Visual Computing for Big Data Analysis in Prostate Cancer Research

Jürgen Bernard[1,2], Thorsten May[1], Dirk Pehrke [3], Thorsten Schlomm [3], and Jörn Kohlhammer[1,2]

[1]Fraunhofer IGD, Darmstadt, Germany
[2]TU Darmstadt, Germany
[3]University Medical Center Hamburg-Eppendorf, Germany

Abstract

Data-centered research is becoming increasingly important in prostate cancer research where a long-term goal is a sound prognosis prior to surgery. We have developed a visual computing technology that contributes to this paradigm change in clinical research and practice for electronic health records (EHR) in this area. This visual-interactive system, developed in close collaboration with medical researchers, helps clinicians efficiently and effectively visualize single and multiple patient histories at a glance, create cohorts of patients for clinical tests, as well as generate and validate hypotheses.

1. Introduction

In the medical domain, the era of big data has the potential to change paradigms in working methods and procedures of physicians and data scientists. There is a rapidly growing interest and practice in using large amounts of patient-related data to advance clinical research and improve clinical quality. Clinics are currently building large patient databases in an effort to digitize as much information about patients, illnesses and treatments as possible. From a visual computing perspective, one of the core challenges in this direction is the development of easy-to-use interfaces and visualizations that allow clinical researchers and analysts to access clinical data bases easily, build and explore patient cohorts in a never seen effectiveness, build and validate hypotheses without effort, and prepare clinical studies efficiently. Recent visualization workshops like the annual VAHC [†] addressed specifically how visual analytics can help in this development. The community is convinced that visual-interactive interfaces are key to put this paradigm change into practice. The solutions lie in jointly developed ideas between visual computing and medicine, i.e., the expertise of both domains is required to create effective and efficient big data solutions.

Our work in this area is focused on cancer research and treatment, where longitudinal information about patient populations are used to further analyze causes, treatments and outcomes on specific cancer types. This entry to the Dirk-Bartz prize reports on a collaboration between experts in visual computing and clinical researchers, who specifically focus on prostate cancer. In this area, surgery is a highly important therapeutical method. However, the decision for or against surgery requires a trade-off between side effects and expected results. The long-term goal of clinical researchers is a sound prognosis of the cancer prior to surgery. For our project we had access to over 20,000 cases that included pre-surgery, histological, clinical, follow-up, and quality-of-life data.

An inherent aspect of such patient histories is that static data (like the age of a patient at the time of surgery) is coupled with time-dependent data. Moreover, it is often important for clinicians how quickly certain patient indications change, for example the pivotal PSA (prostate-specific antigen) value indicating prostate cancer. Furthermore, an effective analysis of cohorts requires the visualization of multiple patient histories at a glance to quickly gain an overview of time-dependent attributes characterizing the cohort of interest. Our result also offers a guided analysis between the current cohort and all static attributes, alleviating the process of building and evaluating in an integrated process.

The initial result of our collaboration was an understanding of their workflow and a visualization design that fits their mental model. We turned this workflow into an integrated visual analytics technique, which is used for cohort building and evaluation and is currently tested for use during patient consultation. Seeing a specific patient's data in relation to a cohort of similar patients helps to better judge potential treatments and outcomes for both doctors and patients.

2. Novelty of the Approach

Our work turns the search for patient stratifications (cohorts) into an integrated solution, which combines an overview over the data of possibly thousands of patients, free user interaction for the flexible selection of cohorts, and immediate feedback on any selection about the strength of potential correlations. This integration effec-

[†] http://www.visualanalyticshealthcare.org/, VA in Healthcare

DOI: 10.2312/egm.20171044

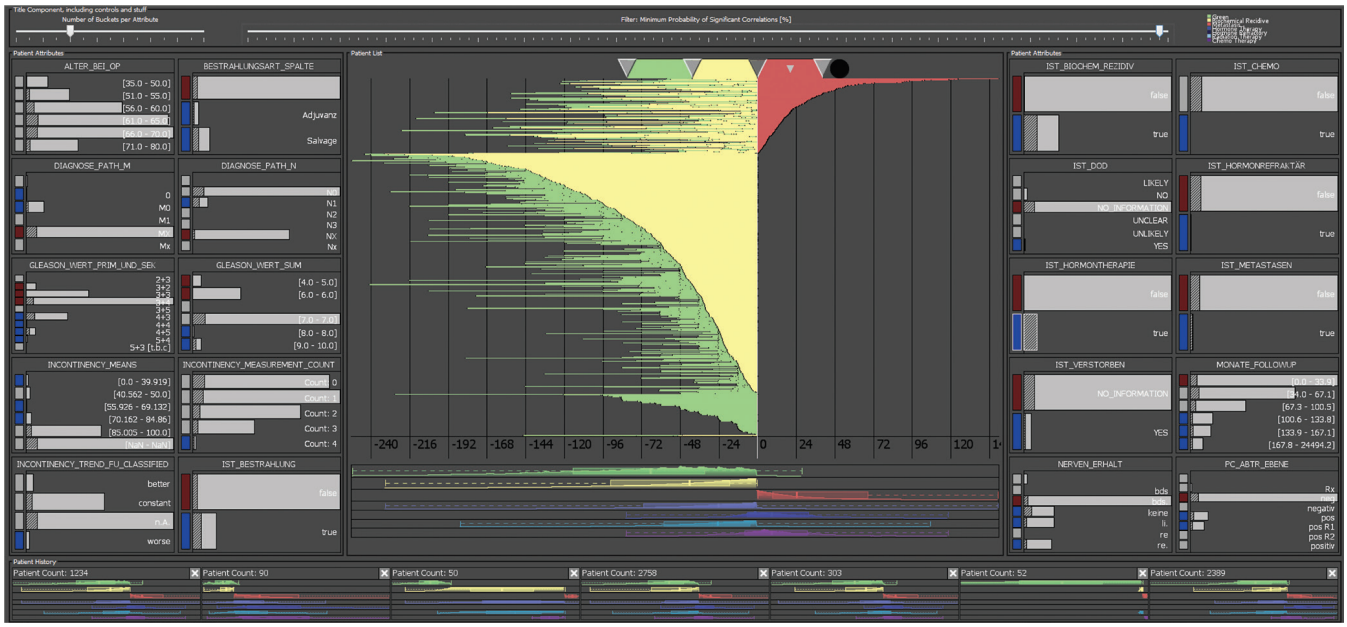

Figure 1: *The visual-interactive system for exploring prostate cancer patient records. It consists of three components. At the center, a list-based overview of temporal patient attributes is shown (green, yellow, and red). The selection history patient subsets is shown at the bottom (seven subsets). To the left and right, the distributions of static patient attributes are shown allowing for the selection of patient cohorts (see arrow). A statistical dependency measure highlights attribute observations related to the current patient selection (blue and red).*

tively closes a loop for the iterative refinement of stratifications. In current medical research the stratification is separated into disjoint steps. The review of individual EHR data and the compilation of patients to cohorts requires tedious and manual effort with no guarantee for significant results. Figure 2 demonstrates one current work practice. We reviewed this stratification workflow to identify opportunities to reduce these efforts. As a result, this allows for a broader evaluation of candidate stratifications.

Related techniques exist for different parts of our solution, but, to our knowledge, it is the first integrated approach to support stratification search. Ben Shneiderman and his colleagues presented seven challenges for visual analytics for healthcare [SPH13]. Our work deals with two of these challenges. First, the characterization and understanding of similarity, which is the basis for the search for patterns in the patient database. Second, visualizing comparative relationships aids the detection of potentially relevant data for prognosis. Rind et al. [RWA*13] compare 14 visualization systems for EHR by their ability to show multiple patient records, their primary datas, the number of variables and the supported user tasks. In terms of these properties, our system visualizes a mixture of snapshot, dynamic event, and measurement data for multiple patient records. Compared to other systems, our design goal was to maximize the number of variables and patients, by visual and statistical abstraction. One of the design goals by Zhang et al. [ZGP15] with the ability to define cohorts without any help of technology experts. We adopted this design goal, but we came to different conclusions. By limiting the number of visualization metaphors, we were able to offer a consistent frame of reference for all variables. Turkay et al. [TLS*14] combined StratomeX with statistical measurement for cohort analysis on the basis of numerical messenger RNA expression data. Much like our own approach, this extension shows the statistical significance of differences due to changes in the interactive selection. In comparison we use a generic goodness-of-fit test to cover nominal and categorized numerical data. Malik et al. [MDM*15] present an approach with the focus of comparing two cohorts based on event sequence data. The approach uses a number a metrics for comparison and to search for relevant properties.

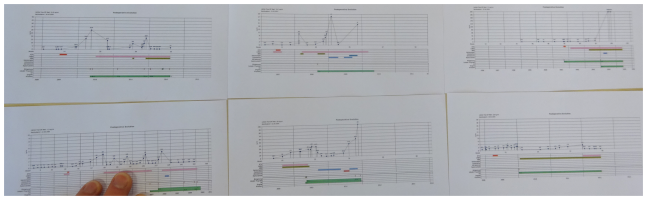

Figure 2: *Original visualization of patients history data showing temporal attributes of the follow-up phase. The current process to define patient cohorts is to go through the plots manually.*

Figure 3: *List of temporal attributes for the visualization of single patients. Colors reflect the common use in the medical domain. The list is sub-divided in biological and therapeutic attributes.*

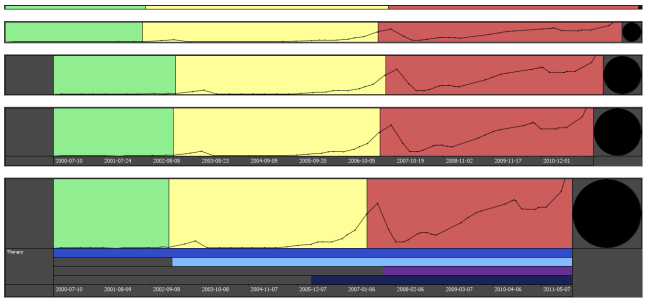

Figure 4: *Final design of a single patient visualization with five different levels of detail. Depending on the available space, the interface switches the LOD. At the bottom all patient attributes are shown in detail: hormone therapy (blue), radiation therapy (light blue), chemotherapy (purple), and hormone refractory (dark blue).*

3. Clinical Value in Prostate Cancer Research and Treatment

The major benefits that stem from our visual analytics technique are the following:

- Scalable overview of thousands of patients
- Interactive hypothesis building and validation
- Faster cohort selection
- Decision support in clinical practice
- Data quality assessment

The original goal and foremost benefit is for the urologists to expand their view from only the patients they personally treated to all patients that are in the database of the clinic. This allows the medical experts to find the patients most similar to a particular patient, but also to perform an exploratory search across all previous cases.

In our joint sessions with our medical partners, this immediately led to new hypotheses that were interactively evaluated with our tool. It should be mentioned that our tool is not intended to replace statistical methods, but to create hypotheses for subsequent statistical testing. However, the time for building hypotheses and cohorts is dramatically reduced from days to minutes. The tool has a built-in statistical dependency testing that covers all available variables and automatically hints at potentially interesting variable correlations. The medical researcher is then able to focus on these results.

A typical scenario is the definition of a cohort including patients of a similar disposition or therapy. A medical researcher may be interested in endpoints, which are reached particularly often by patients of this cohort. In addition, the result also points at potential co-factors or confounding factors, which have to be excluded from further statistical testing.

Another scenario is the definition of a cohort including patients that share a common endpoint. The most interesting variables in this cohort definition are those that are known at the time of the therapeutical decision. Finally, our method can be used for the effective assessment of already established cohorts. This becomes necessary, whenever new variables are introduced to the data. Any dependencies between existing cohorts and new variables offer opportunities for refinement, or for the effective identification of confounding factors.

These scenarios can be adapted to the analysis of and cohort building for other diseases that feature time-related and static attributes. The clinical value also extends to the decision support in clinical practice where the tool is used to explain to a patient his disease stage in relation to similar patients and their treatments and outcomes. Last but not least, our tool has proven to help in the data quality assessment of the patient history repository, where missing entries, wrong entries or implausible time sequences are quickly detectable.

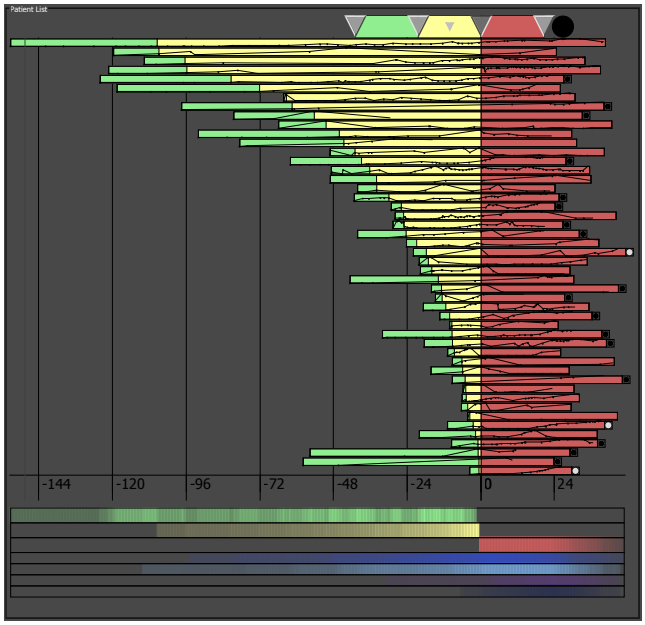

Figure 5: *Cohort of 48 patients all having metastases with a duration of 24 to 48 months. The list is synchronized to the event between phase 'yellow' (biochemical re-occurrence) and phase 'red' (metastasis), sorted by the duration of phase 'yellow'.*

4. Visual-Interactive Analysis of Prostate Cancer Cohorts

Our visual-interactive system for the analysis of patient cohorts is the result of a collaborative effort combining expertises in visual computing with medical research and clinical practice. The user-centered design principle allowed the creation of a usable and useful system that reflects the needs of the clinicians. We adopted a visualization design for single patients (see Figure 2) to offer a familiar point of reference in the new tool. Building on this, a comprehensive domain, data, and task characterization enabled us to create various visual-interactive techniques, assembled to a workable data analysis system [BSM*14, BSB*15, BSM*15]. The resulting system at a glance is presented in Figure 1.

4.1. Detail Visualization for Single Patients

Building on the original patient plot created by the clinicians (see Figure 2), a visualization of single patients depicts details about individual health care records. The visualization includes most relevant attributes as required by the medical experts (see Figure 3) and the most important synchronization points for the comparison of health progression. The design went through several iterations, beginning with paper prototypes and rapid prototypes. The final design of the visualization is shown in Figure 4, where we also

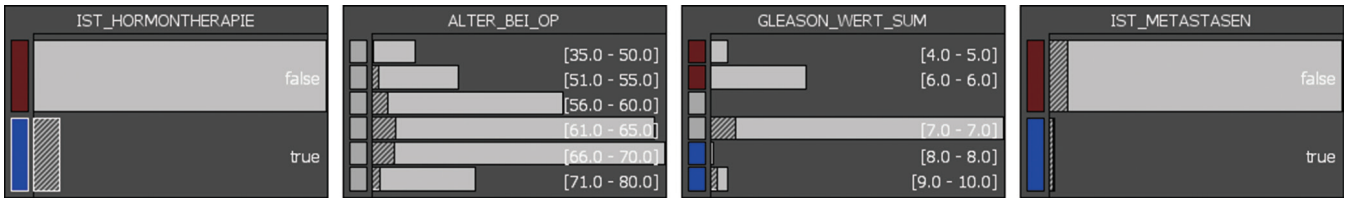

Figure 6: *Attribute views for the definition of cohorts and the guided analysis of relations. The selection of bars updates the cohort shown in the patient list. Currently, the 'true' bar of the left attribute (hormone therapy) is selected. Hatched areas indicate the current patient selection. Blue and red colors guide the user towards interesting relations in the data on the basis of statistical dependency measures.*

demonstrate the level-of-detail concept used to cope with space limitations. Accordingly, the size of the visualization can be reduced to a thin line allowing the assembly of thousands of patients in overview visualizations.

4.2. Overview of Thousands of Patients

Clinicians gain an overview of multiple patients with the list-based patient visualization depicted in Figure 5. The example shows 48 patient histories that all show the biological phases 'green', 'yellow', and 'red', reflecting good condition, recurrence of cancer, and metastases. The emerging pattern in this particular list of patients is a similar duration of phase red (metastases) between 24 and 48 months.

4.3. Browsing Through Numbers of Patient Records

Different interaction designs support clinicians in browsing through large numbers of patient records. According to the clinicians, most useful interactions are a) sorting patient histories with respect to the duration of biological phases (phase yellow in Figure 5), b) synchronizing patient histories with respect to state changes (phases yellow to red in Figure 5), and c) selecting subsets (patient cohorts) from the patient histories. The latter interaction technique particularly mitigates the information overload problem as a means to drill-down to patients of interest.

4.4. Visual-Interactive Cohort Analysis

Techniques for the analysis of patient cohorts are key to derive new knowledge from the underlying patient data base. The interaction techniques are one powerful way to achieve this. Another technique is the visualization of additional attributes about patients provided in the records. Figure 6 demonstrates the barchart visualization by example of four attributes. In this way, up to 100 patient attributes can be added to the analysis system. Every bar represents the number of patients associated to the particular measurement. Users can click any bar, the corresponding subset of patients is automatically loaded in the list view for detailed analysis. As an example, in Figure 6 the left barchart is selected showing all patients having hormone therapy ($IST\_HORMONTHERAPIE = true$). With the ability to select and visualize meaningful subsets of patients clinicians can now create patient cohorts within seconds and thereupon derive new knowledge easily.

4.5. Seeking Hidden Relations in Patient Attributes

Finally, the system reveals hidden relations in the data which can be used to derive and validate hypotheses. Depending on the current patient selection, blue and red colors in the attribute visualizations indicate associations with the patient cohort (see Figure 6). In the example, the selection of patients having hormone therapy reveals associations with the attributes 'GLEASON_WERT_SUM', and 'IST_METASTASEN'. The relation-seeking support is based on statistical testing functions, triggered by easy-to-use interfaces. In this way, we hide the complexity of the statistics, while tuning to mental model of the doctors. Clinicians can simply look-up new associations hidden in the data which build the basis for effective hypotheses generation and testing.

5. Recognition

Jürgen Bernard, who integrated the results in his PhD thesis [Ber15], was therefore awarded with the Hugo-Geiger price [‡] for excellent PhD theses within the entire Fraunhofer society.

References

[Ber15] BERNARD J.: Exploratory search in time-oriented primary data. Technische Universität Darmstadt, December 2015. Dissertation, PhD (Doktorarbeit). URL: http://tuprints.ulb.tu-darmstadt.de/5173/. 4

[BSB*15] BERNARD J., SESSLER D., BANNACH A., MAY T., KOHLHAMMER J.: A visual active learning system for the assessment of patient well-being in prostate cancer research. In *IEEE VIS Workshop on Visual Analytics in Healthcare* (2015), ACM, pp. 1–8. doi:10.1145/2836034.2836035. 3

[BSM*14] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A visual-interactive system for prostate cancer stratifications. In *IEEE VIS Workshop on Visualizing Electronic Health Record Data (EHRVis 2014)* (2014), IEEE Computer Society. 3

[BSM*15] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A visual-interactive system for prostate cancer cohort analysis. *Computer Graphics and Applications, IEEE 35*, 3 (2015), 44–55. doi:10.1109/MCG.2015.49. 3

[MDM*15] MALIK S., DU F., MONROE M., ONUKWUGHA E., PLAISANT C., SHNEIDERMAN B.: Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Intelligent User Interfaces (IUI)* (2015), ACM, pp. 38–49. doi:10.1145/2678025.2701407. 2

[RWA*13] RIND A., WANG T. D., AIGNER W., MIKSCH S., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B.: Interactive information visualization to explore and query electronic health records. *F&T in HCI 5*, 3 (2013), 207–298. doi:10.1561/1100000039. 2

[SPH13] SHNEIDERMAN B., PLAISANT C., HESSE B. W.: Improving healthcare with interactive visualization. *IEEE Computer 46*, 5 (2013), 58–66. doi:10.1109/MC.2013.38. 2

[‡] https://www.igd.fraunhofer.de/en/node/934

[TLS*14] TURKAY C., LEX A., STREIT M., PFISTER H., HAUSER H.: Characterizing cancer subtypes using dual analysis in caleydo stratomex. *Computer Graphics and Applications, IEEE 34*, 2 (Mar 2014), 38–47. doi:10.1109/mcg.2014.1. 2

[ZGP15] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization 14*, 4 (2015), 289–307. URL: http://dx.doi.org/10.1177/1473871614526077, doi:10.1177/1473871614526077. 2