Combining Visual Cleansing and Exploration for Clinical Data

Christoph Schmidt[*] University of Rostock
Martin Röhlig[†] University of Rostock
Bastian Grundel[‡] University of Freiburg
Philipp Daumke[§] Averbis GmbH
Marc Ritter[¶] University of Applied Sciences Mittweida
Andreas Stahl[‖] University of Greifswald
Paul Rosenthal[**] University of Rostock
Heidrun Schumann[††] University of Rostock

Abstract

Clinical data have their own peculiarities, as they evolve over time, may be incomplete, and are highly heterogeneous. These characteristics turn a thorough analysis into a challenging task, especially since domain experts are aware of the data flaws, which may impact their trust in the data. As we obtained anonymized clinical data from more than 3,500 patients with retinal diseases, we have to address these challenges. We define a workflow that integrates data cleansing and exploration in an iterative process, so that users are able to easily find anomalies and patterns in the data at any point in their analysis. We implement our workflow in a user-centered visual analytics tool with dedicated visualization and interaction techniques. In collaboration with experts, we apply our tool to examine the interdependency between patients' visual acuity developments and treatment patterns. We find, that real-life data often have unforeseen incidents which can strongly influence the overall visual acuity development. This differs to study results, which are usually conducted under restrictive conditions and have shown visual acuity improvement with on-schedule treatment.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics

1 Introduction

Since clinics are mapping processes and file handling within data management systems for many years, large amounts of data have been collected. These data could reveal valuable information on patient diagnostics and progression [27] making clinical research possibly more efficient [25]. On the other hand, this poses the challenging task of adequately preparing the large amounts of data to help clinicians in finding answers to their specific questions [33]. Even more, there are many subtasks that might be needed or have to be combined for different aims. These range from the simple selection of data over identifying missing values to the judgment of the semantic within presented data, helping to answer real-world clinical questions [37]. To support these subtasks visual analysis tools can be designed to combine different lower-level approaches. Since medical technology and clinical questions differ between clinical fields, specific solution design is needed, created in close collaboration with practitioners. This also applies to our use case, as we analyze large amounts of heterogeneous and abstract patient data from the ophthalmic domain.

In this paper we depict the design and implementation of our visual analysis tool, which helps retinal physicians to answer their questions. To this end, we have developed a dedicated workflow consisting of the three fundamental steps (i) data preprocessing, (ii) data cleansing, and (iii) data exploration. This workflow is then implemented into a tailored visual analysis tool. We use suitable visualization, interaction, and annotation techniques for retinal physicians to both amend the data and support to locate patterns to possibly generate findings. Subsequently, we gather feedback for the tool, by letting expert users analyze the dependency between treatment and visual acuity development.

2 Background

Our goal is to create a visual analysis tool to be applied to the field of clinical data in ophthalmology. For the development process we follow Tamara Munzners nested model [30]. We reflected this by firstly understanding the use case and secondly analyzing existing literature to find applicable techniques or a gap in existing solutions.

2.1 Use case

The first step in Munzners model is an in-depth understanding of the domain, the data, and the users with their questions.

Medical Background

Visual acuity represents the ability to see one's environment sharply and in detail. It highly depends on the condition of the macula, which is the central part of the retina within the human eye. One of the most common macula impairments is the age-related macular degeneration (AMD), which is an important cause of blindness in industrialized countries [29].

To find treatment for that ailment intense research effort has been invested, resulting in successful therapy methods like the injection of anti-VEGF agents directly into the affected eye. By that, many patients are able to sustain or even improve their visual ability [10]. Short term affect as well as long term impact have been proven beneficial [1, 31]. Yet, up to today different facets of the dependencies between injection frequency, medication and the visual acuity development are not fully understood, especially with respect to real-world patient data. Even though more injections over longer time periods can have positive impact [1], each injection bears the risk of infection (surveyed by Falavarjani and Nguyen [13]). Weighing the chances between potential positive impact and infection risk imposes a heavy responsibility on retinal physicians. Providing meaningful information by means of data analysis of existing patients' developments could effectively support such decision making. This is particularly important, as there are often many extraordinary incidents in a patient's medical history, like cataract operation or pigment epithelial detachment, which may distort the overall visual acuity development.

Data Characterization

Our domain experts work in a clinical environment where most data are gathered and stored electronically. To handle the clinical data, which are management data, device data, examination data,

[*] e-mail: christoph.schmidt2@uni-rostock.de
[†] e-mail: martin.roehlig@uni-rostock.de
[‡] e-mail: bastian.grundel@uni-freiburg.de
[§] e-mail: daumke@averbis.de
[¶] e-mail: ritter@hs-mittweida.de
[‖] e-mail: andreas.stahl@uni-greifswald.de
[**] e-mail: research@paul-rosenthal.de
[††] e-mail: heidrun.schumann@uni-rostock.de

and treatment data, a hybrid system consisting of different commercial tools and a bundling in-house web application exists. The *management data* are recorded by clinic assistant personnel via an in-house developed hospital information system. They mainly have some general information, like age and sex. The eye-related *device data* are recorded by the respective device software, including image data, meta data, and abstract data like visual acuity values and tension values. The *examination data* are recorded via proprietary form sheets, created in accordance with analog form sheets that were previously used. During an examination, an assistant or a retinal physician fills in the information collected. All data are summarized in a semi-structured medical-report letter, which is manually filled in by the retinal physician and completed with *treatment data* as well as free text comments from the retinal physician. Treatment data are additionally generated by the surgery management system that is used to plan and document operations, such as anti-VEGF injections or cataract operations. The data are organized by creating one record for each appointment per patient. For our analysis, ophthalmologists selected relevant patients in their clinical system and extracted the anonymized records. As a prerequisite, each selected patient needed to have received at least one injection of anti-VEGF medication, assuring that treatment in some form had been applied. Altogether, we obtained data from records of more than 3,500 patients between the years 2004 to 2018 spread over 95,000 files stemming from different applications and devices in the clinic. The patients have between one and 131 appointments, distributed over a time span between one day and 13 years. The content of the medical report letters for each appointment has been partly extracted into structured data files by a text mining approach. While the results of the text mining are part of the available data, they are only partially considered, as this work does not cover issues related to text mining.

To increase the level of certainty, there are deliberately included redundancies in the data, as some data points are redundantly recorded by different devices or personnel. On the other hand, there are also incomplete data points, as some of the clinical devices may not have been available over the full period of time. Generally, the data were initially gathered for clinical use. So, to use it for research, we have to cope with large amount of abstract, heterogeneous, incomplete, and redundant data. This implies a high demand for data preparation and data cleansing. As data issues may be various and only discovered during the analysis by the experts, data changes may be necessary even during exploration.

Task Abstraction

Identifying the tasks requires a thorough understanding of the users. To achieve this, we cooperated with retinal experts by using observation on the job, interview, and thinking-aloud techniques. The first goal was to understand the working day of a physician, so we performed an observation on the job over three days in an eye care center, where the visualization experts observed the every day working process of the retinal physicians. This was particularly useful in combination with the thinking aloud technique, as the experts could explain the necessity of a domain related task and their motivation. Yet, especially with patients present, this was not always possible. For that reason we conducted interviews after each domain related task, to clarify questions raised and reflect the task and all involved data acquiring systems like slit lamp examination form, etc.. In the aftermath of the observation, the visual analysis experts designed a first draft of the visual analysis tasks. It took six months and several additional interviews to determine a final version of these tasks. There are several reasons for that: (i) the visual analysis experts recorded the domain related tasks in high detail, including some unrelated and misleading information, (ii) the providable data changed several times, (iii) both the domain experts and the visual analysis experts had to define a common ground of understanding. Especially the last one was challenging, as the domain experts favor straight forward approaches with sparse color and form usage, which has a major influence on our approach.

Combining the techniques above, we learned that physicians are well aware of the available data for an individual patient in the clinical system. They normally use the system to search for and show information on individual patients via a web browser. Here, the user can search the overall database via SQL web queries and display the result in table form. With that information, they commonly transfer their existing knowledge, e.g., pathological information from previously seen patients, to the current case. In combination with their pattern recognition ability and internalized expertise on retinal disease characteristics, they find similarities between the patients. Based on that, they decide on the further course of action. This conventional way of working typically allows to take the last few injections of a patient into account. Yet, they struggle to adequately assess the response to therapy over longer periods of time. Individual factors such as duration of therapy response, which would be apparent from previous treatments, cannot be taken into account for the current therapy decision. This withholds the experts from identifying general relations between the visual acuity development and given injections. Furthermore, the available data are not fully structured and not necessarily complete. Identifying data that are missing within result lists is, however, challenging. Hence, the physicians usually start by analyzing the medical report letters one at a time to see if certain information is missing. While that information for a specific visit of an individual patient is interpretable, a structured export or automated identification and correction of all unstructured data points is not possible. Based on these domain analysis results, we identified three basic tasks that need to be supported:

(i) Efficiently find missing values: Due to the large amount of incomplete data, an efficient way to find and amend missing values is necessary.
(ii) Relate visual acuity values to the applied treatment over time: To analyze the relationship between the visual ability of patients and their treatment patterns, both must be visualized together.
(iii) Find major incidents in visual acuity development: Finding these incidents is important for the overall development prediction, as they are independent from the original disease and treatment.

Supplementary to the task identification we experienced that observation on the job is particularly useful. It supported the detection of low level requirements, like interaction habits, performance, and tool usage experience of ophthalmologists in clinics. For high level task definition and domain understanding, iterative communication is indispensable.

2.2 Related Work

In Munzners nested model, the next step is the integration of a suitable visual encoding and interaction idiom. In this section, we list current solutions that have inspired us for our for the task fulfillment, and outline limitations we had to overcome.

Identify missing values

For dealing with undetected missing values, and value inconsistencies, several approaches with different features and limitations exist. While many are algorithm-focused data cleansing approaches (e.g. [11, 17]), we focused on an iterative human-in-the-loop workflow. This is also recommended by Krishnan et al. [22] as a result of a user survey on the topic of data cleansing with 29 experts from academia and industry. For data preprocessing, we were inspired by the visual analytics tool for epidemiological cohort study data by Alemzadeh et al. [4]. Yet, it does not have the ability to identify completely missing values. For data cleansing Kandel et al. [21] present the tool ”profiler”, which allows users to explore data and gather information on missing values or anomalies. While this approach works for large data, we have missing values with dependencies to existing data points (e.g. due to an existing injection incident it is possible to identify a missing visual acuity measurement). Gschwandtner et al. [15] present the TimeCleanser tool. Based on a previously published taxonomy of errors in time-oriented data [16], the authors integrate an interface for improving correction algorithms with different use-case specific visualizations. For large time-series data the tool ”VisPlause” from Arbesser et al. [5] assesses data quality with the focus on issue identification. They conduct a design study with domain experts in the energy sector, with the goal to visualize and analyze known data issues. As we need to combine the identification and amendment of known and unknown missing values, all of these approaches are inspiring, but not directly transferable to our use case in the field of ophthalmology.

Visualizing Incidents and Sequences in Time-oriented Data

The visualization of time-oriented data has a long history [2] and many approaches for different purposes exist. Visualizing time-oriented clinical data together with indications of medical relevant events has already been introduced by Cousins and Kahn [12]. Based on that, very suitable approaches for incident and sequence visualization in time oriented data are ”LifeLines” from Plaisant et al. [32] and ”LifeLines2” from Wang et al. [39]. While the former introduces a visualization for sequences and incidents in patient data, the latter extends that approach to visualize several timelines at once. Taking that as a base, we examined other literature to identify further techniques adequate to the tasks. The temporal sequence of different motion capture variables is visualized by Müller et al. [28] with the help of different color codings. To provide a global overview of motion capture sequences, Bernard et al. [8, 9] visualize measured time-dependent data together with events and event sequences. The limitation here is, that their sequences are categorical data, while we have numerical data. A similar tool for visually analyzing event data streams is presented by Fischer et al. [14], specifically designed for the field of system administration and monitoring. A closer look into the use of visual analytics in the biomedical domain is provided by a survey by Turkay et al. [38]. Shahar et al. [36] focus especially on browsing through patient data via temporal abstraction and semantic navigation. Bade et al. [6] present a stacking-based overview visualization of health data together with a detail view for clinical treatment support. In a recent work, Zhang et al. [42] superimpose measured data for different patients with symbols for indicating relevant events. In contrast to their data, we are dealing with high variability in the measured time sequences and event gaps.

Annotating Missing Values and Data Relations

To fulfill the tasks from Sec. 2, (i) missing values and other data issues must be marked and corrected before exploration and (ii) the outcome of the exploration must be externalized. A common approach in this regard is the use of annotations, which Lipford et al. [23] and Mahyar et al. [24] point out to be a critical step in visual analytics. While Heer and Shneiderman [19] provide general techniques for annotations to mark peculiarities within a visualization, Schmidt et al. [35] present a general characterization of annotations. They capture different types and reason for annotations as well as ways to gather and communicate them. For each of these aspects, various examples in literature exist. Hellerstein et al. [20] develop a system for domain experts to amend, change, and prepare their data by adding information, mainly based on text and table entries. Annotations within a visualization are shown by Willett et al. [41], who support the integration of textual comments. They can be amended with screenshots and categorized via tags, yet without supporting data changes. Al-Naser et al. [3] enrich the data by integrating user information, which is used to customize the data visualization in a multi-user environment for spatial data. [26]

2.3 Summary

In conclusion, existing visual analysis techniques can facilitate task fulfillment in the highly specific field of ophthalmology. Yet, we didn’t find any description, that combines the various techniques for a comprehensive analysis of clinical ophthalmic data for our tasks. Our goal is therefore to define a workflow that channels and structures the tasks of the experts and allows us to combine and advance existing visualization techniques in an appropriate way.

3 Workflow Development

With the goal to amend and explore the data we design a workflow. It must assure that erroneous data is identified and, where possible, corrected without having a complete list of possible errors or incompleteness. We therefore designed the steps data cleansing and data exploration in addition to the basic step data preprocessing. During implementation we realized that it is hardly possible to detect and correct all errors at once. So, to allow data cleansing at a later stage, we designed the cleansing and exploration steps as iterative and alternately feasible, as shown in Fig. 1.

3.1 Overview

The workflow is divided into three fundamental steps: (i) data preprocessing, (ii) data cleansing loop, and (iii) data exploration loop. The *data preprocessing* ensures that the raw data are correctly imported, structured, merged, and presented to the user in an appropriate way. The user then removes structural. The *data cleansing loop* supports the identification, removal, or amendment of erroneous or missing data on the semantic level. Within the *data exploration loop* the experts can examine interconnections between data dimensions and identify, select, annotate, and/or export their findings. While the workflow foresees an initial run of the data cleansing loop, it deliberately allows a switching between the cleansing and exploration loop at a later stage. This gives the possibility to iteratively amend the data at any point of the analysis.

3.2 Data Preprocessing

The purpose of the data preprocessing is the correct structuring of the raw data, supported by basic filtering functions. It starts with the import of the data either from a specific patient subset or a random sample. For these patients the differently structured JSON files from various sources are converted into a multidimensional data structure with numerical, categorical, and textual dimensions. At this point, the data is automatically evaluated based on rules and, if necessary, converted or sorted out. These rules range from simple thresholding to refining some dimensions, e.g., specifying the ”International Statistical Classification of Diseases and Related Health Problems” (ICD10) information in the raw data with a mapping table created by experts. After these rules have been applied the data is presented to the user, who can then find and remove data elements with structural issues, like empty patients or empty data dimensions.

3.3 Data Cleansing Loop

After the data have been imported and structured with basic data filtering applied, the workflow foresees an initial data cleansing loop as shown in Fig. 1. It allows for identification and correction of common or specifically targeted data issues, and thus supports the first task from Sec. 2. The data cleansing loop has the four steps (1) visualize, (2) adjust view, (3) browse & select, and (4) change data. Step (1) is performed by the computer, which generates an overview visualization of the structured data. To perform the data cleansing task as well as possible, this step already includes preset visualization parameters. Yet, the user still has the possibility to adapt them in step (2) in case of personal preferences. A personalized and task related visualization is an important prerequisite for step

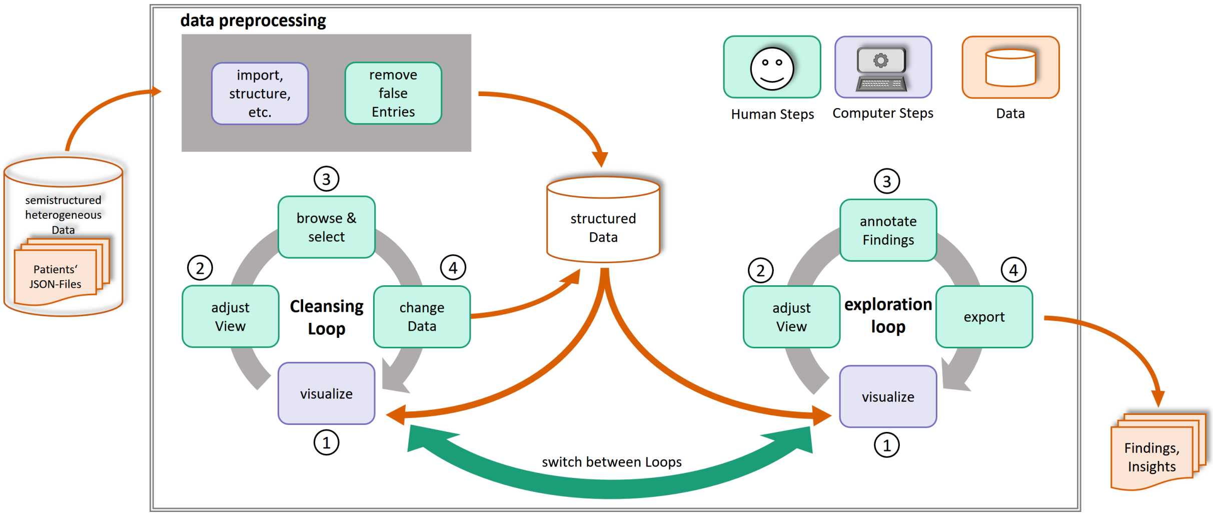


Figure 1: General workflow, developed on the basis of the task abstraction in Sec. 2. It consists of three fundamental steps *data preprocessing* (upper left), *data cleansing loop* (lower left) and *data exploration loop* (lower right). The cleansing and exploration loops are divided into four steps each, which are further detailed in Sec. 3.

(3) the browsing and selecting step. Here the experts' core task is the judgment of the data semantic content. They have to find data-points that need to be added, changed, or removed within the data cleansing loop. The reasons for the data amendment can be discrepancies, swapped, or wrongly assigned data-points, but also missing values, which can only be identified with the tacit knowledge of the experts. If necessary, the user can mark, delete, amend, or add a data-point, representing the final step (4) in the data cleansing loop. Having finished one run of the loop, the user may want to focus on another data discrepancy issue, and thus initiate a new run of the data cleansing loop. Alternatively the user can move on to the exploration loop.

3.4 Data Exploration Loop

The data exploration loop, as shown in Fig. 1, supports task two and three from Sec. 2. It has the goal to relate the visual acuity development to treatment incidents and to support the location of major incidents. This can be done either for a single patient or different patients. The data exploration loop has the four steps (1) visualize, (2) adjust view, (3) annotate findings, and (4) export. The first step provides an overview visualization of the chosen data dimensions for the imported dataset (1). In the second step the visualization can be adapted by the user, so that it is suitable for the data exploration. That means, the visualization parameters have to be set in a way to see interdependencies between different data points. In addition, the user must be able to judge the semantic of the data-points as well as their connection to others.

The exploration loop supports free exploration. In doing so, an interplay between visualization and interaction is created, supporting the relation of the visual acuity values to treatment incidents over time. Additionally, it allows the identification of extraordinary incidents for a patient. In step (3) the user selects, marks, and/or comments the finding via annotations directly within the visualization. These findings can be externalized in step (4) by a dedicated data export of the dataset with the annotations, preserving the reference to the original data-point(s).

During the data exploration loop users usually identify previously overlooked data issues, for example accidentally swapped visual acuity values for the left and right eye of a patient. If this would remain uncorrected the visual acuity development may be corrupted. So, we support the correction of data entries at any time by always allowing to switch between the loops.

4 The Design of the Tool

With the given user tasks (cf. Sect. 2) and established workflow (cf. Sect. 3) we design a tailored visual analysis tool. It supports structuring of heterogeneous data, reading and writing of data, as well as suited visualization techniques that apply to the needs of the data and the domain experts. This includes data operations for improvement as well as the flexibility of switching between cleansing and exploration without disrupting the current mental map of the user. Regarding data preprocessing, we can build upon our previous work about unifying and structuring data in the ophthalmic domain [34]. Regarding data annotations we can build upon our conceptual annotation framework published in [35].

The architecture of our tool consists of the data layer, analytics layer, and user interface layer. The *data layer* reflects the necessity to parse, change, and store the data and annotations. It represents the interface between the internal data storage, the other layers and external data sources and thus controls all data exchange. The *analytics layer* provides different aggregations of the data, sorting/grouping functions and parameter settings for the user interface. In order to support modularity for later extension, it is designed as a separate component with interfaces to the other two layers. With the given data and parameters the *user interface layer* can present the data to the user in the visualization component. Additionally the interaction component ensures that necessary user actions like parameter changing, data amendment, or annotation creation can be supported.

4.1 Data Layer

The data layer facilitates the data preprocessing, and both the data cleansing and exploration loops in the workflow. It is responsible for internal data storage and amendment. With regard to the data preprocessing this means, that the imported data is converted and stored internally in two linked structures. The first structure reflects the appointment-wise way from the clinic to support the data cleansing loop. Each appointment is a heterogeneous n-dimensional vector. The value of n is calculated by counting the number of values recorded on that appointment plus the number of additional dimensions. The additional dimension contain information on the data source, the import date and possible redundancy. The latter is important, e.g., for emphasis during the cleansing loop. With regard to the exploration loop, we create a second data structure. In order to set all values for one patient in context to ease the exploration, the second structure is patient oriented. It holds the generated aggregations from the analytics layer, such as average visual acuity, number of injections, regression information, and so on. In addition to the aggregated information, the patient structure holds lists of incident information for that patient. These include visual acuity measurement lists, that allow to quickly provide time-series data e.g., for regression calculation and visualization.

When the initial parsing is finished and the data is internally stored, all further *data amendment* is only done on user request. Hereby, the amendment is always done for both internal structures and stored in additional dimensions. As soon as a data amendment request is carried out, the updated data is sent to all affected user interfaces.

4.2 Analytics Layer

The analytics layer is designed in two parts. The first part is the parameter definition component and supports both the data cleansing and exploration loops. It controls the visualization-parameter settings through presetting functions, restriction definition, and automatic parameter changing, implementing a rule based specification. So, for instance, if the data dimension changes from linear to logarithmic, the color coding parameter is automatically changed from linear to logarithmic, to ensure correct value encoding. On top of that it provides predefined parameter settings with specific adjustments either for the data cleansing loop or the exploration loop.

The other part is the data analytics component, designed to support tasks two and three from Sec. 2, which are fulfilled in the exploration loop. Both tasks can be performed on the local level (intra-patient) or global level (inter-patient) and need to be seen in the context of the domain experts' measurement units. We thus need specific scaling and aggregation functions to reflect the need in the visualization.

appropriate scaling of data dimensions: To determine scales for both intra-patient and inter-patient exploration it is necessary to derive value boundaries for several dimensions. These can be maximal and minimal values and/or the total number of occurrences (e.g., number of treatments). Furthermore, to reflect the common units of the domain experts, conversion from logarithmic scale units to linear scale units is performed. While domain experts internally communicate using the decimal visual acuity value (V_{dec}), its values do not represent linear changes. Therefore the logMar visual acuity value (V_{logMar}) has been developed [7] to ease calculations. Additionally, we need the letter score (V_{letter}) which is used to measure visual acuity differences (see Wecker et al. [40]), so that a mutual conversion is necessary.

aggregation of patient data: For intra-patient numerical datapoints, like visual-acuity values, *arithmetic means and medians* are calculated. They provide a single data point for each patient and each dimension for comparison. To see the distribution of a data dimension for all incidents for inter-patient comparison, the occurrences of the values are counted. This gives the *absolute frequency*, and is set in relation to the total number of values for that dimension to obtain the *relative frequency*. As the overall visual acuity development over time is one of the major judgment parameters for the domain experts, we have to calculate the difference between the visual acuity value at the beginning and at the end of each patient's monitoring period. Using the first and last visual acuity values is not sufficient, as there is fluctuation depending on the patient's daily form. Instead, we use the *linear regression function*, as it fits the patients values to linear curve, better representing starting and end points.

With the linear regression function it is possible to derive the visual acuity difference between the starting point and end point for each patient. Because this function produces a decimal value that has not been physically measured, a direct encoding could lead to over interpretation by the experts. Additionally, a precise value is not necessary to get an impression on the overall performances. After discussion with the experts, we reduce the outcome in two ways. On the one hand, we use the slope results to visually indicate the development for a patient, which will be detailed in Sub-Sec. 4.3.1. On the other hand, we use the classes, 'gain', 'unchanged' and 'loss', from the ophthalmic domain (see [40]) with the following boundaries:

$$f(x) = \begin{cases} loss & : & & & \Delta V_{letter} & < & -15 \\ unchanged & : & -15 & \leq & \Delta V_{letter} & \leq & 15 \\ gain & : & 15 & < & \Delta V_{letter} & & \end{cases}$$

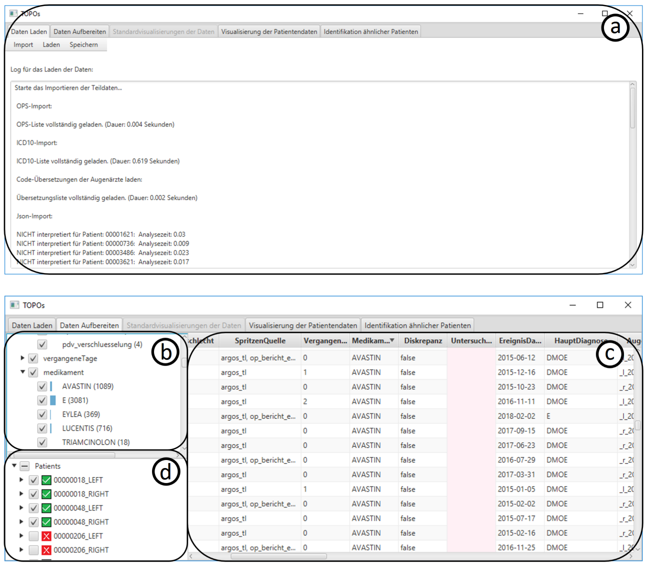
These classes allow users to mentally assign patients into a performer group and aim at preserving the expressiveness of the regression result by reducing the precision.

4.3 User Interface Layer

The user interface layer presents specific visualizations and provides interaction functions that are unique for the different steps in the workflow. We follow the common approach to separate between visualization and interaction, which each represent an own component.

4.3.1 Visualization Component

Having the discussions with the domain experts in mind, we designed a carefully aligned combination of proven techniques and new elements. This allows our tool to use appropriate visualization techniques in accordance with the workflow described in Sec. 3. By that we are able to reach a high level of acceptance among users. To step by step increase visualization complexity, we design three screens. The first two are the data import/export and preprocessing screens and the third is the data cleansing and exploration screen. Both the *data import/export* and *data preprocessing* screens (Fig. 2) have been designed in line with existing screens in the clinic and


TOPOs
Daten Laden
Daten Aufbereiten
Standardvisualisierungen der Daten
Visualisierung der Patientendaten
Identifikation ähnlicher Patienten
Import
Laden
Speichern
a
b
c
d

Log für das Laden der Daten:

```
Starte das Importieren der Teildaten...

 OPS-Import:

 OPS-Liste vollständig geladen. (Dauer: 0.004 Sekunden)

 ICD10-Import:

 ICD10-Liste vollständig geladen. (Dauer: 0.619 Sekunden)

 Code-Übersetzungen der Augenärzte laden:

 Übersetzungsliste vollständig geladen. (Dauer: 0.002 Sekunden)

 Json-Import:

 NICHT interpretiert für Patient: 00001621:  Analysezeit: 0.03
 NICHT interpretiert für Patient: 00000736:  Analysezeit: 0.009
 NICHT interpretiert für Patient: 00003486:  Analysezeit: 0.023
 NICHT interpretiert für Patient: 00003621:  Analysezeit: 0.017
```

☑ pdv_verschluesselung (4)
▸ ☑ vergangeneTage
▾ ☑ medikament
☑ AVASTIN (1089)
☑ E (3081)
☑ EYLEA (369)
☑ LUCENTIS (716)
☑ TRIAMCINOLON (18)

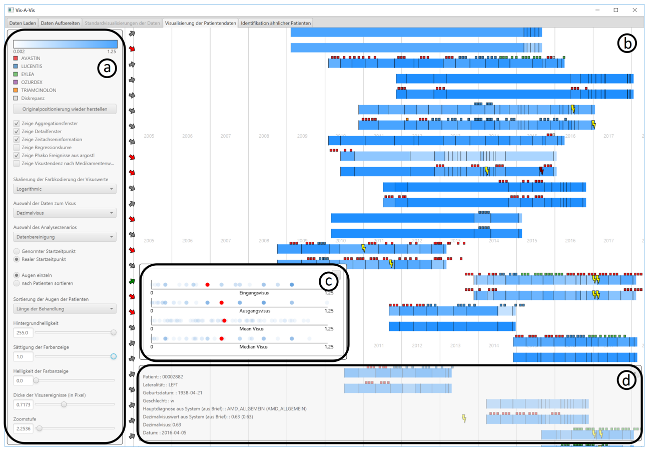
▾ ▭ Patients
▸ ☑ ✓ 00000018_LEFT
▸ ☑ ✓ 00000018_RIGHT
▸ ☑ ✓ 00000048_LEFT
▸ ☑ ✓ 00000048_RIGHT
▸ □ ☒ 00000206_LEFT
▸ □ ☒ 00000206_RIGHT

chlecht	SpritzenQuelle	Vergangen...	Medikam...▼	Diskrepanz	Untersuch...	EreignisDa...	HauptDiagnose	Aug
	argos_tl, op_bericht_e...	0	AVASTIN	false		2015-06-12	DMOE	_20
	argos_tl	1	AVASTIN	false		2015-12-16	DMOE	_l_20
	argos_tl	0	AVASTIN	false		2015-10-23	DMOE	_r_20
	argos_tl	2	AVASTIN	false		2016-11-11	DMOE	_l_20
	argos_tl, op_bericht_e...	0	AVASTIN	false		2018-02-02	E	_l_20
	argos_tl	0	AVASTIN	false		2017-09-15	DMOE	_r_20
	argos_tl	0	AVASTIN	false		2017-06-23	DMOE	_r_20
	argos_tl, op_bericht_e...	0	AVASTIN	false		2016-07-29	DMOE	_r_20
	argos_tl	0	AVASTIN	false		2017-03-31	DMOE	_r_20
	argos_tl	1	AVASTIN	false		2015-01-05	DMOE	_l_20
	argos_tl, op_bericht_e...	0	AVASTIN	false		2015-02-02	DMOE	_r_20
	argos_tl	0	AVASTIN	false		2015-07-17	DMOE	_r_20
	argos_tl	0	AVASTIN	false		2015-02-16	DMOE	_r_20
	argos_tl, op_bericht_e...	0	AVASTIN	false		2016-11-25	DMOE	_l_2

Figure 2: Data import/export screen with log information (a). Data preprocessing screen (b) through (d). (b) the distribution view, showing the distributions for all data dimensions of the appointment, including the absolute frequency and relative frequency. (c) the appointment view, showing all dimensions for one appointment. (d) the patient view, showing the patient oriented structure to show all information for one patient.

proven techniques. This allows users to familiarize themselves with the tool. To increase the trust in the data, it is shown in text form, so that expert have a direct reference to the original data. To make sure, that there's a cut between the external data-source retrieval and further internal processing, we visually separated the data import/export from the data preprocessing.

The main purpose in the data preprocessing is the exclusion of implausible data values for all patients (Fig. 2b), empty patients (Fig. 2d), and/or empty data dimensions (Fig. 2c).



☑ Zeige Aggregationsfenster
☑ Zeige Detailfenster
☑ Zeige Zeitachseninformation
☐ Zeige Regressionskurve
☑ Zeige Phako Ereignisse aus argostl
☐ Zeige Visustendenz nach Medikamentenw...

Skalierung der Farbkodierung der Visuswerte: Logarithmic

Auswahl der Daten zum Visus: Dezimalvisus

Auswahl des Analyseszenarios: Datenbereinigung

○ Genormter Startzeitpunkt
● Realer Startzeitpunkt

● Augen einzeln
○ nach Patienten sortieren

Sortierung der Augen der Patienten: Länge der Behandlung

Hintergrundhelligkeit: 255.0

Sättigung der Farbanzeige: 1.0

Helligkeit der Farbanzeige: 0.0

Dicke der Visusereignisse (in Pixel): 0.7173

Zoomstufe: 2.2536

Patient: : 00002882
Lateralität: : LEFT
Geburtsdatum: : 1938-04-21
Geschlecht: : w
Hauptdiagnose aus System (aus Brief): : AMD_ALLGEMEIN (AMD_ALLGEMEIN)
Dezimalvisuswert aus System (aus Brief): : 0.63 (0.63)
Dezimalvisus: 0.63
Datum: : 2016-04-05

Figure 3: The patient-data visualization-screen. The control panel (a) holds the legend and the parameter setting. The main visualization view (b) shows the time-oriented data with data segments and incidents. The summary panel (c) shows aggregated values for patients and their distributions for all patients. The detail view (d) shows detailed information on a specific data point on demand.

The *data cleansing and exploration* screen as shown in Fig. 3 has the goal to communicate time-oriented data of different types from multiple sources, with redundancies and potential discrepancies. On top of that, the data derived from the analytics layer has to be integrated. In respect to that, we decided to show the content in four separate views.

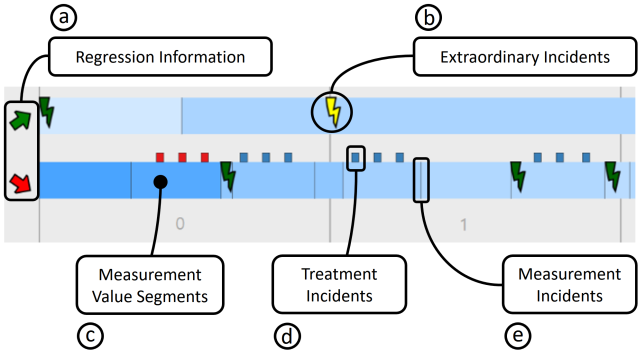


Figure 4: Detailed view of the patient data visualization. At the beginning of each horizontal bar the regression slope information is shown (a). The bar itself consists of color coded segments each representing a visual acuity measurement value (c), the measurement incidents at a certain date (e), the color coded treatment incidents, showing an injection and the medication used. Extraordinary incidents are visualized as color coded flashes (b).

The Main View (3b) has the purpose to allow data cleansing and exploration by visualizing the defined patient data dimensions as detailed in Fig. 4. As the visualized data dimensions are segments and incidents in time for different patients, we divided the visualization space into horizontal rows, which evolve in time from left to right. Each row represents the data for one eye of a patient and displays several data dimensions. The segment visualization aims to show the development over time. We therefore color code each segment and place it in accordance with its time-slot (4c). The color communicates either visual acuity values or visual acuity deviation using appropriate color schemes as published in Harrower and Brewer [18]. As the measured visual acuity values may have some fluctuation, it can be difficult to detect the overall development over the full time period. So, we use the regression data from the analytics layer described in Sec. 4.2 to visualize an additional glyph at the beginning of each row (4a). This arrow shows the visual acuity development by three domain specific classifications: gain (green), unchanged (grey), or loss (red). For a more precise indication, the regression slope for each patient is encoded in the angle of the arrow.

To mentally connect the segments to the incidents, we encoded them locally connected. As incidents (4b), (4d), & (4e) refer to a specific point in time, they are represented as small regular (4d) & (4e) or irregular (4b) glyphs. Since they have varying medical consequences, it is important to differentiate between them. The regular incident glyphs represent regular incidents, like planned measurements or treatments, smoothly connected to the corresponding horizontal bar. Glyphs that represent extraordinary incidents, on the other hand, have an irregular shape, to show the disruption they represent.

The Control Panel (3a) supports steering and understanding of the main view. It is designed to provide an explanatory legend for colors used in the main view. Additionally it holds visualization-parameter controls, depending on the parameter type, to display and control the current parameter state.

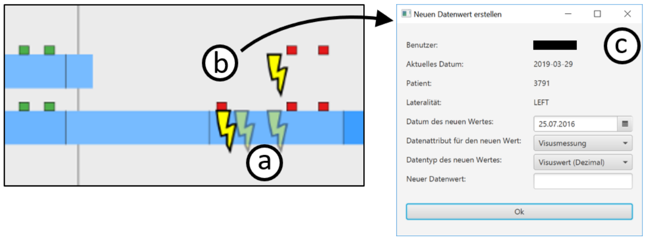
The Summary View (3c) has the goal to provide an overview of the data distribution for all loaded patients. It shows aggregated data for four data dimensions (first, last, mean, and median visual acuity values) for all patients at once and emphasizes the values of a selected patient to evaluate its position in the total. In contrast to the main view, which shows time segments, here the data per time point is displayed. The color intensity encodes the frequency of associated data points.

Finally, **the Detail View (3d)** provides detailed information of just one data point for a patient. Whereas the summary allows for a general overview over all patients, the detail view supports the judgment of a single data point for cleansing or exploration.

4.3.2 Interaction Component

Our interaction concept is mainly driven by two factors: (i) intuitive support of well known techniques and (ii) concurrent interaction purposes, like data cleansing, browsing and annotating. To support intuitive interaction in the *data import/export screen* and the *data preprocessing screen* (Fig. 2), we use the mouse for scrolling, row-selection, sorting, column rearrangement as well as filtering, which is in line with well known techniques from the existing clinical system.

For the views on the visualization screen (Fig. 3), interaction methods get more complex, as they have to support (i) the visualization parameter adaptation, (ii) the visualization navigation, (iii) the data cleansing, and the (iv) data annotation. The visualization parameter adaptation can be performed via the dedicated parameter control panel described above. The user can change specific parameters with their controls or choose preset parameter settings using a combobox. By that, the visualization is changed on the fly. Scrolling and zooming allows to navigate through the visualization, to find data points for cleansing or interest of exploration. Selecting such points through mouse-hovering provides additional information on demand. The cleansing support, which has to take care of marking and/or removing of existing data-points and the creation of new ones, is shown in Fig. 5.



Neuen Datenwert erstellen

Benutzer: [redacted]
Aktuelles Datum: 2019-03-29
Patient: 3791
Lateralität: LEFT
Datum des neuen Wertes: 25.07.2016
Datenattribut für den neuen Wert: Visusmessung
Datentyp des neuen Wertes: Visuswert (Dezimal)
Neuer Datenwert:

Figure 5: Interaction support for cleansing. Segments or incidents can be cleansed with a mouse-click (a). New incidents can be created in the visualization (b) supported by an interactive dialogue (c).

The user can either use the primary mouse button to visually mark a specific data-point as ”cleansed” with transparency (a) or allow the creation of new data points with local reference in the visualization (b) with the secondary mouse button, supported by an interactive dialogue (c). The interaction support for annotations during the exploration loop is similar to the data cleansing type, yet the annotation window shows a comment section instead of data-point amendment fields. Additionally the respective data point is emphasized as annotated. All generated comments for that data point are displayed in the detail view.

5 Expert Feedback

With the tool at hand, we organized two sessions to gather feedback from the experts. The first session was designed as an application session with an expert, usually working in the conventional retrieval and analysis of patient data (cf. Sect. 2.1). The second session was held as a tool demonstration with a mixed group of experts, including ophthalmic practitioners and research scientists.

With the *application session*, our general goal was to test the practical utility of the tool. Particularly, we aimed to assess: (i) the appropriateness of the visual design and (ii) the usability of the interaction functionality to solve the identified tasks (cf. Sect. 2). To answer these points, a senior retinal physician applied the tool supported by a visual analytics expert. The data for the session consisted of an arbitrarily chosen sample of 205 patients with a total of 9790 regular and irregular incidents. All workflow steps were performed within a time-frame of 60 minutes.

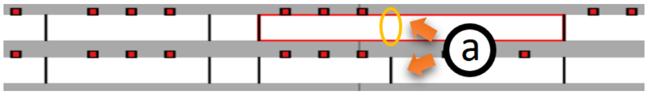


Figure 6: Data cleansing task. The user has adjusted the visualization parameters, so that missing visual acuity measurement incidents can be detected. (a) The arrows point at the left and right eye of the same patient. As the clinic personnel always measures both eyes, it is unusual, that on eye misses a measurement.

Starting with the data preprocessing, the expert's first objective was to identify any flaws in the data. The import log and preprocessing screens helped to reveal several major issues, including missing or irrelevant values (for preprocessing screen see Fig. 2b-d). The expert applied automated filtering rules and used available interaction functions to exclude the corrupted data entries. As a result, a structured dataset with 204 medically relevant patients and 9104 incidents was obtained. Continuing with the data cleansing, the expert aimed at improving the data quality. By looking at the patient-data screen, the expert quickly noticed few instances of missing visual acuity measurements. The visualization parameters were switched accordingly to put further emphasis on these issues (Fig. 6). Browsing through the data helped the expert to classify them as a reoccurring problem, with 73 missing values found in total. Using the annotation function (Fig. 5), all missing values could be directly amended. In this regard, the expert pointed out the interactive manipulation of the visualized data to be particular intuitive and useful. Finally, the expert explored the cleansed data to study the visual acuity development in relation to injections received. An appropriate parameter preset for the patient-data screen was loaded and refined. While browsing through the data, the expert located several extraordinary incidents, such as a patient with a sudden loss of visual ability. The linked summary and detail views allowed to investigate the incidents, check their plausibility, and record gained insights. The expert reassured us that the provided functionality was indeed helpful to draw the conclusion that such incidents are unexpectedly common and strongly influence the visual acuity development. In the end, all tasks of the application session were completed successfully. In retrospect, the expert particularly appreciated the ability to jump back and forth between the exploration and cleansing loops to immediately process every discovered erroneous data point. The expert concluded that reducing the manual data processing effort compared to current procedures while eventually being able to obtain analysis results with higher accuracy are great benefits.

With the *demonstration session*, our general goal was to assess the medical relevance of the design concept. A group of three ophthalmic experts, including a head physician, from an eye care center specialized in the diagnosis and treatment of retinal diseases participated. A live demonstration of the tool and its main components was given by a visual analytics expert based on the described data and workflow. Informal feedback was gathered during the demonstration and in subsequent discussions. The experts particularly appreciated the workflow for enabling a more structured way of working with the clinical data. Regarding the design, one expert stated: ”After looking at the visualization for a while, I begin to recognize patterns, similar to looking at patient images”. In the discussions, this statement was attributed to the new visualization design choices (row-wise arrangement of patient data and applied color presets) as well as to the consideration of familiar presentations known from the conventional data analysis. Overall, the feedback of the second session was very positive. The experts rated the design concept to be effective and the medical outcome to be highly relevant. Based on the demonstration, the experts even decided to present the tool and the generated results at the largest ophthalmology congress in Germany, the DOG Congress in Bonn 2018.

6 Conclusion and Future Work

We have shown that the integration of a suitable workflow with visual analytics methods can be beneficial for the task fulfillment in the domain of ophthalmology in clinical environments. One example is the acknowledgement of the commonness of extraordinary incidents that are not related to the original treatment, yet still influence the visual acuity development. Facilitating the externalization of data amendments and findings, we allow for permanent data improvement and potential knowledge generation. Applying our tool together with the experts triggered discussions and sparked new ideas for further improvements. For instance, the experts suggested to add custom sorting capabilities and task-dependent groupings of patients. Furthermore, it would be interesting to automate the parameterization of our visualization design. This way, we could suggest suitable parameters for given tasks and support users in finding specific patterns in the data. Accordingly, our tool design will be used as a basis for further extension, fine-tuning, and evaluation together with them. Finally, analyzing the captured annotations from previous cleansing and exploration iterations could help in finding patients with similar issues and allow to reduce the overall time needed for completing the workflow.

References

[1] S. D. Adrean, S. Chaili, H. Ramkumar, A. Pirouz, and S. Grant. Consistent long-term therapy of neovascular age-related macular degeneration managed by 50 or more antiVEGF injections using a treat-extend-stop protocol. *Ophthalmology*, 125(7), 2018. doi: 10.1016/j.ophtha.2018.01.012

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2011.

[3] A. Al-Naser, M. Rasheed, D. Irving, and J. Brooke. A visualization architecture for collaborative analytical and data provenance activities. In *Proc. of IV*, 2013. doi: 10.1109/IV.2013.34

[4] S. Alemzadeh, U. Niemann, T. Ittermann, H. Völzke, D. Schneider, M. Spiliopoulou, and B. Preim. Visual analytics of missing data in epidemiological cohort studies. In *Proc. of VCBM*, 2017. doi: 10.2312/vcbm.20171236

[5] C. Arbesser, F. Spechtenhauser, T. Mhlbacher, and H. Piringer. Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE TVCG*, 23(1), 2017. doi: 10.1109/TVCG.2016.2598592

[6] R. Bade, S. Schlechtweg, and S. Miksch. Connecting time-oriented data and information to a coherent interactive visualization. In *Proc. of CHI*, 2004. doi: 10.1145/985692.985706

[7] I. L. Bailey and J. E. Lovie. New design principles for visual acuity letter charts. *Am J Optom Physiol Opt*, 53(11), 1976.

[8] J. Bernard, E. Dobermann, M. Bögl, M. Röhlig, A. Vögele, and J. Kohlhammer. Visual-interactive segmentation of multivariate time series. In *Proc. of EuroVA*, 2016. doi: 10.2312/eurova.20161121

[9] J. Bernard, N. Wilhelm, B. Krüger, T. May, T. Schreck, and J. Kohlhammer. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE TVCG*, 19(12), 2013. doi: 10.1109/TVCG.2013.178

[10] S. B. Bloch, M. Larsen, and I. C. Munch. Incidence of legal blindness from age-related macular degeneration in denmark: Year 2000 to 2010. *Am J Ophthalmol*, 153(2), 2012. doi: 10.1016/j.ajo.2011.10.016

[11] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proc. of SIGMOD*, 2015. doi: 10.1145/2723372.2749431

[12] S. B. Cousins and M. G. Kahn. The visual display of temporal information. *Artif. Intell. Med.*, 3(6), 1991. doi: 10.1016/0933-3657(91)90005-V

[13] K. G. Falavarjani and Q. D. Nguyen. Adverse events and complications associated with intravitreal injection of anti-VEGF agents: a review of literature. *Eye*, 27(7), 2013. doi: 10.1038/eye.2013.107

[14] F. Fischer, F. Mansmann, and D. A. Keim. Real-time visual analytics for event data streams. In *Proc. of SAC*, 2012. doi: 10.1145/2245276.2245432

[15] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy. Timecleanser: A visual analytics approach for data cleansing of time-oriented data. In *Proc. of i-KNOW*, 2014. doi: 10.1145/2637748.2638423

[16] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch. A taxonomy of dirty time-oriented data. In *Proc. of CD-ARES*, 2012. doi: 10.1007/978-3-642-32498-7_5

[17] D. Haas, S. Krishnan, J. Wang, M. J. Franklin, and E. Wu. Wisteria: Nurturing scalable data cleaning infrastructure. *Proc. of VLDB Endow.*, 8(12), 2015. doi: 10.14778/2824032.2824122

[18] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartogr J*, 40(1), 2003. doi: 10.1179/000870403235002042

[19] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2), 2012. doi: 10.1145/2133416.2146416

[20] J. M. Hellerstein, J. Heer, and S. Kandel. Self-service data preparation: Research to practice. *IEEE Data Eng. Bull.*, 41(2), 2018.

[21] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proc. of AVI*, 2012. doi: 10.1145/2254556.2254659

[22] S. Krishnan, D. Haas, M. J. Franklin, and E. Wu. Towards reliable interactive data cleaning: A user survey and recommendations. In *Proc. of HILDA*, 2016. doi: 10.1145/2939502.2939511

[23] H. R. Lipford, F. Stukes, W. Dou, M. E. Hawkins, and R. Chang. Helping users recall their reasoning process. In *Proc. of VAST*, 2010. doi: 10.1109/VAST.2010.5653598

[24] N. Mahyar, A. Sarvghad, and M. Tory. Note-taking in co-located collaborative visual analytics: Analysis of an observational study. *Information Visualization*, 11(3), 2012. doi: 10.1177/1473871611433713

[25] K. A. Mc Cord, H. Ewald, A. Ladanie, M. Briel, B. Speich, H. C. Bucher, and L. G. Hemkens. Current use and costs of electronic health records for clinical trial research: a descriptive study. *CMAJ Open*, 7(1), 2019. doi: 10.9778/cmajo.20180096

[26] N. McCurdy, J. Gerdes, and M. Meyer. A framework for externalizing implicit error using visualization. *IEEE TVCG*, 2018. doi: 10.1109/tvcg.2018.2864913

[27] N. Menachemi and T. Collum. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy*, 4, 2011. doi: 10.2147/RMHP.S12985

[28] M. Müller, A. Baak, and H.-P. Seidel. Efficient and robust annotation of motion capture data. In *Proc. of SCA*, 2009. doi: 10.1145/1599470.1599473

[29] B. Muñoz, S. K. West, G. S. Rubin, O. D. Schein, H. A. Quigley, S. B. Bressler, K. Bandeen-Roche, and the SEE Study Team. Causes of blindness and visual impairment in a population of older americans: The salisbury eye evaluation study. *Arch Ophthalmol*, 118(6), 2000.

[30] T. Munzner. A nested model for visualization design and validation. *IEEE TVCG*, 15(6), 2009. doi: 10.1109/TVCG.2009.111

[31] B. P. Nicholson and A. P. Schachat. A review of clinical trials of anti-VEGF agents for diabetic retinopathy. *Graefes Arch Clin Exp Ophthalmol*, 248(7), 2010. doi: 10.1007/s00417-010-1315-z

[32] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: Using visualization to enhance navigation and analysis of patient records. In *Proc. of AMIA Symp*, 1998.

[33] W. Raghupathi and V. Raghupathi. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*, 2(1), 2014. doi: 10.1186/2047-2501-2-3

[34] M. Röhlig, C. Schmidt, R. K. Prakasam, H. Schumann, and O. Stachs. Visual analysis of retinal changes with optical coherence tomography. *The Visual Computer*, 34(9), 2018. doi: 10.1007/s00371-018-1486-x

[35] C. Schmidt, P. Rosenthal, and H. Schumann. Annotations as a support for knowledge generation - supporting visual analytics in the field of ophthalmology. In *Proc. of IVAPP*, 2018. doi: 10.5220/0006615902640272

[36] Y. Shahar, D. Goren-Bar, D. Boaz, and G. Tahan. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. *Artif Intell Med*, 38(2), 2006. doi: 10.1016/j.artmed.2005.03.001

[37] J. G. Stadler, K. Donlon, J. D. Siewert, T. Franken, and N. E. Lewis. Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards. *Big Data*, 4(2), 2016. doi: 10.1089/big.2015.0059

[38] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser. *On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics*. Springer, 2014. doi: 10.1007/978-3-662-43968-5_7

[39] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE TVCG*, 15(6), 2009. doi: 10.1109/TVCG.2009.187

[40] T. Wecker, C. Ehlken, A. Bühler, C. Lange, H. Agostini, D. Böhringer, and A. Stahl. Five-year visual acuity outcomes and injection patterns in patients with pro-re-nata treatments for AMD, DME, RVO and myopic CNV. *Br J Ophthalmol*, 101(3), 2017.

[41] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala. CommentSpace: Structured support for collaborative visual analysis. In *Proc. of CHI*, 2011. doi: 10.1145/1978942.1979407

[42] Y. Zhang, K. Chanana, and C. Dunne. IDMVis: Temporal event sequence visualization for type 1 diabetes treatment decision support. *IEEE TVCG*, 25(1), 2019. doi: 10.1109/TVCG.2018.2865076