer Healthline component

In iManageCancer, the variable selector and the timelines are organised as a compound component – the Healthline – which is an enhanced timeline group. The Healthline is used to visualise the longitudinal data of multiple biomedical markers. It combines the timeline and the

variable selector to visualise multiple personal health variables. The visual interface of a healthline is composed of the variable selector and a timeline group panel, as shown in Figure 16.

Interactions

Variable selection by Drag and Drop

In the Healthline, there are multiple timelines and a variable list. A solution needs to be provided to select a variable or a group of variables to be visualised by a dedicated timeline. In iManageCancer Healthline, an intuitive Drag and Drop method is designed and implemented. The variable items in the variable list can be directly dragged to the desired timeline and then dropped. The timeline accepts the variable and adds it to its visualisation. Multiple variables may be visualised in the same timeline provides that they have the same units, such as the systolic blood pressure and the diastolic blood pressure. This user interaction is intuitive and natural, and fits better on mobile applications as well.

Overview+Details visualisation of details

Longitudinal data may cover a long period and have different level of details at different time. The visualisation needs to be capable of visualising and revealing the data with different time granularities. Common techniques are zooming, overview+details and focus+context [Cockburn 2009] as introduced in the related work. In the Healthline timeline, zooming and overview+details are employed to provide the user data details with the general trend of the data at the same time. Each timeline is composed of the top detail panel and the bottom overview panel. A brushing tool is provided in the overview panel to select the desired data period. When the user moves the brush, the content in the detail panel changes accordingly. In addition, zooming is supported in the detail panel and the brush in the overview panel is automatically synchronised.

6.4. Cohort Visual Analysis

6.4.1. Cohort Visual Analysis Requirements

In iManageCancer, the cohort data analysis and visualisation help the study of the status and treatments of cancer patients. In this section the requirements of cohort visual analysis are introduced and discussed.

Cohort definition

In statistics and demography, a cohort is a group of subjects selected by dedicated criteria for observation, data recording and analysis.

In medical practice, a cohort might be a group of patients who share a defining characteristic (e.g. similar health conditions and experienced a common treatment in a selected time period). Cohort comparisons and studies are commonly used in medical practice to study risk factors and the efficacy of treatments and medicines. A cohort study is normally a longitudinal study that examines a cohort at intervals through time and is widely used in medicine, demography, psychology, social science, business analytics, and ecology, etc.

Traditionally, the medical cohort study is more in the statistical sense and many of the cohort details cannot be revealed due to the large amount of longitudinal data. In iManageCancer, the visual analysis method is incorporated to meet the challenge of cohort study. Visual analysis is

highly important for the analysis and comparison of the status of patients in the iManageCancer system.

In the following subsections the requirements of iManageCancer cohort visual analysis are introduced.

Cohort Summary

As a cohort comprises a number of subjects (patients), the users need to know the general properties of the group, which is normally statistical summaries of the group. For example, number of patients, gender, age distribution, average values of medical biometrics, etc. The summary information is very important for the users to discover the main distinctive properties of the cohort and to gain insights easier in the process of cohort comparison and visual analysis.

Cohort Details

In cohort analysis, cohort summary describes the general characteristics of the cohort, but it is far from revealing the details of subjects in the cohort. Cohort detail visual analysis allows the users to have a close view of a dedicated subject or subject group in the selected time range. Without the details of cohort members, it is not possible to provide the users the detailed information of the cohort.

Cohort Visualisation

A cohort is associated with a large volume of data which may be heterogeneous and time-varying. Without appropriate visualisation it is not possible to present, utilise, compare and understand the massive data. Visualisation is capable of convert the huge amount of data into graphical elements that can be perceived and processed in a massive parallel mode by human brains as visual signals. Insights can be gained by interacting and analysing the visualisation, which is normally very difficult or impossible with the raw data.

Clustering Result Visualisation

The SAF data analysis service provides k-means clustering of the cohort data to find the "similar" cases/records. It is required that the results of the clustering be presented in a graphical form that can be intuitively recognised.

Feature Selection Visualisation

To reduce variable dimensions and find the most important variables, SAF provides service to mine the most informative variables of a cohort to the end user. The results contain list of features with a weight that represents how informative is each feature. The information needs to be visually presented.

User Interaction

Visual analytics is a process with human in loop. The iterative process of data mining -> visualisation -> end users implies that the user interaction is a key phase in the whole visual analytics loop. The user interactions change the visualisation and provide multiple aspects of the data. With effective visualisation, the users can narrow down the data search and highlight the attributes for their specific analysis purpose. Through iterative user interactions and visualisation, insights may be gained from the process.

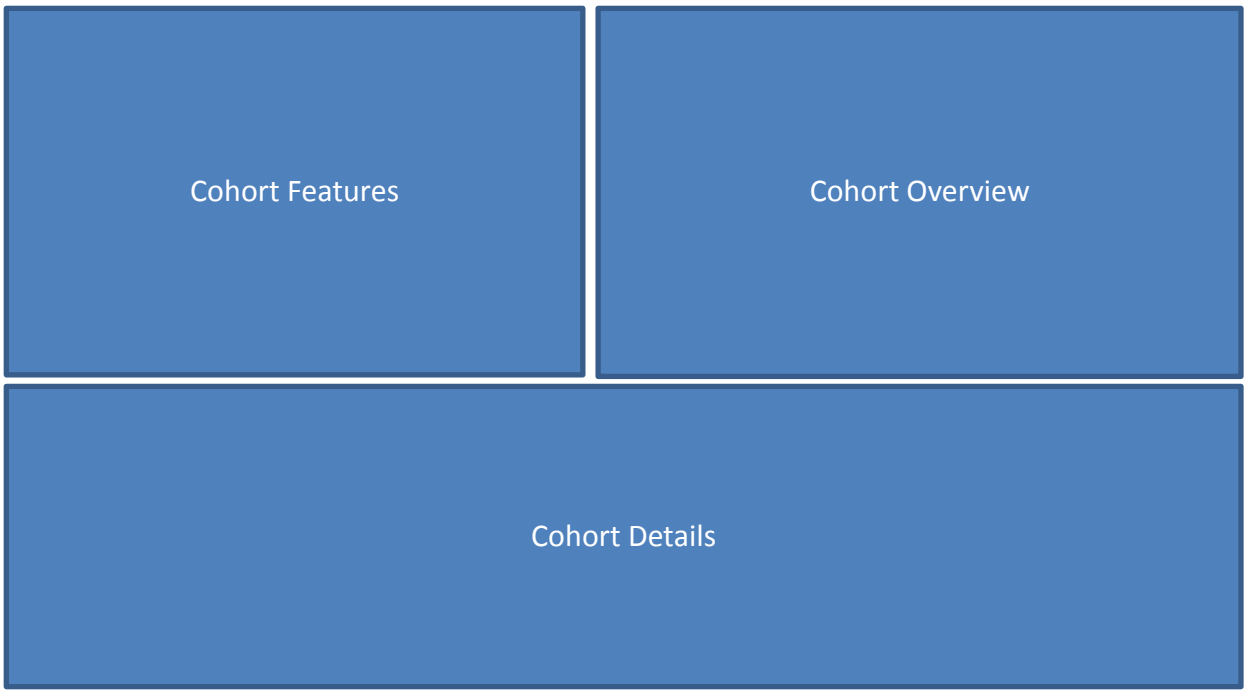
Cohort Comparison

In cohort study, cohort comparison is a critical method which reveals the common and different properties of different cohorts, thus leading to the discovery of knowledge and insights. For example, cohort comparison based study is widely used in statistical hypothesis tests.

Without visual analysis, cohort comparison is performed more on the statistical level of the certain key factors such as survival rate, survival time, etc. With interactive visual analysis, more capable comparison of the cohorts at a finer level can be achieved. For example, key properties of cohorts can be directly compared via interactive visualisation.

6.4.2. Visual Interface Layout

As introduced in Section 6.4.1, the required functions of iManageCancer cohort visual analysis include cohort definition, cohort summary visualisation, cohort detail visualisation and cohort comparison. Accordingly, in the design of the visual interface layout of our iManageCancer cohort visual analysis, we include the cohort feature panel, cohort overview panel and the cohort detail panel (as shown in Figure 17) which are introduced in detail in Section 6.4.3.



6.4.3. Visual Interface Components

As introduced in Section 6.4.2, the cohort visual analysis user interface include a cohort feature panel, a cohort overview panel and a cohort detail panel. In each functional panel, multiple choices may exist for the selection of the corresponding component. The components in the current implementation are introduced in the following subsections.

Figure 17 iManageCancer Cohort visual analysis interface layout design

Cohort Feature Panel

A direct implementation of the cohort filter can be a list of range sliders to select the range of health indicators to define the cohort. However, sliders are weak at visual representation of the data. Our solution is to use parallel coordinates as a tool to define cohorts. The axes of parallel coordinates can achieve the same functions as range sliders. Meanwhile, parallel coordinates itself can dynamically visualise the selected cohorts and show patterns of these cohorts. This helps the user to define and adjust cohorts interactively.

If the cluster information is available, variable axes of different clusters will be displayed in different colours.

If the information of important feature is available, the corresponding axes will be highlighted and visualised in the thickness representing their weight.

In the user interaction, the user selects interested variables and then brushes on the corresponding axis of the parallel coordinates to specify the criteria of data queries. The range values are automatically converted to the data queries and sent to the data server.

Cohort Summary Panel

The cohort summary panel employs data charts for cohort summary visualisation. The screenshot in Figure 18 shows a bar chart visualisation for patients in a cohort. Other types of data charts can also be used, depending on the application requirements. For example, the cluster information can be visualised in different colours in the scatterplot.

Cohort Detail Panel

The cohort detail panel is based on the Healthline component. However, the Healthline component initially designed for single patient visualisation has been extended to display data from multiple patients to accommodate cohort visualisation and cohort comparison, as shown in Figure 18.

For single cohort visualisation, the user can click on the legend of patients to highlight the data of specific patients. For cohort comparison, data in different cohorts are shown in different colour or legends. The statistical distribution of the cohort data can be visualised in the Healthline as well. Brushing on the overview panel naturally supports overview+details visualisation of the data within the selected period. Dragging an item from the left variable list into the timeline allows visualisation of the selected biomarker.

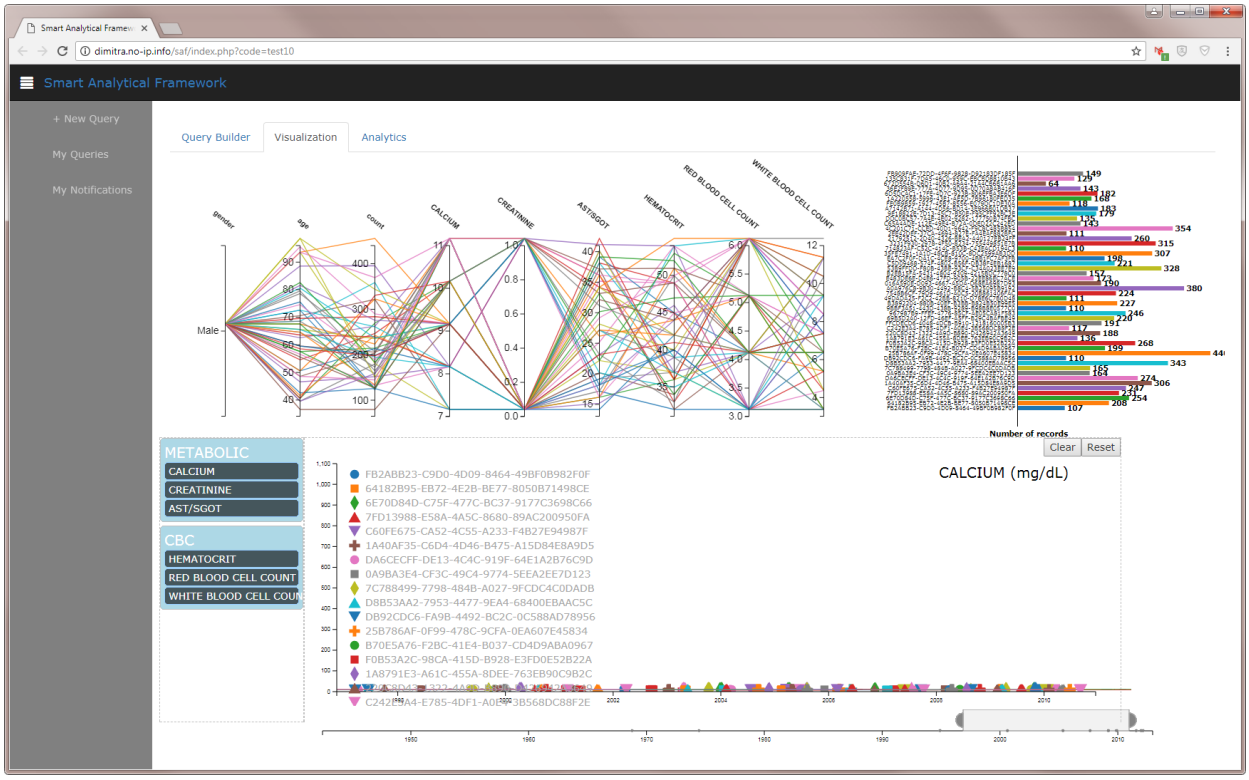


Figure 18 iManageCancer visual analysis user interface

7. Implementation

The web-based visual analysis framework is developed in Javascript, PHP and HTML and deployed on a Linux environment. It uses several open source library, including D3.js [d3] for interactive web-based visualisation, jQuery [jQuery] for front end javascript programming. The backend is based on PHP programming which is a popular language for server side web programming and runs on all major platforms include Windows, Linux and Mac OS. The frontend web-based UI is mainly based on HTML and jQuery [jQuery] which facilitates javascript programming. The interactive visual analysis components are implemented in javascript and the javascript based scalar vector graphics (SVG) library d3.js [d3]. Each component is placed in a DIV element in the component panel and the whole visualisation interface is managed automatically as tabs in the main SAF user interface.

The health data of patients are fetched from the FORTH server via the dedicated query APIs. jQuery is used to help shape the frontend logic, and interact with the REST service provided by backend. The project is hosted on FORTH server and is available at http://dimitra.no-ip.info/saf/index.php.

8. Conclusion

Visual analytics is an integral approach with visualisation, human factors, and data analysis involved. This process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data.

The target of WP8 in iManageCancer project is to provide effective visual analysis tools to empower users to access, view, compare and understand the health data of patients and cohorts.

In this deliverable report of D8.2, the design and implementation of Task 8.2 "Visualisation" is presented. It is based on the work reported in D8.1: "Implemented data analysis and data mining services". Visualisation components including charts, timeline, parallel coordinates are used to select and visualise the health data of patients and cohorts.

A web-based iManageCancer Smart Analysis Framework is provided as the hosting system of the visual analysis components. A visual analysis interface is designed for cohort visualisation and analysis, which includes the definition, overview and detail of the cohort.

The design and implementation meets the demands of visual analysis of health and medical data of a single patient as well as cohort study and analysis. Cohorts can be interactively defined on the visual interface and data can be queries from the server accordingly. The overview of the cohort is visualised and the details of the cohort can be interactively visualised and studied.

9. References

[Aigner 2008] W Aigner, S Miksch, W Muller, H Schumann, C Tominski, Visual Methods for Analyzing Time-Oriented Data, IEEE TVCG Vol.14(1), 2008

[Aigner 2011] W Aigner, S Miksch, H Schumann, C Tominski, Visualisation of Time-Oriented Data, Springer, 2011

[Rind 2011] A Rind, TD Wang, W Aigner, S Miksch, K Wongsuphasawat, C Plaisant, B Shneiderman, Interactive information visualization to explore and query electronic health records. Foundations and Trends in HCI 5(3):207–298, 2011

[AJAX] AJAX, https://www.w3schools.com/xml/ajax_intro.asp

[Balzer 2007] M Balzer, O Deussen, Level-of-detail visualisation of clustered graph layouts, Asia-Pacific Symposium on Visualisation (APVIS), 2007

[Bewick 2004] V Bewick, L Cheek, J Ball. Statistics review 12: survival analysis. Critical Care, 8(5):389-94, 2004

[Borgo 2013] R Borgo, J Kehrer, DHS Chung, E Maguire, RS Laramee, H Hauser, M Ward and M Chen, Glyph-based Visualisation: Foundations, Design Guidelines, Techniques and Applications, Eurographics State of the Art Report, pp. 39-63, 2013

[Calendar Heatmap] Calendar Heatmap, http://kamisama.github.io/cal-heatmap/

[Cassandra] Apache Cassandra, http://cassandra.apache.org

[CouchDB] CouchDB, http://couchdb.apache.org/

[Cockburn 2009] A Cockburn, A Karlson, BB Bederson, A review of overview+detail, zooming, and focus+context interfaces. ACM Computing Surveys (CSUR) Surveys 41(1), 2009

[d3] d3.js, http://d3js.org/

[Fails 2006] JA Fails, AK Karlson, L Shahamat , et al., A visual interface for multivariate temporal data: Finding patterns of events across multiple histories, in Wong, P.C., Keim, D.A. (Eds.) 'IEEE VAST' (IEEE Computer Society, 2006) pp 167–174, 2006

[GapMinder] GapMinder, http://www.gapminder.org

[Gleicher 2011] M Gleicher, D Albers, R Walker, I Jusufi, CD Hansen, and JC Roberts. Visual comparison for information visualization. Information Visualization, 10(4):289-309, September 2011

[GoogleMap] Google Maps API Family, http://www.google.com/apis/maps

[Gotz 2014] D Gotz, H Stavropoulos, Decisionflow: Visual analytics for high-dimensional temporal event sequence data. IEEE Transactions on Visualization and Computer Graphics 20(12), 2014

[Groves 2013] P Groves, B Kayylai, D Knott, S Van Kuiken, The big-data revolution in us health care: Accelerating value and innovation, Centre for US Health System Reform; Business Technology Office, 2013

[Havre 2000] S Havre, B Hetzler, L Nowell, ThemeRiver: visualizing theme changes over time, Proc. IEEE Symposium on Information Visualisation, 2000

[Holten 2006] D Holten, Hierarchical Edge Bundles: Visualisation of Adjacency Relations in Hierarchical Data, IEEE TVCG, Volume 12 Issue 5, Pages 741-748, September 2006

[Holten 2009] D Holten, Jarke J. van Wijk, A user study on visualizing directed edges in graphs, Proceeding CHI '09, Pages 2299-2308, 2009

[Inselberg 1990] A Inselberg, B Dimsdale, Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proc. the 1st IEEE Symposium on Visualization. pp.361–378,1990

[Jamsen 2016] KM Jamsen, J Ilomäki, SN Hilmer, N Jokanovic, EC Tan, JS Bell, A systematic review of the statistical methods in prospective cohort studies investigating the effect of medications on cognition in older people, Res Social Adm Pharm., Jan-Feb;12(1):20-8, 2016

[Joy 2009] KI Joy, Massive data visualisation: a survey, In Mathematical Foundations of Scientific Visualisation, Computer Graphics, and Massive Data Exploration, pp 285-302, Springer, 2009

[jQuery] jQuery, https://jquery.com/

[Kaplan 1958] EL Kaplan, P Meier, Nonparametric estimation from incomplete observations, Journal of the American Statistical Association, 53 (282): 457–481, 1958

[Keim 2010] D Keim, J Kohlhammer, G Ellis, et al. Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association, 2010

[Kincaid 2010] R Kincaid, SignalLens: Focus+Context Applied to Electronic Time Series, IEEE Transactions on Visualization and Computer Graphics (TVCG), 16(6), 2010

[Klemm 2014] P Klemm, S Oeltze-Jafra, K Lawonn, K Hegenscheid, H Völzke, B Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. IEEE Transactions on Visualization and Computer Graphics (TVCG), pp. 1673-1682, 2014.

[Klimov 2010] D Klimov, Y Shahar, M Taieb-Maimon, Intelligent visualization and exploration of time-oriented data of multiple patients, Artificial Intelligence in Medicine, 49(1):11–31, 2010

[Lipsa 2011] DR Lipsa, RS Laramee, RD Bergeron, TM Sparr, Techniques for large data visualisation, Int. J. of Research and Reviews in Computer Science 2(2):315-322, 2011

[Malik 2015] S Malik, F Du, M Monroe, E Onukwugha, C Plaisant, B Shneiderman, Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics, in Proceedings of ACM Intelligent User Interfaces (IUI) 2015. Atlanta, GA, USA. pp. 38-49, 2015

[Malik 2016a] S Malik, B Shneiderman, F Du, C Plaisant, M Bjarnadottir, High-Volume Hypothesis Testing: Systematic Exploration of Event Sequence Comparisons, ACM Transactions on Interactive Intelligent Systems (TiiS), 6(1), May 2016

[Malik 2016b] S Malik, A Visual Analytics Approach to Comparing Cohorts of Event Sequences, PhD Disseration, University of Maryland, 2016, http://drum.lib.umd.edu/handle/1903/18585

[MongoDB] MongoDB , http://www.mongodb.org

[Monroe 2013] M Monroe, R Lan, C Plaisant, B Shneiderman, Temporal Event Sequence Simplification, TVCG: IEEE Transactions on Visualization and Computer Graphic, 2013.

[Munzner 2014] T Munzner, Visualization Analysis and Design, AK Peters, 2014

[PHP] PHP, http://www.php.net/

[Plaisant 1996] C Plaisant, B Milash, A Rose, S Widoff, B Shneiderman, Life Lines: Visualizing personal histories, Proc. of ACM CHI '96, pp. 221-227, Vancouver, April 13-18, 1996

[Plaisant 1998] C Plaisant, R Mushlin, A Snyder, J Li, D Heller, B Shneiderman, KP Colorado, Lifelines: Using visualization to enhance navigation and analysis of patient records. In: In Proceedings of the 1998 American Medical Informatic Association Annual Fall Symposium. pp 76–80, 1998

[Pokorny 2013] J Pokorny, NoSQL databases: a step to database scalability in web environment, Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, pp. 278-283

[Pudil 1998] P Pudil, J Novovičová, Novel Methods for Feature Subset Selection with Respect to Problem Knowledge. In Liu, Huan; Motoda, Hiroshi. 1998.

[Reddy 2015] CK Reddy, CC Aggarwal, Healthcare Data Analytics. Chapman & Hall/CRC, 2015

[Rokach 2010] L Rokach, Chapter 14 A survey of Clustering Algorithms, Data Mining and Knowledge Discovery Handbook (2nd ed), 2010

[Riehmann 2005] P Riehmann, M Hanfler, B Froehlich, Interactive sankey diagrams. In: Proc. IEEE Symposium on Infomation Visualization, InfoVis 2005. pp 233–240., 2005

[SAF] iManageCancer Smart Analysis Framework, http://dimitra.no-ip.info/saf/index.php

[Shahar 1999] Y Shahar, C Cheng, Intelligent visualization and exploration of time-oriented clinical data. Proc. 32nd Annual Hawaii Int. Conf. on Systems Sciences, Hawaii, 1999

[Shahar 2006] Y Shahar, D Goren-Bar, D Boaz, et al., Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, Artificial Intelligence in Medicine, 38(2):115–135, 2006

[Shi 2009] L Shi, N Cao, S Liu, W Qian, HiMap: Adaptive visualisation of large-scale online social networks, Visualisation Symposium, IEEE PacificVis '09, 2009

[Shneiderman 1992] B Shneiderman, Tree visualisation with tree-maps: 2-d space-filling approach, ACM Transactions on Graphics, 11(1): 92-99, Jan. 1992

[Shneiderman 2013] B Shneiderman, C Plaisant, BW Hesse, Improving health and healthcare with interactive visualization methods. IEEE Computer 46(1): 58–66, 2013

[Tanahashi 2012] Y Tanahashi and KL Ma, Design Considerations for Optimizing Storyline Visualisations, IEEE TVCG 18(12), 2012

[Thomas 2005] J Thomas, K Cook, Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, 2005

[Tominski 2004] C Tominski, J Abello, Axes-Based Visualisations with Radial Layouts, Proceedings of ACM Symposium on Applied Computing, 2004

[Tukey 1977] JW Tukey, Exploratory Data Analysis. Addison-Wesley, Reading, MA. 1977

[Vrotsou 2014] K Vrotsou, A Ynnerman, M Cooper. Are we what we do? exploring group behaviour through user-defined event-sequence similarity. Information Visualization, 13(3):232-247, 2014

[Wang 2009] TD Wang, C Plaisant, B Shneiderman, N Spring, D Roseman, G Marchand, V Mukherjee, M Smith, Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. IEEE Transactions on Visualization and Computer Graphics 15(6):1049–1056, 2009

[West 2015] VL West, D Borland, WE Hammond, Innovative information visualization of electronic health record data: a systematic review. Journal of the American Medical Informatics Association 22(2), 2015

[Wijk 1999] JJ Van Wijk, ER Van Selow, Cluster and calendar based visualisation of time, series data, Proceeding of INFOVIS '99, 1999

[Wong 2004] PC Wong and J Thomas, Visual Analytics, 2004

[Wongsuphasawat 2011] K Wongsuphasawat, JA Guerra Gómez, C Plaisant, TD Wang, M Taieb-Maimon, B Shneiderman, Lifeflow: Visualizing an overview of event sequences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1747–1756. CHI ’11, ACM, New York, NY, USA, 2011

[Wongsuphasawat 2012] K Wongsuphasawat, D Gotz, Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization, IEEE Transactions on Visualization and Computer Graphics 18(12):2659–2668, 2012

[Zhao 2015] J Zhao, Z Liu, M Dontcheva, A Hertzmann, A Wilson. Matrixwave: Visual comparison of event sequence data. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI'15, pages 259-268, New York, NY, USA, 2015

[Zhang 2014] Z Zhang, D Gotz, A Perer, Iterative cohort analysis and exploration. Information Visualization, Mar. 2014