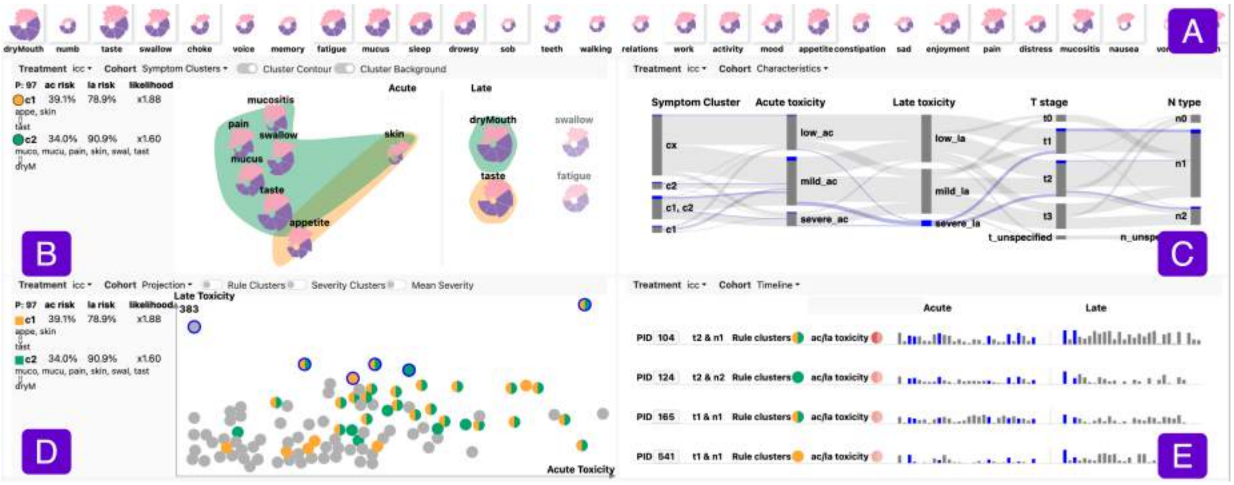
Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining

C. Floricel, A. Wentzel, A. Mohamed, C.D. Fuller, G. Canahuate, and G.E. Marai



P: 97	ac risk	la risk	likelihood
c1	39.1%	78.9%	x1.88
appe, skin tast			
c2	34.0%	90.9%	x1.60
muco, mucu, pain, skin, swal, tast dryM			

Fig. 1: Longitudinal symptom analysis and prediction for head and neck patients (ICC treatment). A) Overall severity over time for each symptom, across treatments. B) Sequential mining component, showing two clusters that use *acute* symptoms (left) to predict *late* symptoms (right). Lower opacity indicates other *late* prevalent symptoms, not selected by the current model. C) Cohort characteristics, showing symptom cluster results against patient attributes. D) Scatterplot showing patients projected based on the total symptom score for *acute* (X axis) and *late* (Y axis) stages. E) Cohort timeline, displaying cluster labels, clinical details, and mean symptoms burden.

Abstract—Personalized head and neck cancer therapeutics have greatly improved survival rates for patients, but are often leading to understudied long-lasting symptoms which affect quality of life. Sequential rule mining (SRM) is a promising unsupervised machine learning method for predicting longitudinal patterns in temporal data which, however, can output many repetitive patterns that are difficult to interpret without the assistance of visual analytics. We present a data-driven, human-machine analysis visual system developed in collaboration with SRM model builders in cancer symptom research, which facilitates mechanistic knowledge discovery in large scale, multivariate cohort symptom data. Our system supports multivariate predictive modeling of post-treatment symptoms based on during-treatment symptoms. It supports this goal through an SRM, clustering, and aggregation back end, and a custom front end to help develop and tune the predictive models. The system also explains the resulting predictions in the context of therapeutic decisions typical in personalized care delivery. We evaluate the resulting models and system with an interdisciplinary group of modelers and head and neck oncology researchers. The results demonstrate that our system effectively supports clinical and symptom research.

Index Terms—Temporal Data; Life Sciences; Mixed Initiative Human-Machine Analysis; Data Clustering and Aggregation

1 Introduction

Recent advancements in oncology have resulted in a greater variety of personalized cancer treatment outcomes for head and neck cancer (HNC) patients. Despite the increase in survival outcomes ("roses"), in many patients treatment leads to side effects that can greatly affect quality of life even after the completion of treatment ("thorns"). These symptoms can often be mitigated through preventative therapies, but the preventative treatment can also be an additional burden to patients. Thus, there is a growing interest in understanding how symptoms develop, stratifying patients into high-risk and low-risk cohorts, and studying the relationship between symptoms and treatment decisions, with an effort on identifying long-term symptoms that affect the patient's quality of life.

In HNC, identifying symptom risk is particularly challenging due to the composite effects of specific treatments and clinical factors [51]. Furthermore, some symptoms are correlated, either due to direct influence, or by shared root causes. These factors make predicting treatment outcomes difficult, and hamper personalized care decision making and delivery. Thus, there is a need for alternative human-machine analysis tools that can leverage computational and human effort to help modelers better understand HNC symptoms.

Current computational symptom research is focused on symptom clustering [29,36], however, there is little work [23] to explore symptom patterns across time or to compare the outcomes of different treatments. Sequential rule mining (SRM) is a promising unsupervised learning approach for discovering common temporal patterns in symptom data, but it can produce many repetitive, or even misleading results for predicting outcomes. Our work uses SRM modeling in combination with other unsupervised machine learning (ML) methods to predict treatment-related toxicities. At the same time, the model results have to also make sense in a clinical setting, and so they need to be interpreted

- *C. Floricel, A. Wentzel, and G.E. Marai are with the University of Illinois Chicago. E-mail: cflori3@uic.edu | gmarai@uic.edu.*
- *G. Canahuate is with the University of Iowa.*
- *A. Mohamed and C.D. Fuller are with the MD Anderson Cancer Center at the University of Texas.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

by domain experts with real patient data. Beyond helping modelers, visual analysis can further help with model interpretation in the context of clinical patient data.

Visual computing with temporal symptoms has several challenges. First, the large size of patient cohort, number of symptoms, and time-points requires scalable encodings, as well as meaningful aggregation techniques. Second, interpreting symptom trajectories in a clinical setting requires access to secondary clinical features for the cohort. Third, because domain experts are interested in identifying which symptoms are caused by treatments or other symptoms, a visual system needs to allow comparison between symptom groups and between treatments. Fourth, since the interpretation of association structures requires both data mining and clinical expertise, the systems need to allow for multiple workflows and levels of details to analyze both symptom and patient sub-cohorts. Finally, drawing conclusions from high-dimensional cohort data requires the use of interpretable algorithms, such as rule mining, to help extract patterns that are both useful and simple.

To address these challenges, we introduce a visual computing system to support the analysis of treatment-related toxicities and to predict post treatment symptoms based on during treatment symptoms. Our system uses an unsupervised, multivariate method that incorporates sequential rule mining, hierarchical clustering, and factor analysis to assess temporal interrelationships among multiple symptoms in the context of personalized care delivery. Our main contributions are: 1) a description of the modeling problem, data, and tasks; 2) a hybrid human-machine approach for identifying symptom profiles in HNC patients, stratified by treatment methods; 3) the design and implementation of this approach in a system which allows for the exploration of HNC cohort data at both the symptom and patient level, with an emphasis on capturing longitudinal patterns in symptom and patient cohorts; 4) a clinically-validated evaluation by domain experts; 5) the lessons learned from this multidisciplinary collaboration.

2 Related Work

Patient Cohort and Clinical Pathway Visualization. Visual analysis for patient cohorts often relies on finding connections between different patient attributes from medical records. This implies human interpretation of patterns within heterogeneous, and even multidimensional clinical information from patient records. In explainable AI (XAI) medical applications, cohort analysis tackles clinical statistics from patient records [30, 66], cohort history comparison [5, 12, 70], cohort medical image attribute comparison [10, 37, 44, 54, 62], or survival risk analysis [43]. The use of visual encodings vary largely among these applications, from custom histograms [4], to time-series plots [26, 35], matrices [18, 41], and radial charts [27]. When working with large cohorts where the focus is on finding outlier patients and understanding why they are showing unexpected clinical attributes, scatterplot projections are a common way to interpret cohort clusters [19, 22, 46, 47, 64]. Similarly, we use scatterplots for cohort interpretation, however we customize these plots to capture multivariate patient attributes, while also supporting treatment comparison.

Patient longitudinal medical records data are often visualized using clinical pathway summaries for individual patients [9], or cohort temporal summaries for cohorts [67]. Visual abstractions for temporal cohort data have mostly used matrix-based [18], flow-based representations [26, 66], or timelines [4, 28, 53]. Tree-based representations have been used for event sequence summarization, ordering, and statistics in temporal, clinical data [41, 49, 60, 67]. Other systems have used PCPs or flow-based representations with line bundling [4, 47]. While we adapt some of these encodings, we also support cohort summaries of both temporal and categorical attributes.

A popular method for visual temporal cohort analysis focused on clinical event sequences is sequential pattern mining [11, 16, 59]. However, sequential pattern mining can be misleading as there is no assessment of the probability that a pattern will be followed. In contrast, our proposed work uses sequential rule mining (SRM), which takes into account the probability that a temporal pattern will be followed.

Rule Visualization. Rule-based modeling is a common approach for creating explainable models [38, 68]. In XAI, rule-based explanations are often used to interpret black-box models such as neural networks [48], support vector machines [45], and latent factor models [52]. Rules have been adopted in medical data visualization as well, with applications in clinical risk prognosis [3, 39] and disease or treatment toxicity prediction [23, 48, 57, 63]. Surveys and recent visualization systems have shown that rule sets are usually visualized using node-links, tree-based representations, matrices, scatterplots, or PCPs [8, 31, 34, 71]. In a further departure from previous work, our approach combats scalability issues for large rule sets, and emphasizes the temporal separation between the rule antecedent and consequent.

Alongside rule sets items, visualization systems have to also integrate relevant rule metrics such as the support and confidence to denote the relevance of the rules. Yuan et al. [69] found that feature alignment and predicate encoding are influential visual factors for representing rules, arguing that the interpretability and decision making process are highly influenced by the different rule structures. Applications that support rule itemsets and rule metrics explanation in disease progression have used matrix-based representations accompanied by barcharts and tree-based circular glyphs [3, 48], while others employed node-links to represent temporal rules from diagnosis codes [50]. Our previous rule mining work [23] explored symptom associations in a given treatment stage independently (*acute* or *late*), and could not capture dependency between stages. In this work we tackle a different modeling problem and XAI challenges, where we focus on late symptoms, we model the symptom burden evolution sequentially, we apply rule clustering to reduce complexity and tackle scalability, and we support per-treatment sub-cohort analysis.

3 Background

HNC treatment can involve surgery, radiation treatment, induction chemotherapy, radiation and chemotherapy together, or a combination of these treatments. These treatment modalities often result in symptom burden both during the treatment period, and even after the completion of treatment [20], i.e., "roses have thorns". The MD Anderson Cancer Center documents and quantifies these symptoms through a standardized monitoring program based on MDASI (MD Anderson Symptom Inventory) [13], a patient-reported outcome measure for clinical and research use. The program uses questionnaires that are collected weekly at the time of the treatment appointment (acute stage), and at longer intervals post treatment (*late* stage), during cancer recurrence monitoring.

Counter-intuitively, the two-stage monitoring frequency is not driven by current modeling interests: clinicians have a good understanding of symptom values and trends during treatment (*acute* stage), which are sampled with higher-resolution, but not of the after-treatment symptoms (*late* stage), collected at lower-resolution. Data sampling is in fact constrained by the standard of care practice, which is targeted at detecting cancer recurrence based on recommended guidelines while reducing the patient burden of required clinic visits, and the fact that in-clinic questionnaires yield higher reliability and patient compliance than at home self-reports.

Since cancer patients can experience a multitude of symptoms that can co-occur or can cause other symptoms, oncologists are interested in modeling clusters of frequently co-occurring symptoms and in how symptoms are correlated with the diagnosis [61, 65] and prescribed treatment [2, 17, 21, 58, 63]. However, existing research does not focus on the temporal association between symptoms, changes in symptom severity over time, or the prediction of post-treatment symptoms. Current approaches that analyze the association between symptoms include methods such as factor analysis (FA) [55, 56], hierarchical cluster analysis (HCA) [29], latent class profile analysis [32, 36], and rule mining [6, 23]. In this work, we study how *acute* symptoms predict *late* symptoms, using a combination of SRM and HCA.

4 Design

This project is part of a multiyear, interdisciplinary collaboration between research groups with cancer symptom modeling experience from three research sites, composed of: three radiation oncology experts with clinical and research experience, one senior data mining expert, one senior visual computing expert, and several junior researchers in visual computing. The team held weekly remote meetings to discuss various clinical data analyses, during which our visual computing research group collected feedback.

Our design process followed an Activity-Centered-Design (ACD) approach [42], focusing on user activities and workflows. This paradigm has shown higher success than traditional human centered design for scientific, interdisciplinary collaborations. In this work, we used ACD to build workflows around the evaluation of clinically applicable models and complementary clinical data analysis.

The visual computing and data mining research groups met weekly to define functional specifications, prototype the interface for the clinically applied models, and evaluate the interface. This was an interactive process that, following the ACD approach, proved to be effective in the context of this remote collaboration [42]. Alternative designs for visual prototypes are available in the supplemental materials.

4.1 Activity and Task Analysis

Our system serves model builders in cancer symptom research. Our collaborators have experience in ML approaches for symptom analysis, but were interested in alternative approaches for temporal analysis that are centered on exploring the differences between patients receiving different treatment modalities, with particular emphasis on *late* symptoms. There was also a need to efficiently present and interpret the results from the proposed model to our clinician collaborators. Additionally, it was imperative to compare the toxicities found by the model across different treatment groups. Based on these considerations, and following the ACD paradigm, we split the requirements for this project into two main activities and we list their corresponding tasks:

A1 Support temporal symptom analysis for a given treatment

- **T1.1** Predict *late* symptoms based on *acute* symptoms
- **T1.2** Identify temporal patterns in the overall symptom severity
- **T1.3** Correlate clinical cohort details and symptom patterns
- **T1.4** Facilitate the analysis of a subset of patients within a cohort

A2 Support temporal symptom analysis across multiple treatments

- **T2.1** Compare temporal symptom profiles across treatments
- **T2.2** Evaluate the likelihood of experiencing a symptom profile compared to alternative treatments
- **T2.3** Identify temporal patterns in severity across treatments
- **T2.4** Facilitate the comparison of clinical patient data for multiple treatments

Our evaluation describes examples of preferred workflows concentrated on these activities, while the results are clinically validated by oncology domain experts. Non-functional requirements included clarity in the model results, scalable visualizations that can display symptom and patient statistics, and intuitive visual abstractions.

4.2 Data

The data used for building the proposed work is from a cohort of 823 HNC patients that underwent treatment at the MD Anderson Cancer Center in Houston, TX. Demographic and diagnostic information was recorded for this cohort, spanning ordinal attributes (tumor stage, lymph node stage), quantitative attributes (age, radiation dose), nominal attributes (treatment modality), and time-series attributes with quantitative values (symptom ratings) collected in two stages, *acute* (baseline and during treatment, higher frequency) and *late* (after treatment, lower frequency).

Self-reported longitudinal symptom data was extracted from patient questionnaires [13] over the span of 12 time points: before starting the treatment, weekly for 7 weeks during treatment, 6 weeks after treatment, and 6, 12, and 18 months post treatment. Symptoms were rated on a 0-to-10 scale, from "not present” (0), to "as bad as you can imagine” (10). A total of 28 symptoms were considered in this longitudinal assessment, split into HNC specific symptoms (swallow, speech, mucus, taste, constipation, teeth, mouth sores, choking, and skin problems), general cancer symptoms (fatigue, sleep, distress, pain, drowsiness, sadness, memory, numbness, dry mouth, appetite, breath, nausea, and vomiting problems), and daily life interference symptoms (work, enjoyment, general activity, mood, walking, relationships problems). The 12 timepoints belong to one of two categories: the *acute* stage (once before the treatment's start date, or week 0, and all 7 weeks throughout the treatment), and the *late* stage (the remaining 4 post treatment assessment dates). Not all features were available for every patient. Missing clinical variables were marked as “unspecified”, and missing symptom ratings were considered a rating of 0, which were not considered when building the models.

This cohort presents six possible treatment combinations: induction with concurrent chemotherapy and radiation therapy (ICC) (n = 97), concurrent chemotherapy and radiation therapy(CC) (n = 329), induction and radiation therapy (IRT) (n = 66), radiation therapy alone (RT) (n = 199), surgery and other treatments (S_and_others) (n = 75), and surgery alone (S) (n=57). Patients were stratified by treatment during the sequential rule mining analyses. Patients receiving surgery alone were removed from the model building because this sub-cohort did not report weekly symptom scores during treatment.

4.3 SRM Modeling for Medical Data

Association Rule Mining (ARM) [1] is an unsupervised method that identifies frequent patterns, correlations, or association structures in transactional data sets. Association rules are most commonly found in the form $X \rightarrow Y$ (the appearance of X implies the appearance of Y), with X called the antecedent and Y the consequent of the rule. Because rule mining is more transparent than the black box models used in diverse applications, it has caught attention in medical research as well [3, 50, 57]. We applied ARM in our previous work [6, 23] in the context of cancer symptoms $\{taste\} \rightarrow \{dryMouth\}$ (if the patient suffers from taste, then they will more likely suffer from dryMouth as well) by transforming our longitudinal symptom records into a transactional data set. This helped to find common symptom combinations at different stages in the patient observation period, but it did not help us predict *late* symptoms based on symptoms during treatment.

One interesting extension of association mining for temporal data is sequential rule mining (SRM) [15]. SRM uses the antecedent of a rule to predict the consequent of the rule with the condition that the antecedent precedes the consequent. We applied SRM to our longitudinal symptom data considering the during- and post-treatment time frames as temporal sequences of symptom toxicity as follows:

$$R1 : \{taste, nausea\} \rightarrow \{dryMouth\} \tag{1}$$

meaning that if a patient suffers from taste and nausea problems during treatment, they will more likely suffer from dryMouth problems after the completion of the treatment. However, the disadvantage of rule mining in clinical applications is that typically a large number of rules may be required to make knowledge actionable. Moreover, prediction should reflect a strong association relationship between the antecedent and the consequent of a rule. Fortunately, useful knowledge can be quickly identified using rule metrics such as support, confidence, and lift. In the case of the previous rule $R1$, the support of the rule is the ratio of patients that have taste and nausea problems during treatment followed by dryMouth problems after treatment:

$$sup(R1) = \frac{|\{(taste, nausea) \cup (dryMouth)\}|}{|S|} \tag{2}$$

where $|S|$ is the total number of patient symptom sequences.

The confidence of the rule predicts the risk of a patient to develop *late* symptoms (dryMouth in our example), given a certain symptomatology during treatment (taste and nausea in our example) and is reported as:

$$conf(R1) = \frac{sup(R1)}{sup(\{taste, nausea\})} \tag{3}$$

The lift of a sequential rule denotes the strength of the rule, or in other words, denotes whether the antecedent and the consequent are dependent on each other or not, and is computed as follows:

$$lift(R1) = \frac{sup(R1)}{sup(\{taste, nausea\}) \times sup(\{dryMouth\})} \tag{4}$$

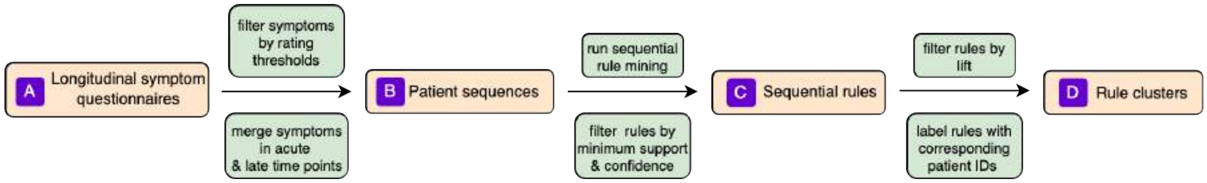

A Longitudinal symptom questionnaires
filter symptoms by rating thresholds
merge symptoms in acute & late time points
B Patient sequences
run sequential rule mining
filter rules by minimum support & confidence
C Sequential rules
filter rules by lift
label rules with corresponding patient IDs
D Rule clusters

Fig. 2: SRM Modeling. A) Patient-reported symptom ratings are recorded as longitudinal records. B) Records are processed into patient symptom sequences. C) Patient sequences are provided as input to the SRM algorithm. D) The sequential rules are filtered and clustered into rule clusters based on their corresponding patient IDs.

A lift value ≤ 1 indicates that the rule is not able to predict the consequent more accurately than what could be predicted by chance.

As already noted, rule mining can result in a multitude of rules that can show overlapping patterns. It is important to filter these results based on the previous metrics in order to get useful, easy to interpret, and meaningful information regarding the patterns within the data.

4.3.1 Back-end Design

We use Sequential Rule Mining (SRM) to identify temporal patterns in symptoms and to predict *late* symptoms. We discretize treatment ratings into two bins: before treatment and weekly ratings taken during treatment for up to 7 weeks (the *acute* stage), and ratings 6-18 months after treatment (the *late* stage) (Figure 2.A). Patients are stratified based on treatment modality, and the rule mining algorithm is run separately for each sub-cohort, as we are interested in identifying treatment-related symptoms.

We used the CMDeo algorithm [24] to compute the sequential rules, which is an adaptation from Deogun et al.'s algorithm [15] for multiple sequences of events. We followed the documentation from the open source data mining library called SPMF [25] that supports the CMDeo algorithm. The Python wrapper from this library was used for the model, which required us to pre-process our data to correspond to the input structure from the documentation.

In the first step of the data pre-processing, we computed sequences from the patient timelines (Figure 2.B). Each sequence corresponds to the temporal ratings of one patient across both the *acute* stage (baseline and during treatment) and the *late* stage (after treatment). Accordingly, we abstract the sequences into two-stage patterns, *acute* and *late* (Sec. 3). In the *acute* pattern, we include a symptom only if the patient provided a rating above a given severity threshold (e.g. ≥ 5) during any of the *acute* time points. Similarly, in the *late* pattern, we include a symptom only if the patient provided a rating above a given threshold (e.g. ≥ 3). Clinically, a rating ≥ 5 is considered a moderate-to-high severity, while 3 is considered mild severity. The same threshold is not enforce for the two stages because in general, ratings are lower in the *late* stage than in the *acute* stage. The use of a severity threshold helps to minimize patient variability and individual symptom severity ratings.

Next, the SRM algorithm was applied on these sequences to identify sequential rules (Figure 2.C). Similarly to traditional association rule mining, two input parameters, namely support and confidence, need to be specified by the user to generate the rules. In our experiments, we used minimum support (i.e. percent of patients that show the resulting patterns) of 30% or 40% depending on the number of sequences, as we consider patterns experienced by a third of the patients to be significant. The minimum confidence (i.e. risk of *late* symptoms) was set to 50%. From the initial set of rules, only rules with a lift threshold higher than 1 were selected to ensure the rules can be used for the prediction of *late* symptoms. The lift of a rule indicates the degree of dependency between the antecedent and consequent of the rule. The resulting rule sets varied from 9 to 46 rules, depending on the number of sequences for each treatment and the variety of occurring symptoms per sub-cohort.

As could be expected, the extracted rules within each treatment cohort showed a lot of similarities in terms of the symptom patterns (often differing in only one or two symptoms) and in the set of patients supporting the rules (over 90% of the same patients appearing in two or more rules). To minimize redundancy among the rules, we decided to cluster the rules into rule clusters that would then be used for visualization. We labeled each rule with the corresponding patient IDs supporting the rule. Next, we computed the similarity between rules based on their common patient IDs using Jaccard's index [33]. We used this method because we work with sets (i.e. patient ID sets) for which we wish to compute rule similarity based on the patients the rules affect. We then applied hierarchical clustering using the complete linkage [14] on the resulting similarity matrices. We used the complete linkage since the point of reducing a group of rules to a single rule was to yield cohesive rule clusters while avoiding in-cluster outliers. We used hierarchical clustering since we have found it yields highly interpretable results through the use of dendrograms [40] which allows us to manually adjust the clusters and identify outliers. We decided upon the number of clusters after inspecting all treatment results. We created rule clusters (Fig. 2.D) by merging the antecedent symptoms and consequent symptoms from all rules within a cluster. Thus, each cluster is formed by a set of *acute* symptoms and a set of *late* symptoms.

We attached to each cluster all the patient IDs from that cluster's corresponding rules. This is helpful for visually connecting the cluster information with the patient cohort. We report the following measurements per each cluster: 1) the probability (support) of developing the *acute* symptoms given a treatment method; 2) the probability of the *acute* symptoms to develop the cluster's corresponding *late* symptoms, given by the confidence of the rule cluster; 3) the likelihood that the temporal pattern shown by the rule cluster will appear more frequently as compared to the rest of the treatment modalities, given by the support of the cluster within the treatment over the support of the cluster outside the treatment (i.e. for all the alternative treatment modalities).

4.4 Front-end Design

Our system uses Python for the back-end and React with D3.js for the front-end. The top design is based on coordinated multiple views, which support diverse analysis workflows. The interface consists of 5 panels that support 6 types of visual components. The top panel (Figure 1.A) serves deliberately as an anchor which cannot be configured by the modeler, and displays the stratified overall symptom severity for the entire patient cohort. The remaining quadrants can be configured with any of the following five visual components: symptom clustering (Figure 1.B) - which denotes temporal symptom clusters for one treatment; patient clustering (Figure 1.D) - which shows patient cohort symptomatology attributes for one treatment; cohort characteristics (Figure 1.C) - which correlates diagnostic data to symptom clusters and symptom overall severity over time; cohort timeline (Figure 1.E) - which displays an in-depth view of each patient's longitudinal and diagnostic features; and the symptom query component (Figure 7.C) - which provides overall statistics regarding the appearance of symptoms during (*acute*) and after treatment (*late*). Excluding the top view, which uses the entire cohort, each component displays the data for one treatment modality. The quadrants have treatment and visual component queries attached to their top-left to facilitate workflow configurations. This front-end design supports our modeling goals, which are centered on treatment comparison and cluster outlier analysis, but not on patient comparison, nor on alternative symptom rankings.

Our design uses custom Rose Glyphs (Figure 3) to encode the trajectory of a single symptom severity within either the entire cohort (Figure 1.A), or subgroups in the data (Figure 1.B). For the selected subgroup, the mean symptom rating at each time interval is encoded

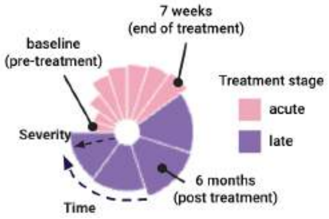


Fig. 3: Rose glyph. Color-coded petals aggregate the mean severity for patients for symptom dryMouth. Petals in the radial layout start at 9 o'clock and proceed clockwise. Pink "petals" encode *acute* time intervals while purple encodes *late* time intervals. *Late* petals are wider to depict longer time intervals, while *acute* petals depict shorter intervals.

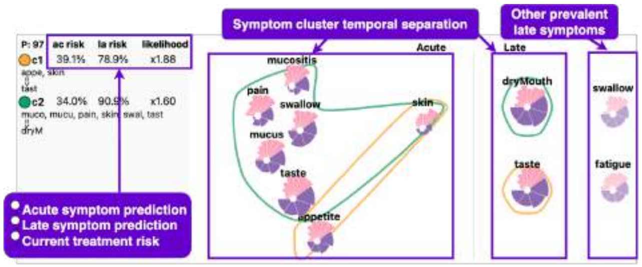


Fig. 4: Symptom clustering for treatment ICC. The clusters in orange and green predict in the *late* stage dryMouth and taste problems. Cluster 1 (orange) shows a higher risk to develop these toxicities for ICC rather than the other existing treatment modalities (i.e. 1.88 times more likely).

using variable-radius slices (petals). The symptom trajectory starts with the baseline ratings at 9 o'clock, and progresses clockwise in order of increasing time intervals, showing rating details for each of the twelve time intervals of the patient observation period. Pink petals encode *acute* treatment time intervals, while purple petals encode *late* intervals. The width of the petals is driven by the two-stage data sampling, and by the modelers' interests in late symptoms (Sec. 3). The color scheme was chosen based on perceptual criteria, and to emphasize late intervals. The flat interval mapping was in alignment with the clinical assumption that symptom variation within an observation interval is not significant. We took inspiration from Florence Nightingale's Rose Diagram [7] for this glyph, but instead of focusing on comparing events within a time frame, we concentrate on temporal trajectory and comparing trajectories across symptoms. We employed the radial glyph design because it provided a compact way to display symptom burden across cohorts and treatments, while supporting rapid similarity detection [43]. Our timeline histogram supports symptom value comparison better than the glyph layout; however, the glyph's purpose is not accurate symptom value comparison, but compactly encoding each symptom's trajectory and comparing symptom trajectories.

4.4.1 Overall Symptom Severity

This component (Figure 1.A) displays the mean severity (i.e. rating) distribution for each of the 28 symptoms for the entire patient cohort (i.e. all treatment modalities) (T2.3) using rose glyphs. The symptom list starts with dryMouth, which is the one of the most severe symptoms throughout the observation period, and it is the most persistent symptom after treatment across patient sub-cohorts. The rest of the symptoms are ordered based on the cosine similarity to dryMouth, computed using the mean temporal ratings per each symptom. We used cosine similarity because we are more interested in relative frequency of symptom occurrence, as there can be a large variation in self-reported symptoms among items that may not correlate to their impact on quality-of-life. The symptoms are grouped based on temporal similarity, in support of our modeling goals. While other grouping options are possible, they were g priority in this project. Symptoms predicted by SRM in at least one of the treatments are highlighted with a shadowed border. This encoding provides a compact way of showing overall symptom burden for the entire cohort and it serves as a reference point for evaluating treatment-specific symptom patterns.

4.4.2 Symptom Clustering

This component (Figure 4) provides a visual abstraction for the symptom clusters found in Section 4.3.1 through a 2D projection of the corresponding symptoms using rose glyphs (T1.1, T2.1). The view is split into two halves to facilitate the temporal separation between *acute* and *late* stages. The X and Y axis in the *acute* half correspond to the first two principal components after applying PCA on the Jaccard's similarity between symptoms, based on the common patient IDs they share. Because many of these symptoms have underlying association, we used PCA, as opposed to other projections, as it works better for correlated attributes. We use a force-directed layout to ensure the symptom glyphs are not overlapping in the projection. Symptoms are represented using rose glyphs to show the mean severity distribution over time across the patients that correspond the clusters. This also enhances temporal symptom severity comparisons between a selected treatment and the overall cohort or another treatment.

In the *late* half (Figure 4), the clustering results are not part of the PCA projections because these clusters usually resulted in one or two different symptoms in this stage for a given treatment. Additionally, we listed on the right edge of the view the *late* symptoms that appeared in our rule mining results, but were not part of the rules filtered for the prediction or the clustering of the symptoms due to low metrics results (i.e. lift < 1). We chose to visualize these additional *late* symptoms to highlight the fact that, although the data shows many common treatment-related toxicities, these cannot be accurately predicted using *acute* symptoms with the data at hand. We mark these symptoms using a low opacity for the rose glyphs as opposed to the predicted symptoms.

The left legend of the component shows the details for each symptom cluster (Figure 4): the cluster ID, the corresponding antecedent (*acute* symptoms) and consequent (*late* symptoms), he support of the *acute* stage (i.e. how many patients display the symptom patterns from the *acute* stage), the confidence of the cluster (i.e. the risk of developing *late* symptoms given the *acute* symptoms), and the support of the cluster within the treatment cohort over the support of the cluster for the other treatment cohorts (i.e. the likeliness that this cluster might appear more frequent for the given treatment as opposed to all the other treatments) (T2.2). Each cluster is highlighted using an envelope (Figure 7.A,B) categorically color-coded. The envelopes' background can be turned on (Figure 1.B), which can better emphasize the symptom correspondence to clusters, using a Venn diagram-like illustration. From the legend panel, the clusters can be unselected, which will result in the removal of the highlight for those cluster envelopes. Selecting a symptom glyph from the projection or a cluster label from the legend will result in highlighting the complementing information in the other interface components (e.g. cohort attributes that correspond to the selected item).

In our previous work with rule mining for symptom analysis, we used node-link diagrams to represent the symptoms' inter-relationships [21]. Domain experts preferred this 2D visual abstraction due to the small number of rules that we displayed. However, in this project we work with a larger number of temporal rules. Early prototypes relied on a combination of network-based encodings and barplots. However, this resulted in clutter due to the large number of edges between nodes that did not capture well the temporal nature of the rulesets. As a result, we detached from displaying actual rules and opted for a cleaner projection that uses rule clusters, using envelopes to show relationships between symptoms and horizontal separation to denote temporal direction. We opted for the rose glyph, as opposed to circles, for symptom interpretation, to enhance trajectory comparison between symptoms.

4.4.3 Cohort Symptom Query

This component (Figure 7.C) provides an overview of all the 28 symptoms from the cohort for the *acute* and *late* stages, and guides the analysis of symptom clusters, using a vertical barchart (T2.1). For a selected treatment, tumor and lymph node stage, *acute* and *late* symptom rating thresholds; this view returns the percentage of the patients

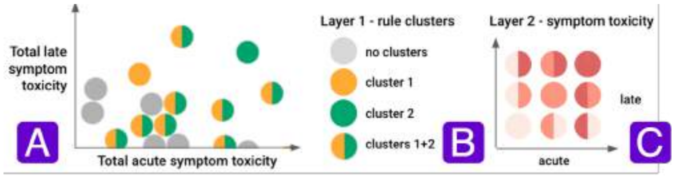

Total late symptom toxicity
A
Total acute symptom toxicity
Layer 1 - rule clusters
no clusters
cluster 1
cluster 2
clusters 1+2
B
Layer 2 - symptom toxicity
late
acute
C

Fig. 5: Patient Clustering View. A) Scatterplot showing patient glyphs. Two options for patient encodings in the scatterplot: B) encodes cluster memberships and C) encodes temporal symptom burdens.

that have reported symptoms above the given thresholds at least once during and after treatment for each symptom. Symptoms are ordered from top to bottom by the highest cumulative percentages for *acute* and *late* occurrences, which highlights symptoms of high prevalence among patients. Symptoms from SRM clusters are colored in blue.

We proposed this encoding in early prototyping iterations, to show statistics for rule symptom occurrences. Our collaborators quickly adopted it into their analysis due to its low complexity, so we chose to follow this design for explaining symptom prevalence.

4.4.4 Patient Clustering

This component (Figure 1.D) provides a custom 2D scatterplot projection of the patient cohort, with axes corresponding to the total symptom severity scores for the *acute* time points (X axis) and the *late* time points (Y axis) (Figure 5.A). We chose this orientation to better highlight patient outliers for the *acute* and *late* stages (T1.2). We use a force-directed layout to remove overlap and ensure that each individual patient can be selected from this projection for further analysis. Patient similarity comparison is supported by the scatterplot projections.

This component has two interchangeable layers: the first layer (Figure 5.B) colors the points based on the patients' rule cluster labels. If a patient is not included in any of the rule clusters, their corresponding point is gray. Otherwise the point is split into as many sections as the number of clusters it belongs to, where each section is colored to match the clusters colors from the symptom clustering component. The second layer (Figure 5.C) splits the points into two sections, representing, from left to right, the *acute* and *late* treatment periods, respectively. The color of each section is mapped to the overall symptom severity for its corresponding period, with lighter red encoding low severity and dark red encoding high severity. This layer can be applied upon selecting a subset of symptoms from the top rose glyph row (Figure 1.A), and it will be updated to show the *acute*/*late* severities of the selected symptoms. Brushing operations will highlight or filter information in the rest of the views based on the patient selection (T1.4).

Alternative designs experimented with other projection methods and glyph encodings. However, we found most projection methods like PCA [23] and T-SNE did not capture the rule clusters and associations. In contrast, we found moderate-to-high *acute* and *late* symptom ratings were consistently correlated with more cluster membership, which made the glyph encoding more intuitive to collaborators. Using symptom severity made it easier for collaborators to identify patients with increases or decreases in treatment severity between the *acute* and *late* stage. For the scatterplot glyph design, we considered alternative shapes instead of circles for different clusters, but we found that it was difficult to capture an arbitrary number of cluster memberships across treatment modalities using shape. For the symptom toxicity layer, we considered splitting circles into more than two time periods (i.e. *baseline*, *acute*, *late*) or using rose glyphs, but that cluttered the view and made it difficult to find patterns. This component ensures a better understanding of the model results and clinical statistics as it connects that cohort information to actual patients for the given treatment.

4.4.5 Cohort Timeline

This component (Figure 1.E) functions as a detailed view of the attributes of each patient (T1.4), using timelines and small multiples to show mean symptom ratings over time, patient cluster labels, and diagnostic information (T1.3). The left half of the view shows the patient's ID, tumor (T) stage, lymph node (N) stage, symptom clusters labels, and temporal symptom severity using their corresponding points from the scatterplot (Figure 8.C). The right half uses a barchart timeline, split by the *acute* and *late* stages, showcasing mean ratings for each of the 28 symptoms (Figure 8.C). The symptom bars are ordered by the top interface row symptom order. They are colored in blue when they represent symptoms that are present in at least one of the rule clusters for the selected treatment, or in gray vice versa. The patient timelines are listed in descending order based on the cumulative *acute* and *late* symptom severity, and based on symptom cluster membership in case of equality for the former metric. Brushing from the scatterplots will filter this view by the selection. Clicking on the patient IDs will highlight the corresponding patients in the scatterplot and in flows in the cohort characteristics component.

Oncology experts are often interested in analyzing a single patient and comparing them to the rest of the cohort. As a result, we designed this component to make individual-patient analysis possible. Previous prototyping iterations explored matrix-based encodings which included all timepoints from the symptom data. This resulted in cluttered components which took the majority of screen space due to the large number of timepoints, making the inclusion of diagnostic patient data difficult. Thus, we adopted this custom, simplified view of the temporal symptom data, deciding to aggregate the *acute* and *late* time points while also integrating the diagnostic and symptom cluster/severity labels. The timeline component can also be used to observe how a patient's symptomatology trajectory compares to other patients, or to observe the overall burden of symptoms for a given set of patients (T1.2, T2.3).

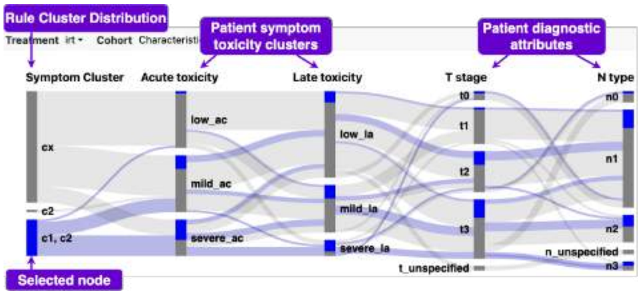
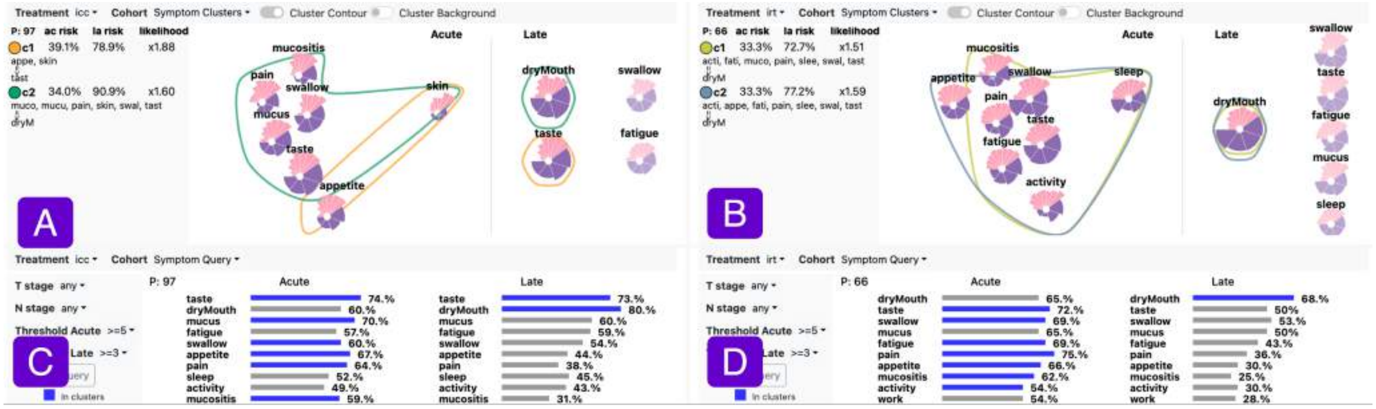


Fig. 6: Sankey Diagram for IRT treatment. Node $c1, c2$ is selected, showing that a very small part of the patient cohort from with this cluster combination is linked to low symptom severity in the *acute* stage.

4.4.6 Cohort Characteristics

This component (Figure 6) connects symptom cluster memberships, overall symptom burden for the *acute* and *late* stages, and diagnostic patient data (T stage, N stage) using a Sankey Diagram (T1.3, T2.4). Apart from showcasing the possible symptom cluster combinations, we stratify the patients into low, medium, or high symptom burden for the *acute* and *late* stages using K Means clustering on the total symptom toxicity scores for both stages. This further emphasizes how *acute* symptom burden transposes into *late* symptom burden. The nodes from the diagram can be selected and the corresponding nodes and flows are highlighted in blue (Figure 6), while filtering options in the other views highlight with blue the selection in this component as well.

When we prototyped this component, we kept in mind that we needed to showcase the distribution of categorical cohort attributes while also considering time directionality for our temporal attributes (i.e. *acute* and *late* symptom toxicity). We opted for a Sankey design as it has shown adoption in both categorical and temporal attributes in previous work [66]. This design was easily adopted by our collaborators and became a pivotal component in their analyses. All of the diagram's ordinal axes are ordered from top to bottom (i.e. T/N stage, *acute*/*late* toxicity), as per the suggestion of our oncology domain experts. Due to the limited number of attributes, this component can clearly show the distribution of a particular attribute's values for a treatment modality and how it is connected to the distributions of the other cohort attributes.



Treatment icc ▾ Cohort Symptom Clusters ▾ Cluster Contour Cluster Background

P: 97	ac risk	la risk	likelihood
c1	39.1%	78.9%	x1.88
appe, skin ⇓ tast			
c2	34.0%	90.9%	x1.60
muco, mucu, pain, skin, swal, tast ⇓ dryM			

Treatment irt ▾ Cohort Symptom Clusters ▾ Cluster Contour Cluster Background

P: 66	ac risk	la risk	likelihood
c1	33.3%	72.7%	x1.51
acti, fati, muco, pain, slee, swal, tast ⇓ dryM			
c2	33.3%	77.2%	x1.59
acti, appe, fati, pain, slee, swal, tast ⇓ dryM			

Treatment icc ▾ Cohort Symptom Query ▾

T stage any ▾

N stage any ▾

Threshold Acute >=5 ▾

Late >=3 ▾

P: 97

Acute		Late	
taste	74.%	taste	73.%
dryMouth	60.%	dryMouth	80.%
mucus	70.%	mucus	60.%
fatigue	57.%	fatigue	59.%
swallow	60.%	swallow	54.%
appetite	67.%	appetite	44.%
pain	64.%	pain	38.%
sleep	52.%	sleep	45.%
activity	49.%	activity	43.%
mucositis	59.%	mucositis	31.%

Treatment irt ▾ Cohort Symptom Query ▾

T stage any ▾

N stage any ▾

Threshold Acute >=5 ▾

Late >=3 ▾

P: 66

Acute		Late	
dryMouth	65.%	dryMouth	68.%
taste	72.%	taste	50%
swallow	69.%	swallow	53.%
mucus	65.%	mucus	50%
fatigue	69.%	fatigue	43.%
pain	75.%	pain	36.%
appetite	66.%	appetite	30.%
mucositis	62.%	mucositis	25.%
activity	54.%	activity	30.%
work	54.%	work	28.%

Fig. 7: Treatment comparison. A) Overall cohort toxicities for all timepoints. B) and C) Symptom clusters for treatments ICC and IRT. Both treatments show two clusters, with similar *acute* symptoms, but ICC presents taste as a *late* symptom (B), as opposed to IRT(C). Although the rose glyphs are projected based on similar patients in the *acute* half, both treatments have outliers (i.e. skin and sleep in *acute*). D) and E) Symptom queries showing the prevalence of all symptoms for the two selected treatments. These bottom views show that, although there are many *late* common toxicities, not all can be predicted by the *acute* symptoms in B) and C) (i.e. mucus in *late* ICC is prevalent but not predicted in the symptom clusters).

4.4.7 Flexible Workflow Support

Due to the variation of requirements that would support analysis at both the patient cohort level, as well as at the symptom cohort level, we designed these visual components to provide a balance between flexibility and guidance across analysis workflows. Our modelers were interested in understanding and interpreting the SRM model results in the context of treatment decision making and treatment-related symptoms. However, they were also looking for common symptoms across treatments that may develop independent of treatment strategy. Moreover, they were interested in predicting what a new patient should expect given a selected treatment to better assist future treatment decisions. To support these workflows, the afferent components can be flexibly swapped.

5 Evaluation

We evaluated the system and the resulting models through multiple demonstrations and case studies involving two senior model builders, two junior model builders, and three senior clinical oncology co-modelers with ML experience. Beyond the six domain expert co-authors of this paper who provided feedback, an additional oncology researcher also evaluated the system. The model builders were active co-designers, whereas the oncology experts provided occasional input and feedback. Although our system is dedicated to model builders in cancer symptom research, we also needed to clinically validate the results we had found. The evaluators participated in several hour-long demo sessions over 4 months, followed by the case studies. They also explored the tool on their own while providing feedback. We illustrate two case studies that were conducted through focus groups via Zoom, using screen sharing and note taking. During these sessions, the first author navigated the interface under the guidance of the model builders and oncology co-modelers, using the think-aloud method. These studies used a cohort of 766 HNC patients that presented five treatment modalities: RT, IRT, CC, ICC, and Surgery_and_other. We present below, in abbreviated form, these case studies.

5.1 Case Study I: Multi-treatment Analysis

The model builders wanted to find temporal symptom patterns across multiple treatment modalities and compare the results. The oncologists were hoping they would find specific symptoms highly correlated to specific treatment strategies. After examining the top row of the interface (Figure 1.A), the evaluators noted that, unsurprisingly, common toxicities such as dryMouth, taste, swallow, and mucus were the highest overall (T 2.3). In general, symptoms usually followed a gradually increased toxicity during treatment and decreased post treatment, which was expected. However, symptoms related to daily life activities, such as mood, enjoyment of life, distress, and sadness showed severity peaks before the start of the treatment (i.e. first pink petal), implying that mental health improved when the patients started the treatment (i.e. the severity decreased). Next, the interface was used to show the symptom clusters for ICC (Figure 7.A) and IRT (Figure 7.B) in conjunction with the symptom queries (Figure 7.C,D). Using the symptom queries, the evaluators found similar prevalent symptoms for both treatments (T2.3). In the symptom cluster components, both treatments showed two temporal clusters each, with similar overall symptom profiles (T2.1). Although the symptom queries showed many prevalent *late* toxicities (Figure 7.B,D), they were not all predicted by the model. These symptoms appeared as common *late* toxicities in the rule mining results, as shown by the low opacity *late* symptoms in the clusters panels (Figure 7.A,B). Taste was predicted as a *late* toxicity for ICC, correlated with the loss of appetite, and, surprisingly, with skin problems (Figure 7.A). DryMouth showed obvious severe toxicity in *late* when compared to the whole cohort (Figure 1.A), more so for IRT (Figure 7.B) (T2.1,3). The evaluators appreciated how the rose glyph projection kept symptoms with similar trajectories together. For instance, in the ICC symptom clusters, pain and mucositis showed strikingly similar trajectories (Figure 7.A). They hypothesized this might be a sign that pain, being such a general symptom, was highly correlated with mucositis problems for this cohort. The evaluators also showed particular interest in the outliers of the *acute* projections, namely problems with sleep in IRT and skin in ICC.

Checking the Sankey diagrams for the two treatments (Figure 1.C, Figure 6), the evaluators observed that that IRT showed N3 stage (advanced) for node lymphs, while ICC did not present such a high attribute value (T2.4). Although the evaluated cohort had missing data, the oncology co-modelers appreciated the model's ability to find common longitudinal patterns for small sub-cohorts which show increased risk of developing those patterns within the given treatment (T2.2). For example, although only 97 patients were given ICC (Figure 7.A), the model predicted a higher likelihood (i.e. almost two times more likely) that appetite and skin problems could cause dryMouth as opposed to all the other treatments. The evaluators concluded that the symptom clustering component was an effective way to understand the impact of *late* symptoms in a sub-cohort. They are excited to analyze the SRM results with more symptom rating data for this patient sub-cohort.

5.2 Case Study II: Single Treatment Analysis

For the second study, the oncology co-modelers wished to better understand the mechanisms between symptom clusters. They started with a treatment example, ICC. The interface was configured as follows: the symptom cluster component (Figure 8.A), patient projection component using the symptom cluster layer (Figure 8.B), the patient timeline component (Figure 8.C), and the cohort characteristics component (Figure 1.C). At first glance, the patients that usually suffered from

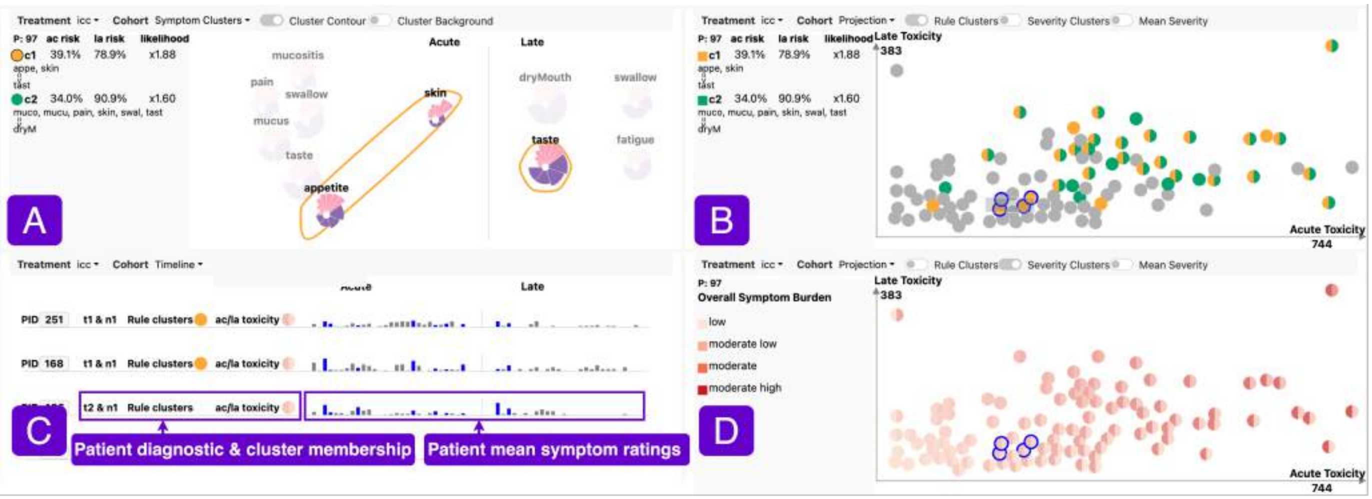


Fig. 8: Single-treatment analysis. A) ICC treatment symptom clusters with cluster 1 (orange) selected. B) ICC patient projection with the cluster label layer. The cluster 1 outlier patients from the lower-left side are selected and highlighted with blue in the scatterplot filtered in the other views. C) patient timelines for the selection from B) showing low mean temporal toxicities. D) Patient projections with the toxicity layer. The selection from B) is highlighted with blue in this view, and shows moderate total severities for both *acute* and *late*.

moderate-to-high symptom burden overall showed patterns among the two existing symptom clusters (Figure 8.A,B) (T1.1). Patients usually showed problems from both clusters, with lower burden patients sharing mostly cluster 1 (appetite, skin ->taste) (Figure 8.B). Selecting the previously mentioned sub-group of patients with cluster 1 from the scatterplot (T1.4), the evaluators looked at their timelines (Figure 8.C), and observed low mean symptom ratings for both treatment stages, with peaks among the symptoms from the clusters (T1.1). Moving to the cohort characteristics component, the modelers observed that most cluster combinations among this cohort showed higher symptom burden for the *acute* and *late* stages, but the symptom cluster 1 patients showed only below T3 stage problems (T1.3). While evaluating the cohort characteristics component, the oncology co-modelers commented that they expected severe symptom burden in *late* to be correlated with higher T stage (i.e. T3) (Figure 8.C), which this view proved that it was not the case.

Next, the evaluators wished to understand the overall temporal toxicity among patients better. The cohort characteristics component was changed to show the patient projection with the overall temporal severity layer applied (Figure 8.D) (T1.2). This way, they could better understand the relationship between symptom cluster labels and *acute-late* toxicity. They noted that almost half of the patients within this treatment often showed severe toxicity during *acute*, but low severity after the completion of treatment, which was received with relief. Selecting the top-left outlier (T1.4), the evaluators observed that the given patient did not have reported data for the *acute* stage, making it an outlier, and agreed that the medical records for this patient needed further analysis. The oncology co-modelers expressed that the scatterplot was really efficient in detecting outliers in patient data, while also connecting the cohort to symptom burden characteristics. After finding the outliers and unexpected diagnostic patient details connected to symptom clusters, the evaluators decided that their future studies should focus further on diagnostic patient data.

5.3 Expert Feedback

The visual system and model results received positive feedback. One of the senior model builders affirmed: *"The interface is extremely useful for navigating through the data patient-reported outcome data and generating hypotheses. Evaluating the effect of different thresholds for symptom severity and rule mining would be overwhelming without these visualizations [...] Using the rose glyphs gives a quick overview of the symptom trajectory for a group of patients and it is easy to compare between different therapeutic combinations [...] The sequential rules provide a way to identify acute symptoms that can be predictive for late toxicity. The rule clustering dramatically reduces the complexity of the analysis by reducing the number of relevant rules and highlighting interesting metrics to compare the different treatments."*

The oncology co-modelers were really impressed, one of them affirming: "The app is very good and combines all the information in one place, so that is very interesting", while another commented: "I really like this..I feel very strongly about this, the utility for exploring the data here is very high" and "if you're talking about quantitative decision-making, this is very strong". The appreciation for multiple data-driven analyses was further emphasized by the oncologist co-modelers: "First, we can stratify by treatment, [...] second, we can see that patients who have certain patterns of symptoms like those more impacted by skin and appetite are more likely to get taste problems later on than [...] third, you can stratify the patients by T stage, N stage, and different clinical parameters [...] so for me, it is really, really helpful, it is a really cool tool". When considering wider adoption of the models in the clinic, the oncologists wished for additional workflows that started with data from the patient they are treating, and to analyze similar patients from the dataset to predict the *late* symptoms for that patient.

Our visual encodings were designed through a parallel prototyping process, with frequent feedback and suggestions from collaborators, often aimed at reducing encoding complexity and increasing alignment with the clinical intuition (see Supplemental materials). In the end design, the modelers appreciated the usefulness and many tasks that the rose glyphs accomplish, from single-symptom, single-treatment analysis to multi-symptom, multi-treatment analysis. One oncologist co-modeler commented when analyzing the rose glyphs: *"Fascinating that taste is so prevalent [...] we don't understand why it's so bad. The kinetics are fascinating"*. The oncologists responded well to the inclusion of the rose glyphs as a fixed anchor at the top of the interface, and were able to immediately spot and comment on trends in different symptoms of interest. In an earlier iteration, a collaborator was able to identify a trend where patient symptoms decreased after the first week of treatment, which quickly led to finding an issue in the data preprocessing. Similarly, during the review, they immediately identified and commented on interesting trajectories for taste, and noted that they should explore taste-related issues in future studies (T2.1,T2.3). The glyphs' horizontal separation for *acute* and *late* facilitated interpretation of sequential rule clusters (T1.1), and their compact design was particularly appreciated when comparing trajectories (T2.1, T2.3). Secondly, they appreciated the Sankey diagram: *"this one is going to help if you want to connect the dots between staging and toxicity, and symptom clusters, so it gives an overall connection"* because they could compare symptom burden and clinical data across treatments more easily (T2.4). The diagram was an intuitive way of analyzing temporal and categorical patient attributes (T1.3) and it revealed surprising results: *"I expected that the more advanced staging you have (T stage), the more toxicity you get - it corrected my assumptions"*. They found the scatterplot useful to observe symptom burden temporal trends at the cohort level while detecting outliers, and to compare symptom burden trends across treatments (T1.4, T2.3). As our collaborators routinely analyze cohorts using scatterplots, they found the scatterplot temporal abstraction intuitive to analyze overall symptom burden and its relation to symptom clusters (T2.2). The other components served as useful complements to the model analysis.

6 Discussion

This work was developed as a collaborative project alongside oncologists and data scientists to create explainable rule mining and clustering of temporal patient symptoms. The evaluation with domain experts in symptom research demonstrates that our visual system successfully explains the SRM model results in the context of several aspects of the patient and symptom cohort data. Our results show that our visual system is an effective tool for collaboratively analyzing treatment-related symptom patterns in clinical patients. Our combination of SRM and rule clusters allows for a comprehensible explanation of common co-occurring symptoms and predicting *late* stage symptoms for different treatment groups. While we focus our design on model building, our case studies and feedback suggest that our interface is able to provide usable insights for clinical practitioners. Although we target radiation oncology patients, we generalize design insights to a wide range of approaches when dealing with complex, temporal patient outcomes and when working with clinical explainable ML models. Next, we present the lessons learned from this multi-disciplinary collaboration:

L1. *Use visual scaffolding to introduce new visual encodings.* At the beginning of the design process, we started by visually listing sequential rules to the senior modelers, which were hard to interpret due to too many existing patterns. Thus we worked on a model that would summarize the mined rules, but we needed the means to convey the rule results in a meaningful way. That is how we came up with the rose encoding, which in the beginning, provoked some skepticism from the domain experts who were used to the rule abstraction from our previous work, which used node-link encodings for rules, their antecedents, and consequents. However, this abstraction did not work in the present work because node-link representations were not able to deal with large numbers of rules or capture temporal patterns. Replacing the original node-link with 2D projections, whose items were temporally separated and grouped using envelopes, proved to be an intuitive solution. After a couple of sessions throughout the interface prototyping stage, the oncology co-modelers got to rely on this encoding the most, and during the evaluation session, it ranked the highest among their preference.

L2. *Focus on actionability and transparency when working with clinical XAI applications.* When developing our model, we focused on rule-mining based on the positive reception the approach received from clinical researchers due to its simplicity and transparency. However, we found that large rule sets with overlapping results made the model lack actionability. We addressed this issue by producing rule clusters that could be easily interpreted. Moreover, adapting rule metrics (i.e. support, confidence, lift) to clinical context (i.e. symptom risk) helped the team identify interesting results, adding to the actionability of the SRM models. This drastically improved reception from collaborators.

L3. *Use highly configurable interfaces in XAI modeling.* Although our designed focused on visually interpreting the SRM clusters and evaluate how they are impacted by treatments, we found that properly analyzing the data required varied workflows and orientations, such as analyzing individual patients, rules, symptoms, treatments, and clusters, in order to fully understand the underlying algorithm and assess where issues may arise. Our human-machine visual system supports a variety of workflows with the help of six visual components. At the same time, the use of configurable layouts allowed us to minimize cognitive load when working collaboratively by hiding unnecessary components.

L4. *Account for multiple levels of details when working with collaborative workflows.* Our system was designed in coordination with multiple domain experts, who approach the problem with different viewpoints, which required different forms of data abstraction. For example, a senior model builder was more interested in identifying the rules with the highest confidence and support, and thus benefited from views with higher levels of aggregation such as the symptom query view, along with a layout that was more focused on showing multiple different panels. In contrast, the oncologists gave insights into potential mechanisms behind symptom clusters, and others were interested in exploring single patients to identify and explain outliers or assess the value of the rules when explaining results to patients, and thus benefited more from the inclusion of the scatterplots alongside the symptom cluster view. In addition, since clinicians were more interested in the impact of different treatment groups, they benefited more from configuring the layout to allow for side-by-side comparisons between panels showing results for different treatment groups. By providing a configurable interface that allowed for analysis of both rule sets and sub-groups of patients with different granularities, we were able to better accommodate different insights and workflows from experts in data mining and oncology.

Because our system aims to visualize individual patients in the cohort, some of our visual components such as the scatterplot and individual patient timelines can show scalability issues if they must support a large number of patients (e.g. n > 800). However, this may be addressed by increasing the granularity of the sub-cohorts used to reason about the data. On the other hand, the Sankey diagram, rose glyphs, and symptom query barcharts can support any cohort sizes. The timeline component aims to support analysis of only one or a handful of patients at a time. Moreover, if having to support more data attributes, the Sankey diagram would become harder to understand, although brushing operations can uncover the necessary connections. On the other hand, given the difficulty in collecting large homogenous cohorts of symptom data, we felt that it was more important to provide a highly configurable interface, supporting several workflows, at the cost of some scalability issues. Each visual component and view of the interface can be initialized with a given sub-cohort, with consistent layouts across the views, in order to support side-by-side symptom cluster, symptom burden, or treatment comparison. Single-cohort, symptom, treatment, and outlier analysis is further supported across views through brushing and linking operations. Pairwise sub-cohort comparison was, however, not a modeling goal.

Notably, some of the patients used in the model building were still under the observation period, and as a result, they were missing symptom ratings for many after treatment time points. This impacted the results of the model's predictions. Our modeling approach is generalizable, although it is a clinical-practice based model. Future work includes supporting SRM model refinement once surveillance is complete, and applying rule mining to longitudinal treatment plans even after potential cancer recurrence.

7 Conclusion

In this work, we described the activity-centered design of a visual analytics system that helps to explain and validate our proposed multivariate model for longitudinal symptom analysis. While we examine a cohort of HNC patients, our approach can be generalized to other disease applications that study cohort toxicities. Our back-end uses SRM in conjunction with other unsupervised methods to predict and find temporal patterns in cancer symptoms, while our front-end supports the analysis of these models in the context of real patient data. We propose SRM to find temporal symptom clusters and a new visual encoding, the rose glyphs, to describe the resulting clusters and predictions. Our visual system supports various workflows through configurable components, which guide to a better understanding of treatment-related symptoms for multiple treatments. The evaluation with domain experts in cancer symptom model building demonstrates the usefulness of our approach in clinical research. Lastly, we summarize the lessons learned from this multidisciplinary collaboration, and we hope they will guide towards better XAI applications in healthcare.

Acknowledgments

Our work is supported by NIH awards NCI-R01-CA258827 and NLM-R01-LM012527, and NSF awards CDSE-1854815 and CNS-1828265.

References

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, p. 487–499. Morgan Kaufmann Publishers Inc., 1994. 3

[2] A. Aktas, D. Walsh, and L. Rybicki. Symptom Clusters: Myth or Reality? *Palliative Med.*, 24(4):373–385, 2010. doi: 10.1177/0269216310367842 2

[3] D. Antweiler and G. Fuchs. Visualizing rule-based classifiers for clinical risk prognosis. In *2022 IEEE Vis. and Visual Anal. (VIS)*, pp. 55–59. IEEE, 2022. doi: 10.1109/tkde.2008.131 2, 3

[4] T. Baumgartl, M. Petzold, et al. In Search of Patient Zero: Visual Analytics of Pathogen Transmission Pathways in Hospitals. *IEEE Trans. Vis. Comp. Graph.*, 27(2):711–721, 2021. doi: 10.1109/tvcg.2020.3030437 2

[5] J. Bernard, D. Sessler, et al. A visual-interactive system for prostate cancer stratifications. In *Proc. IEEE VIS Workshop Visualizing Electronic Health Record Data*, 2014. doi: 10.1109/mcg.2015.49 2

[6] M. Biggs, C. Floricel, et al. Identifying Symptom Clusters from Patient Reported Outcomes through Association Rule Mining. *19th Int. Conf. Artif. Intel. in Med. (AIME)*, 2021. doi: 10.1007/978-3-030-77211-6_58 2, 3

[7] L. Brasseur. Florence nightingale's visual rhetoric in the rose diagrams. *Technical Communication Quarterly*, 14(2):161–182, 2005. 5

[8] D. Bruzzese and C. Davino. Visual Mining of Association Rules. In *Visual Data Mining*, p. 103–122. Springer, 2008. doi: 10.1007/978-3-540-71080-6_8 2

[9] H. S. G. Caballero, A. Corvo, et al. Visual analytics for evaluating clinical pathways. In *2017 IEEE Workshop on Vis. Anal. in Healthcare (VAHC)*, pp. 39–46. IEEE, 2017. doi: 10.1109/VAHC.2017.8387499 2

[10] G. Canahuate, A. Wentzel, et al. Spatially-aware clustering improves ajcc-8 risk stratification performance in oropharyngeal carcinomas. *Oral Oncology*, 144:106460, 2023. doi: 10.1016/j.oraloncology.2023.106460 2

[11] B. C. Cappers and J. J. van Wijk. Exploring multivariate event sequences using rules, aggregations, and selections. *IEEE Trans. Vis. Comp. Graph.*, 24(1):532–541, 2017. doi: 10.1109/tvcg.2017.2745278 2

[12] K. K. H. Chui, J. B. Wenger, S. A. Cohen, and others. Visual Analytics for Epidemiologists: Understanding the Interactions Between Age, Time, and Disease with Multi-Panel Graphs. *PLoS one*, 6(2):1–8, 2011. doi: 10.1371/journal.pone.0014683 2

[13] C. S. Cleeland, T. R. Mendoza, et al. Assessing symptom distress in cancer patients: The M.D. Anderson Symptom Inventory. *Cancer*, 89:1634–46, 11 2000. doi: 10.1046/j.1533-2500.2001.01023-30.x 2, 3

[14] D. Defays. An efficient algorithm for a complete link method. *The Comp. Journ.*, 20(4):364–366, 1977. doi: 10.1093/comjnl/20.4.364 4

[15] J. Deogun and L. Jiang. Prediction mining–an approach to mining association rules for prediction. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31-September 3, 2005, Proceedings, Part II 10*, pp. 98–108. Springer, 2005. doi: 10.1007/11548706_11 3, 4

[16] S. Di Bartolomeo, Y. Zhang, et al. Sequence braiding: Visual overviews of temporal event sequences and attributes. *IEEE Trans. Vis. Comp. Graph.*, 27(2):1353–1363, 2020. doi: 10.31219/osf.io/mq2wt 2

[17] S. T. Dong, D. S. Costa, et al. Symptom Clusters in Advanced Cancer Patients: An Empirical Comparison of Statistical Methods and the Impact on Quality of Life. *Pain and Symptom Manag.*, 51(1):88–98, 2016. doi: 10.1016/j.jpainsymman.2015.07.013 2

[18] F. Du, C. Plaisant, et al. Eventaction: Visual analytics for temporal event sequence recommendation. In *2016 IEEE Conf. on Vis. Anal. Sci. and Tech. (VAST)*, pp. 61–70. IEEE, 2016. doi: 10.1109/vast.2016.7883512 2

[19] M. Elshehaly, K. Sohal, et al. Creative visualisation opportunities workshops: A case study in population health. In *2022 IEEE Eval. and Beyond-Methodological Approaches for Vis. (BELIV)*, pp. 11–19. IEEE, 2022. doi: 10.1109/beliv57783.2022.00006 2

[20] S. A. Eraj, M. K. Jomaa, et al. Long-term patient Reported Outcomes Following Radiation Therapy for Oropharyngeal Cancer. *Rad. Onco.*, 12(1):150, 2017. 2

[21] G. Fan, L. Filipczak, and E. Chow. Symptom Clusters in Cancer Patients: A Review of the Literature. *Curr. Onco.*, 14(5):173–179, 2007. doi: 10.3747/co.2007.145 2

[22] C. Floricel, J. Epifano, S. Caamano, S. Kark, R. Christie, A. Masino, and A. D. Paredes. Opening access to visual exploration of audiovisual digital biomarkers: an opendbm analytics tool. *arXiv preprint arXiv:2210.01618*, 2022. 2

[23] C. Floricel, N. Nipu, et al. Thalis: Human-machine analysis of longitudinal symptoms in cancer therapy. *IEEE Trans. Vis. Comp. Graph.*, 28(1):151–161, 2021. doi: 10.1109/tvcg.2021.3114810 1, 2, 3, 6

[24] P. Fournier-Viger, U. Faghihi, et al. Cmrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, 25(1):63–76, 2012. doi: 10.1016/j.knosys.2011.07.005 4

[25] P. Fournier-Viger, A. Gomariz, et al. Spmf: a java open-source pattern mining library. *J. Mach. Learn. Res.*, 15(1):3389–3393, 2014. 4

[26] D. Gotz and H. Stavropoulos. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Trans. Vis. Comp. Graph.*, 20(12):1783–1792, 2014. doi: 10.1109/tvcg.2014.2346682 2

[27] S. Guo, F. Du, et al. Visualizing uUncertainty and Alternatives in Event Sequence Predictions. In *Proc. CHI Conf. Human Factors in Comput. Sys.*, p. 1–12, 2019. doi: 10.1145/3290605.3300803 2

[28] S. Guo, K. Xu, et al. Eventthread: Visual summarization and stage analysis of event sequence data. *IEEE Trans. Vis. Comp. Graph.*, 24(1):56–65, 2017. doi: 10.1109/tvcg.2017.2745320 2

[29] C. K. Gwede, B. J. Small, et al. Exploring the Differential Experience of Breast Cancer Treatment-Related Symptoms: a Cluster Analytic Approach. *Support. Care in Canc.*, 16(8):925–933, 2008. doi: 10.1007/s00520-007-0364-2 1, 2

[30] C. W. Huang, R. Lu, et al. A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records. *BMC Med. Infor. and Dec. Making*, 15:92, 2015. doi: 10.1186/s12911-015-0218-7 2

[31] K. Huat Ong, K. Leong Ong, et al. Crystalclear: Active visualization of association rules. In *Int. Workshop on Act. Min., AM2002*, 2002. 2

[32] J. Illi, C. Miaskowski, B. Cooper, et al. Association Between pro- and anti-Inflammatory Cytokine Genes and a Symptom Cluster of Pain, Fatigue, Sleep Disturbance, and Depression. *Cytokine*, 58(3):437–447, 2012. doi: 10.1016/j.cyto.2012.02.015 2

[33] P. Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x 4

[34] W. Jentner and D. A. Keim. *Visualization and Visual Analytic Techniques for Patterns*. Springer, 2019. doi: 10.1007/978-3-030-04921-8_12 2

[35] Z. Jin, S. Cui, et al. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthcare*, 1(1), 2020. doi: 10.1145/3344258 2

[36] H.-J. Kim, I. Abraham, and P. S. Malone. Analytical Methods and Issues for Symptom Cluster Research in Oncology. *Curr. Opi. in Supp. and Pall. Care*, 7(1):45–53, 2013. doi: 10.1097/spc.0b013e32835bf28b 1, 2

[37] P. Klemm, S. Oeltze-Jafra, K. Lawonn, et al. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Trans. Vis. Comp. Graph.*, 20(12):1673–1682, 2014. doi: 10.1109/tvcg.2014.2346591 2

[38] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proc. 22nd ACM SIGKDD Int. Conf. Know. Disc. and Data Min.*, p. 1675–1684. ACM, New York, NY, USA, 2016. doi: 10.1145/2939672.2939874 2

[39] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *The Annals of App. Stat.*, 9(3):1350 – 1371, 2015. doi: 10.1214/15-AOAS848 2

[40] T. Luciani, A. Wentzel, et al. A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine. *Jour. of Bio. Info.*, 112:100067, 2020. doi: 10.1016/j.yjbinx.2020.100067 4

[41] S. Malik, F. Du, et al. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In *Proc. 20th Int. Conf. Intel. UI (IUI)*, p. 38–49. ACM, 2015. doi: 10.1145/2678025.2701407 2

[42] G. E. Marai. Activity-Centered Domain Characterization for Problem-Driven Scientific Visualization. *IEEE Trans. Vis. Comp. Graph.*, 24(1):913–922, 2018. doi: 10.1109/tvcg.2017.2744459 3

[43] G. E. Marai, C. Ma, et al. Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots. *IEEE Trans. Vis. Comp. Graph.*, 25(4):1732–1745, 2019. doi: 10.1109/tvcg.2018.2817557 2, 5

[44] A. Maries, N. Mays, et al. GRACE: A Visual Comparison Framework for Integrated Spatial and Non-Spatial Geriatric Data. *IEEE Trans. Vis. Comp. Graph.*, 19(12):2916–2925, 2013. doi: 10.1109/tvcg.2013.161 2

[45] D. Martens, B. Baesens, and T. Van Gestel. Decompositional rule extraction from support vector machines by active learning. *IEEE Trans. Knowl. Data Eng*, 21(2):178–191, 2008. doi: 10.1109/tkde.2008.131 2

[46] T. Metsalu and J. Vilo. ClustVis: A Web Tool for Visualizing Clustering of Multivariate Data Using Principal Component Analysis and Heatmap. *Nucleic Acids Research*, 43(W1):W566–W570, 2015. doi: 10.1093/nar/gkv468 2

[47] M. Meuschke, U. Niemann, et al. Gucci-guided cardiac cohort investigation of blood flow data. *IEEE Trans. Vis. Comp. Graph.*, 2021. doi: 10.1109/tvcg.2021.3134083 2

[48] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Trans. Vis. Comp. Graph.*, 25(1):342–352, 2018. doi: 10.1109/tvcg.2018.2864812 2

[49] M. Monroe, R. Lan, et al. Temporal event sequence simplification. *IEEE Trans. Vis. Comp. Graph.*, 19(12):2227–2236, 2013. doi: 10.1109/tvcg.2013.200 2

[50] D. Nguyen, W. Luo, et al. Ltarm: A novel temporal association rule mining method to understand toxicities in a routine cancer treatment. *Knowledge-Based Systems*, 161:313–328, 2018. doi: 10.1016/j.knosys.2018.07.031 2, 3

[51] B. O'Sullivan, S. H. Huang, et al. Development and validation of a staging system for hpv-related oropharyngeal cancer by the international collaboration on oropharyngeal cancer network for staging (icon-s): a multicentre cohort study. *The Lancet Onco.*, 17(4):440–451, 2016. doi: 10.1016/S1470-2045(15)00560-4 1

[52] G. Peake and J. Wang. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proc. 24th ACM SIGKDD Int. Conf. on Knowl. Disc. & Data Min.*, pp. 2060–2069, 2018. doi: 10.1145/3219819.3220072 2

[53] C. Plaisant, B. Milash, et al. LifeLines: Visualizing Personal Histories. In *Proc. SIGCHI Conf. Hu. Fact. in Comp. Sys.*, p. 221–227, 1996. doi: 10.1145/257089.257391 2

[54] R. G. Raidou, U. A. van der Heide, et al. Visual analytics for the exploration of tumor tissue characterization. In *Comp. Graph. Forum*, vol. 34, pp. 11–20. Wiley Online Library, 2015. doi: 10.1111/cgf.12613 2

[55] D. I. Rosenthal, T. R. Mendoza, et al. Measuring head and neck cancer symptom burden: the development and validation of the md anderson symptom inventory, head and neck module. *Head & Neck: J. for the Sci. and Spec. of the Head and Neck*, 29(10):923–931, 2007. doi: 10.1002/hed.20602 2

[56] H. M. Skerman, P. M. Yates, and D. Battistutta. Multivariate Methods to Identify Cancer-related Symptom Clusters. *Res. in Nursing & Health*, 32(3):345–360, 2009. doi: 10.1002/nur.20323 2

[57] M. Tandan, Y. Acharya, et al. Discovering symptom patterns of covid-19 patients using association rule mining. *Computers in biology and medicine*, 131:104249, 2021. doi: 10.1016/j.compbiomed.2021.104249 2, 3

[58] L. V. van Dijk, A. S. Mohamed, et al. Head and neck cancer predictive risk estimator to determine control and therapeutic outcomes of radiotherapy (hnc-predictor): development, international multi-institutional validation, and web implementation of clinic-ready model-based risk stratification for head and neck cancer. *European Journal of Cancer*, 178:150–161, 2023. doi: 10.1016/j.ejca.2022.10.011 2

[59] Q. Wang, T. Mazor, et al. Threadstates: State-based visual analysis of disease progression. *IEEE Trans. Vis. Comp. Graph.*, 28(1):238–247, 2021. doi: 10.31219/osf.io/vcskm 2

[60] T. D. Wang, C. Plaisant, et al. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Trans. Vis. Comp. Graph.*, 15(6):1049–1056, 2009. doi: 10.1109/tvcg.2009.187 2

[61] Y. Wang, G. Canahuate, et al. Predicting late symptoms of head and neck cancer treatment using lstm and patient reported outcomes. IDEAS '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3472163.3472177 2

[62] A. Wentzel, G. Canahuate, et al. Explainable spatial clustering: Leveraging spatial data in radiation oncology. In *IEEE Vis. Comp. Graph.*, p. 281–285, 2020. doi: 10.1109/vis47514.2020.00063 2

[63] A. Wentzel, C. Floricel, et al. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Comp. Graph. Forum*, 2023. doi: 10.1111/cgf.14830 2

[64] A. Wentzel, P. Hanula, T. Luciani, et al. Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration. *IEEE Trans. Vis. Comp. Graph.*, 26(1):949–959, 2020. doi: 10.1109/tvcg.2019.2934546 2

[65] A. Wentzel, T. Luciani, et al. Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas. *Radiotherapy and Oncology*, 161:152–158, 2021. doi: 10.1016/j.radonc.2021.06.016 2

[66] K. Wongsuphasawat and D. Gotz. Outflow: Visualizing Patient Flow by Symptoms and Outcome. In *IEEE VisWeek Workshop Vis. Anal. in Health.*, p. 25–28. American Medical Informatics Association, 2011. 2, 6

[67] K. Wongsuphasawat, J. A. Guerra Gómez, et al. LifeFlow: Visualizing an Overview of Event Sequences. In *Proc. SIGCHI Conf. Hu. Fact. in Comp. Sys.*, p. 1747–1756, 2011. doi: 10.1145/1979742.1979557 2

[68] H. Yang, C. Rudin, and M. Seltzer. Scalable bayesian rule lists. In *Int. Conf. on ML*, pp. 3921–3930. PMLR, 2017. 2

[69] J. Yuan, O. Nov, and E. Bertini. An exploration and validation of visual factors in understanding classification rule sets. In *IEEE Trans. Vis. Comp. Graph.*, pp. 6–10. IEEE, 2021. doi: 10.1109/vis49827.2021.9623303 2

[70] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Info. Vis.*, 14(4):289–307, 2015. doi: 10.1177/1473871614526077 2

[71] X. Zhao, Y. Wu, et al. Iforest: Interpreting random forests via visual analytics. *IEEE Trans. Vis. Comp. Graph.*, 25(1):407–416, 2018. doi: 10.1109/tvcg.2018.2864475 2