EuroVis Workshop on Visual Analytics (2019)
C. Turkay and T. von Landesberger (Editors)

SurviVIS: Visual Analytics for Interactive Survival Analysis

A. Corvò, H.S. Garcia Caballero, and M.A. Westenberg

Eindhoven University of Technology, Eindhoven, The Netherlands

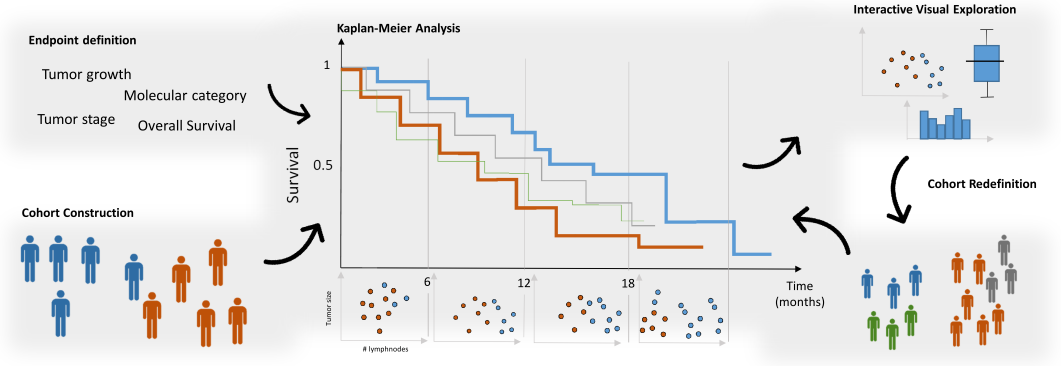

Figure 1: *An overview of the main workflow on SurviVIS. The user can start by defining one or more cohorts. The Kaplan Meier analysis is performed automatically and it guides the exploration of clinical data over successive time frames. Interaction on the different views allows the user to redefine the cohorts and explore the survival rates.*

Abstract

The increasing quantity of data in biomedical informatics is leading towards better patient profiling and personalized medicine. Lab tests, medical images, and clinical data represent extraordinary sources for patient characterization. While retrospective studies focus on finding correlations in this sheer volume of data, potential new biomarkers are difficult to identify. A common approach is to observe patient mortality with respect to different clinical variables in what is called survival analysis. Kaplan-Meier plots, also known as survival curves, are generally used to examine patient survival in retrospective and prognostic studies. The plot is very intuitive and hence very popular in the medical domain to disclose evidence of poor or good prognosis. However, the Kaplan-Meier plots are mostly static and the data exploration of the plotted cohorts can be performed only with additional analysis. There is a need to make survival plots interactive and to integrate potential prognostic data that may reveal correlations with disease progression. We introduce SurviVIS, a visual analytics approach for interactive survival analysis and data integration on Kaplan-Meier plots. We demonstrate our work on a melanoma dataset and in the perspective of a potential use case in precision imaging.

CCS Concepts

• ***Applications*** → *Visual Analytics;*

1. Introduction

The recent advances in biomedical informatics, software and algorithmic power have provided clinicians and researchers with large volumes of data that characterize individuals in great detail. Typically, individuals become patients when a certain event of interest occurs. For instance, the onset of a disease or a first screening at the hospital are common situations. After that, the clinical course of the patient is accompanied by test results, image acquisitions and follow-ups that provide insights on the disease and the treatment effects. The collection of such clinical information has turned to be extremely valuable in observing the prognosis of groups of patients (cohorts) that share similar characteristics. An intuitive way to estimate the disease risk is represented by the Kaplan-Meier (KM) curve [RNP*10]. The KM curve derives from the published meth-

DOI: 10.2312/eurova.20191128

ods by Kaplan and Meier [KM58] and it is broadly used in the medical field. Besides the large variety of quality in these plots [Bol03], the KM curve remains easy to read and interpret. The typical use of a KM curve is to estimate the mortality of cohorts with respect to prognostic factors (*e.g.*, tumor stage, overall survival, specific pathological conditions, etc.). However, KM curves are only visualized and interpreted in a static way despite their potential to describe disease progression over time in much more detail. In the last years, many visual analytics approaches have showed that interactive graphs of temporal data can help the user explore cohorts and clinical information [LPK*15,BSM*15,WGP*11]. In a similar way, interactive KM curves can provide more information on patient mortality and lead to insights. These aspects make KM curves an appealing visualization that may benefit visual analytics tools in the medical domain.

In this paper, we propose a visual analytics approach for interactive survival analysis by means of KM curves as the reference plotting system. We conveyed our concept in SurviVIS, an approach for exploration and analysis of survival data.

2. Background

Many visualization approaches for exploration of clinical records and analysis of longitudinal data of patients have been presented in the last years. Most of them focus on the construction of cohorts by enabling interactive selection of specific clinical information. CAVA [ZGP14] and COQUITO [Jos16] are some examples that successfully support the reasoning process of researchers. In the last years, medical data has become even more complex as it comes from many different sources and in vast quantities [DZAD17, CG15]. For instance, cancer-research is an expanding field where data comes from many medical imaging modalities, algorithms, molecular and genomic tests. In this field, Bernard et al. [BSM*15] presented a visual-interactive system to support clinicians to investigate prostate cancer patients. Event data of the patients is explored together with a large amount of features in the process of biomarker discovery. At the same time, advanced algorithms are generating hundreds of features from medical images. The increasing evolution of such algorithms initiated the field of radiomics [LRVL*12] and computational pathology [LFC*16] that promise to unveil new biomarkers and to lead towards better diagnosis and treatment. Nevertheless, getting insights from automatically extracted features is becoming a hard task, and visual analytics can be valuable as shown by Klemm et al. [KOJL*14]. In their work, integration of automatically extracted image data and clinical records in a visualization helped epidemiologists in hypothesis generation and validation. Yu et al. [YJY*17] address the field of radiomics with iVAR, an interactive visual analytics system for large-scale medical image exploration. It supports statistical analysis in a multi-view dashboard that empowers researchers with more means to study patient prognosis. In the work of Lex et al. [LSS*12], KM plots are used to provide a static overview of the survival rates of cancer subtypes. However, no interaction is provided.

Surprisingly, survival curves are often not available in these visual interactive dashboards despite their ordinary use in the clinical domain. In our work, we present an approach to exploit the KM-curves for characterization of patients and feature exploration.

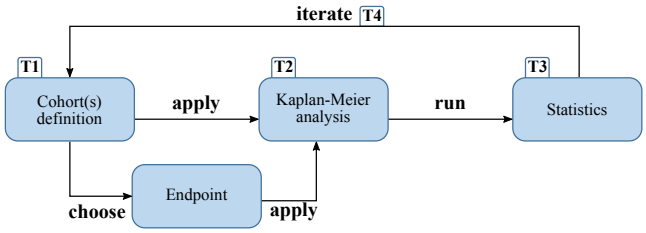

Figure 2: *The typical workflow of a researcher on survival analysis. After defining one or more cohorts, the KM analysis is conducted. Statistical methods are applied to extract insights from the curve data points.*

3. Data and Problem definition

On the basis of a literature review and the indication of the work of Bollschweiler et al. [Bol03], we define the data involved and the problem definition. The data used for survival analysis includes retrospective datasets of large records of patients. Each patient must be characterized by the following information:

- Serial time.
- Status at serial time (1=event of interest; 0= censored).
- Study group (e.g., Group A, B, etc.).

However, the definition of the group becomes complex in studies where many variables are available and the combination of them may lead to many relevant subgroups. A typical workflow for survival analysis involves the tasks depicted in Figure 1. The researcher begins the study by defining one or more cohorts **(Task 1)**. Usually, the first operation is to perform a KM analysis **(Task 2)** on the entire population to obtain an *overall survival* plot. In *overall survival* curves, the event of interest is death from any cause [RNP*10]. The researcher can look at different types of curve by selecting a different event of interest (*e.g.* disease onset, relapse). For instance, *Disease free survival curves* are generated by considering the relapse as the event of interest [RNP*10]. After generating the KM curves, researchers typically use conventional statistical methods such as the *Log-rank test* and the hazard ratio calculation between two curves [RNP*10] **(Task 3)**. These two methods indicate whether or not there are differences in the survival rate of two groups and which group has a more favorable survival rate. Then, the researcher can decide to focus on one particular group and to redefine the cohort (**Task 4**) according to specific clinical variables (*e.g.* tumor stage, lab tests)

We list a set of problems to address when using KM curves:

P1 **No data integration**. The focus is typically on generating the survival curve for the identified cohort. However, variation and redefinition of the cohort by filtering variables is a manual task.

P2 **Lack of interaction**. Survival curves are used to generate evidence in clinical papers, but standard libraries (*e.g.*, [Cam19]) do not provide interaction.

P3 **Poor consistency**. As Bollschweiler [Bol03] states: "*mistakes and distortions frequently arise in the display and interpretation of survival plots*".

From these problems, we derived the requirements for a proto-

type that leverages survival curves for interactive analysis of clinical data.

4. SurviVIS

To design our approach, we collected the statements presented in [Bol03]. The central part of SurviVIS is the KM plot (Fig. 5.4). We use the fact that the chart is typically divided in time frames given by the steepness of the curve to link the other views (Fig. 3). We decide to use the time frames to provide more insights into the distribution of the clinical variables in that specific time range.

SurviVIS consists of four main components (see Fig. 5). We provide the dataset information (Fig. 5.1) regarding the population of the study such as the number of patients, death events and censored cases. After data import, SurviVIS automatically computes the KM plot on the basis of the data fields provided. Also, we indicate the median survival in the dataset information.

Variable List. Variables are automatically classified as categorical or numerical and shown in a list (Fig. 5.2) where small charts shows the distribution of the data. We present them sorted by their skewness. This type of statistics can lead the user to select the variables that may discern the population better **(Task1)**. A selected variable can be plotted on one of the three views of SurviVIS . A click on the first circle (Fig. 4) plots the data on the 1D-variable view. A click on the KM rectangle will show the data overlaid on the KM curve. The last circle respectively will combine the variable with another selected variable in the bottom view (Fig. 5.5).

Kaplan-Meier Plot. We preserve the main elements of standard survival curves to perform **Task 2**. The cumulative probability of surviving a given time is shown on the Y-axis. The x-axis is divided in serial times that are also used as guides for the other visualizations of the tool (Fig. 6). Users can select a maximum of five serial times to generate the time windows. More windows may lead to confusion and cluttered view. A vertical bar in the KM plot enables the user to split the curve into two halves. This helps the user to redefine the cohort, resulting in an automatic update of the rest of the views that display corresponding data just for the two halves. The distribution of a selected variable can be plotted in dedicated box-plots directly on the KM plot area. Each time frame includes a box-plot (see Fig. 5.4) with the data distribution of the subcohort in that range.

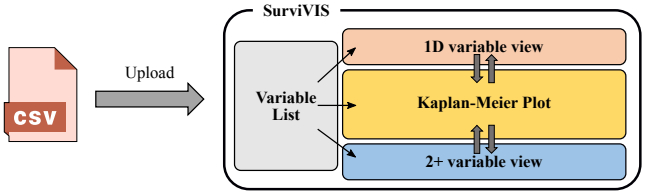

Figure 3: *An overview of our system. A csv file can be imported in SurviVIS by defining the main variables needed to create the KM Plot. The curve is then automatically generated. By selecting the variables in the left side bar, the 1D variable view and the 2+ variable view are shown. The three views are linked with dedicated interaction.*

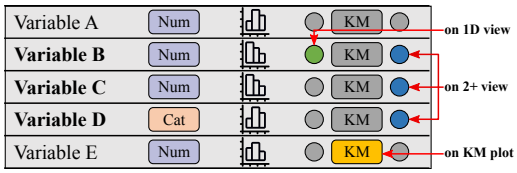

Figure 4: *The variable list presents patient data discriminated as categorical and numerical. A small chart displays the data distribution. Variables can be sorted by their skewness. A click on the circles or the KM label enables the user to plot the data in the corresponding view of SurviVIS.*

1D-variable view. The 1D-variable view (Fig. 5.3) represents the data distribution of numerical or categorical data divided by the corresponding time frames of the *KM Plot*. The user can choose to visualize the data of the patients for which an event occurred (death/recurrence) or not (alive). An example in cancer research would see the user to assess the distribution of the tumor stage for the sub-cohorts of patients in time frames of one year (Fig.6). For early death, high tumor stage categories should be predominant.

2+ variable view. This view (Fig. 5.5) is dedicated to the data exploration and allows the user to explore multiple variables with respect to the time frames given by the survival plot. As the previous plot, it provides insights for cohort redefinition **(Task 4)** . By selecting two numeric variables from the *variable list* the user is automatically provided with a succession of scatterplots below the *KM Plot*. By selecting an additional categorical variable, the user is given with an enriched version of the *2+ variable view*.

5. Use case: Malignant Melanoma data

To demonstrate the usefulness of SurviVIS, we use a publicly available dataset from [ABGK93] on 205 Malignant Melanoma patients. Each patient had their tumor removed by surgery. Among the measurements taken were the thickness of the tumor and whether it was ulcerated or not. These information were shown to indicate an increased chance of death from melanoma [BBSC*14].

The *overall survival* of the population is shown with an orange line in the central part of SurviVIS (Fig. 5). By clicking on the cohort splitting option for the ulcer variable in the *variable list*, two subgroups are generated: patients with ulcerated tumor and without **(Task 1)**. The KM curve **(Task 2)** for the first group shows a more favourable survival rate than for the second group (Hazard Ratio = 1.89) **(Task 3)**. The user can investigate the thickness of the tumor. The measurements are visualized as boxplots (Fig. 5) integrated in the KM curve. Contrary to the hypothesis that high thickness values could be associated with lower prognosis, the researcher can observe the data in the successive time frames and notice that thicker tumors seem to be present for long survivals. The user can then select two numerical variables (age and tumor thickness) that are plotted on the scatterplots. By adding the information on the ulcer status, the user observes that ulcerated tumors are actually characterized by lower values in tumor thickness after four years. Hence, the user decides to look at the time range between 4 and 6 years. The provided interactions on the scatterplot allow the user to zoom

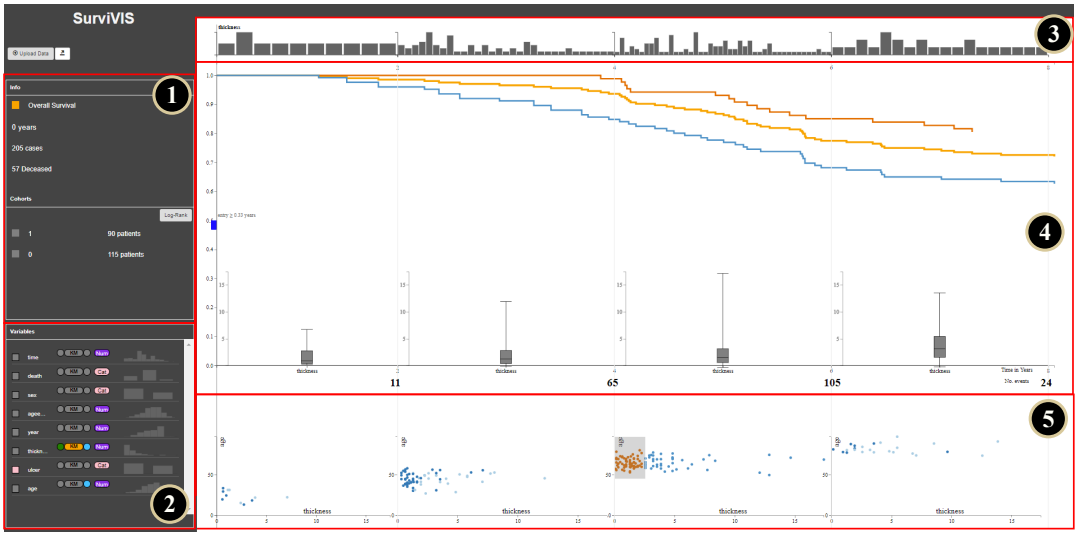

Figure 5: *The interface of SurviVIS with the main components and data from the Malignant Melanoma dataset. The researcher is interested in the thickness of the tumor. On the left sidebar, an overview of the dataset (1) and a list of variables (2) are shown. The thickness variable distribution is shown for the patients of each time window (3) of the main KM plot (4). Two additional curves (orange and blue) are generated by selecting two groups of patients in the scatterplot (5).*

to the data points of the specific time frame of the overall survival curve. The user selects a specific group of patients ((**Task 4**).) to restrict the variable range to the size of vertices of the rectangle (Fig. 6). These last steps would have required several manual iterations in standard survival analysis. This kind of exploration can trigger hypothesis generation and new observation for further analysis and better patient stratification.

6. Generic use case

A generic use case for SurviVIS is represented by the field of precision imaging in medical research [FTG16]. In the specific, radiomics [LRVL*12] and computational pathology [LFC*16] are

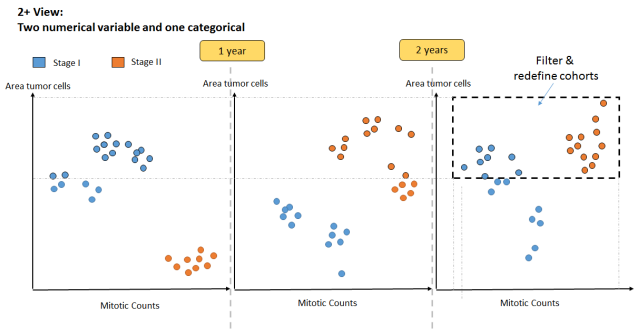

Figure 6: *A scenario of pathology image based features: The 2+ view can be used to inspect the survival rate with respect to quantitative image-based features (e.g. mitotic counts and tumor cells area) and the tumor grading. Similarly, other untapped prognostic factors could be explored in an interactive way.*

two fields of research that generate large quantity of image-based features and opportunities for research. Many studies already demonstrated the relevance of computing morphological and architectural features for biomarker discovery [DZAD17]. The common approach is to explore and analyze automatically extracted features with respect to patient mortality, disease progression or classification [VPvDV14, FTG16, GKH16]. Because of the large quantity of features, exploration may be complex and tedious [LFC*16,KI18,TP18] that could be eased with visual analytics. For instance, in pathology we can imagine to explore the information derived from image analysis on the mean area of the tumor cells and the mitotic count (indicator of tumor proliferation [VPvDV14]) over successive time frames (Fig. 6). Subgroups of patients could be identified and investigated with respect to other clinical variables (*e.g.* molecular tests). In light of this, we aim at facilitating the current workflow and at obtaining new insights by using SurviVIS.

7. Conclusion

In this paper, we presented a first approach to make use of KM curves for interactive survival analysis and clinical data exploration. SurviVIS combines several charts arranged with the time frames of the KM plot. We designed a general interactive approach to explore numerical and categorical variables in multiple ways. We illustrated a use case where the user explores a malignant melanoma dataset to investigate standard diagnostic and prognostic factors. More options to redefine the cohorts, filtering the data and to track data provenance can quickly increase the experience on SurviVIS.

References

[ABGK93] ANDERSEN P. K., BORGAN Ø., GILL R. D., KEIDING N.: *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US, New York, NY, 1993. 3

[BBSC*14] BØNNELYKKE-BEHRNDTZ M. L., SCHMIDT H., CHRISTENSEN I. J., DAMSGAARD T. E., MØLLER H. J., BASTHOLT L., NØRGAARD P. H., STEINICHE T.: Prognostic Stratification of Ulcerated Melanoma. *American Journal of Clinical Pathology 142*, 6 (dec 2014), 845–856. 3

[Bol03] BOLLSCHWEILER E.: Benefits and limitations of Kaplan?Meier calculations of survival chance in cancer surgery. *Langenbeck's Archives of Surgery 388*, 4 (sep 2003), 239–244. 2, 3

[BSM*15] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A Visual-Interactive System for Prostate Cancer Cohort Analysis. *IEEE Computer Graphics and Applications 35*, 3 (may 2015), 44–55. 2

[Cam19] CAM DAVIDSON PILON: Lifelines. https://lifelines.readthedocs.io/en/latest/, 2019. Online; accessed 3 March 2019. 2

[CG15] CABAN J. J., GOTZ D.: Visual analytics in healthcare - opportunities and research challenges. *Journal of the American Medical Informatics Association 22*, 2 (mar 2015), 260–262. 2

[DZAD17] DJURIC U., ZADEH G., ALDAPE K., DIAMANDIS P.: Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology 1*, 1 (dec 2017), 22. 2, 4

[FTG16] FRANGI A. F., TAYLOR Z. A., GOOYA A.: Precision Imaging: more descriptive, predictive and integrative imaging. *Medical Image Analysis 33* (oct 2016), 27–32. 4

[GKH16] GILLIES R. J., KINAHAN P. E., HRICAK H.: Radiomics: Images Are More than Pictures, They Are Data. *Radiology 278*, 2 (feb 2016), 563–577. 4

[Jos16] JOSUA KRAUSE, ADAM PERER H. S.: Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 91–100. 2

[KI18] KOMURA D., ISHIKAWA S.: Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal 16* (jan 2018), 34–42. 4

[KM58] KAPLAN E. L., MEIER P.: Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 53*, 282 (jun 1958), 457. 2

[KOJL*14] KLEMM P., OELTZE-JAFRA S., LAWONN K., HEGENSCHEID K., VOLZKE H., PREIM B.: Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (dec 2014), 1673–1682. 2

[LFC*16] LOUIS D. N., FELDMAN M., CARTER A. B., DIGHE A. S., PFEIFER J. D., BRY L., ALMEIDA J. S., SALTZ J., BRAUN J., TOMASZEWSKI J. E., GILBERTSON J. R., SINARD J. H., GERBER G. K., GALLI S. J., GOLDEN J. A., BECICH M. J.: Computational Pathology: A Path Ahead. *Archives of Pathology & Laboratory Medicine 140*, 1 (jan 2016), 41–50. 2, 4

[LPK*15] LOORAK M., PERIN C., KAMAL N., HILL M., CARPENDALE S.: TimeSpan: Using Visualization to Explore Temporal Multi-Dimensional Data of Stroke Patients. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 1–1. doi:10.1109/TVCG.2015.2467325. 2

[LRVL*12] LAMBIN P., RIOS-VELAZQUEZ E., LEIJENAAR R., CARVALHO S., VAN STIPHOUT R. G., GRANTON P., ZEGERS C. M., GILLIES R., BOELLARD R., DEKKER A., AERTS H. J.: Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer 48*, 4 (mar 2012), 441–446. 2, 4

[LSS*12] LEX A., STREIT M., SCHULZ H., PARTL C., SCHMALSTIEG D., PARK P. J., GEHLENBORG N.: Stratomex: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. *Comput. Graph. Forum 31*, 3 (2012), 1175–1184. 2

[RNP*10] RICH J. T., NEELY J. G., PANIELLO R. C., VOELKER C. C. J., NUSSENBAUM B., WANG E. W.: A practical guide to understanding Kaplan-Meier curves. *Otolaryngology–head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery 143*, 3 (sep 2010), 331–6. 1, 2

[TP18] TIZHOOSH H., PANTANOWITZ L.: Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics 9*, 1 (2018), 38. 4

[VPvDV14] VETA M., PLUIM J. P. W., VAN DIEST P. J., VIERGEVER M. A.: Breast cancer histopathology image analysis: a review. *IEEE transactions on bio-medical engineering 61*, 5 (may 2014), 1400–11. 4

[WGP*11] WONGSUPHASAWAT K., GÓMEZ J. A. G., PLAISANT C., WANG T. D., TAIEB-MAIMON M., SHNEIDERMAN B.: Lifeflow: visualizing an overview of event sequences. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011* (2011), pp. 1747–1756. 2

[YJY*17] YU L., JIANG H., YU H., ZHANG C., MCALLISTER J., ZHENG D.: iVAR: Interactive visual analytics of radiomics features from large-scale medical images. In *2017 IEEE International Conference on Big Data (Big Data)* (dec 2017), IEEE, pp. 3916–3923. 2

[ZGP14] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization* (2014), 1473871614526077. 2