Patient Cohort Visual Analytics for Post-Treatment Care

by

Carla Gabriela Floricel
B.S., University Politehnica of Bucharest, 2019
M.S., University of Illinois Chicago, 2025

Dissertation

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
at the Graduate College of the University of Illinois Chicago, 2025

Chicago, Illinois

Defense Committee:
Dr. Michael E. Papka, *Chair and Advisor*, University of Illinois Chicago and Argonne National Laboratory
Dr. Fabio Miranda, University of Illinois Chicago
Dr. Khairi Reda, University of Illinois Chicago
Dr. Andrew Johnson, University of Illinois Chicago
Dr. Victor A. Mateevitsi, Argonne National Laboratory and University of Illinois Chicago
Dr. Guadalupe Canahuate, University of Iowa

Acknowledgments

I want to thank my collaborators at the University of Iowa: Guadalupe Canahuate and Yaohua Wang, as well as my collaborators from the MD Anderson Cancer Center: David Fuller and Abdallah Mohamed, for their professionalism and dedication to our collaboration.

Many thanks to the members of the Electronic Visualization Laboratory, especially my peers: Juan Trelles Trabucco, Andrew Wentzel, Stefan Cobeli, Nafiul Nipu, Sanjana Srabanti, Kazi Shahruk Omar, Gustavo Moreira, Carolina Veiga, Leonardo Ferreira, Hossein Fathollahian, Ashwini Naik, Arthur Nishimoto, and Abari Bhattacharya – thank you for making this experience better. Dana, Lance, and Luc have been great contributors to making me feel like I belong to the EVL family. I also want to add my amazing group of friends and fellow students who have joined the program at the same time as myself: Wenshao Zhong, Kostas Solomos, Shubham Singh, and Alex Politowicz.

I want to thank my committee members: Guadalupe Canahuate, Fabio Miranda, Andrew Johnson, Khairi Reda, and Victor Mateevitsi; I have been fortunate to receive so much emotional and professional support from all of you over the years. I especially want to thank my advisor, Michael Papka – you have given me unconditional support and I can never thank you enough for how you have reshaped my academic experience. Furthermore, I want to thank the previous DGS, Barbara Di Eugenio, for her gracious guidance during my academic challenges.

Finally, I want to thank my family: my parents, brother, aunt, uncle, cousin, and in-laws, as well as my friends from home; I lost count of the days you have encouraged me to keep going. I couldn't have done it without your unconditional love. I want to add my puppy, Daisy, she has been a beacon of joy and happiness since she entered my life. Last but not least, I want to thank my husband, Tibi – we were both kids when we left our home country and decided to embark on this journey of grad school. I am so grateful for having you by my side through my highs and lows and for giving me a safe space to be myself. I could not have asked for a better life partner.

Contributions of Authors

This thesis contains the work from three published papers and a paper under review.

1. The paper from Chapter 2, *Opening Access to Visual Exploration of Audiovisual Digital Biomarkers: an OpenDBM Analytics Tool* [59] was completed during a research internship. Andre Paredes, my internship manager, helped with task gathering, which included interviewing users, drafting and editing the paper, and providing the data. All coauthors provided feedback on the prototype and paper.

2. The paper from Chapter 3 *THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy* [60] was based on a class project prototype, where Naveen Kumar helped with the design and development of the prototype. The interface from the paper has changed almost completely, and Nafiul Nipu helped with finishing touches on the front-end code. The rule-mining modeling is based on the code from Mikayla Biggs. Nafiul Nipu, Andrew Wentzel, and Guadalupe Canahuate helped with the editing on the manuscript. Guadalupe Canahuate, Clifton David Fuller, Abdallah Mohamed, and Lisanne Van Dijk provided the data and feedback on the interface. G.E. Marai helped with the research direction and with paper edits.

3. In the paper from Chapter 4, *Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining* [61], Andrew Wentzel provided feedback on the design and helped with paper edits. Guadalupe Canahuate and G.E. Marai helped with the research direction and editing of the paper. All coauthors provided feedback on the prototype.

4. For the paper from Chapter 5, *L-VISP: LSTM Visualization for Interpretable Symptom Prediction in Patient Cohorts*, Guadalupe Canahuate and Yaohua Wang provided the data and modeling results, and together with Andrew Wentzel and Michael Papka, helped with paper editing. Guadalupe Canahuate and G.E. Marai provided guidance on the research direction. All coauthors gave feedback on the prototype.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
OpenDBM	Visual Analytics System from Chapter 2
THALIS	Visual Analytics System from Chapter 3
Roses	Visual Analytics System from Chapter 4
L-VISP	Visual analytics System from Chapter 5
HNC	Head and Neck Cancer
DBM	Digital Biomarker
ACD	Activity-Centered Design
XAI	Explainable AI
ARM	Association Rule Mining
SRM	Sequential Rule Mining
LSTM	Long Short-Term Memory
Bi-LSTM	Bi-directional LSTM
IMV-LSTM	Interpretable Multi-Variable LSTM
PRO	Patient-Reported Outcome
MDASI	MD Anderson Symptom Inventory
MDASI-HN	MD Anderson Symptom Inventory - Head and Neck Module
CC	Concurrent Chemotherapy
IC	Induction Chemotherapy
ICC	Induction and Concurrent Chemotherapy
RT	Radiation Therapy
IRT	Induction Chemotherapy and Radiation Therapy
B	Baseline
W0	Week 0
W6	Week 6
M6	Month 6
M12	Month 12

Summary

Post-treatment decision-making is a process that depends on longitudinal studies of patient cohorts, in which identifying the risk of adverse treatment outcomes is critical to improving personalized care. This process benefits from visual analytics because it can overcome many of the challenges that arise with complex patient cohort data. Visual analytics can help to interpret treatment progressions and outcomes, and stratify cohorts by risk categories, which are crucial for improving treatment decision-making. However, post-treatment cohort analysis uses large-scale, heterogeneous, temporal, multivariate datasets with associated attributes and missing values. Visualization needs to provide scalable, effective methods for cohort analysis at different levels of detail that can uncover patterns and associations among patient attributes that correspond to negative treatment outcomes. Moreover, post-treatment care planning relies on computational cohort data modeling and, as a result, uses both objective and subjective evidence, namely, the clinician's interpretation of the modeling results. Consequently, cohort modeling and analysis depend on collaborations between clinicians and data modelers. Therefore, visual analytics solutions need to facilitate these collaborations and the interpretation and evaluation of modeling results in a clinical context.

This dissertation explores visual analytics techniques for cohort modeling and analysis and applies these techniques to post-treatment decision-making. This work addresses the challenges identified above by designing, developing, and evaluating four application-specific visualization systems in collaboration with clinical researchers and data modelers. I first identified the design requirements for a family of cohort modeling problems in cancer symptom and digital biomarker research. Next, I design several systems that integrate unsupervised modeling for the computational back-end and data visualization for the front-end. I propose novel, custom visual encodings for multivariate temporal cohorts that enable iterative risk assessment across cohort stratifications. A first system, OpenDBM, uses visual analytics for behavioral risk assessment in digital biomarker research, using cohorts with hundreds of modeled attributes, and it was designed for the open-source community. This work proposes

a novel encoding that aggregates multivariate, spatial, and non-spatial temporal attributes on anatomical locations to explain behavioral biomarkers. A second system, THALIS, shifts the focus to clinician-modeler collaborations in head and neck cancer cohort modeling, and to a multi-stage patient monitoring process, namely, during and post-treatment. This system uses scalable visual encodings to interpret attribute associations and introduces a new encoding for evaluating patient outcomes in multivariate, multi-stage time series. A third system, Roses, builds on the previous work, using custom visualizations for the interpretation and evaluation of outcome risk predictions, this time accommodating configurable analytical workflows for clinicians and modelers. The system introduces a visual encoding to summarize multi-stage networks, with temporal nodes, which helps to evaluate patterns and associations in modeled outcome risk components. A fourth system, L-VISP, explores visual analytics for understanding and assessing black-box models in cohort risk prediction, with an emphasis on the design requirements for data modeler activities. To support model evaluation, the system visualizes results for machine-derived (cluster) or user-specified cohort stratifications and introduces custom encodings for weighted associations in multivariate attributes. Together, these systems contribute to data visualization and modeling solutions for the challenges that data modelers and clinicians face during collaborations.

Patient records were used for the cancer research projects. These records contained demographic and diagnostic details, as well as longitudinal symptom ratings. The records were anonymized by our clinical expert collaborators and stored on a private institution cloud. Access was limited and given by our collaborators on a case-by-case basis. For the digital biomarker project, the collaborators provided longitudinal biomarker records. These records were extracted from actor, not patient, videos using a feature-extraction toolkit and stored in a private cloud. There were no personal identifiers in these data, and no research involving human subjects was conducted in this dissertation.

Chapter 1

Introduction

1.1 Motivation

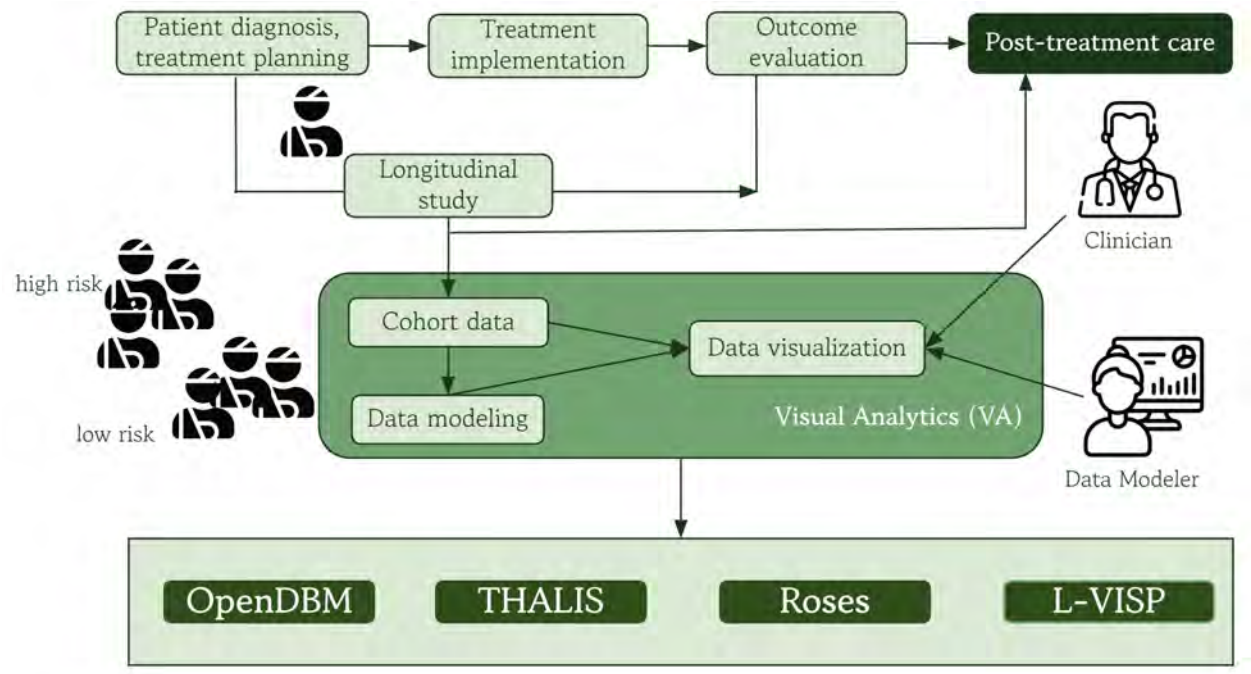

Figure 1.1: Simplified framework for patient cohort post-treatment decision-making and dissertation contributions (i.e., OpenDBM, THALIS, Roses, and L-VISP) within the framework.

Post-treatment care is an important approach to ensure a better quality of life in patients and to combat health relapses after treatment implementation. The one-size-fits-all approach to post-treatment care delivery, based on data from an average patient, does not work well for most health conditions [167]. Post-treatment care research categorizes patients by clinical and demographic factors, as well as treatment progression and outcome, to help tailor personalized plans and anticipate adverse clinical outcomes. This research typically relies on longitudinal studies of existing cohorts, following patients' pathways from initial assessment

and diagnosis through treatment planning, implementation, and outcome evaluation, and even into longitudinal post-treatment evaluations (Fig. 1.1). Consequently, it uses a big data, cohort-based approach, and patient data are generally extracted from multiple sources and have a variety of attributes that must be considered concurrently. In particular, post-treatment cohort analysis is a multidisciplinary, two-step human-machine approach that uses a mix of objective data and subjective clues. More specifically, the first step relies on cohort data modeling and thus on data scientists/modelers' expertise, while the second step depends on clinicians' interpretation of the modeling outcomes (see Fig. 1.1). Thus, this is a complex process, and advanced analytical tools, such as visual analytics tools, that can overcome the challenges mentioned earlier, are in high demand.

An example of an application for post-treatment care is symptom risk research in head and neck cancer. At the MD Anderson Cancer Center in Texas, patients prescribed with radiation therapy often suffer from treatment-induced symptoms. The domain experts who help treat this cohort collect patient-reported outcomes to understand the risk of symptom appearance and severity during and after oncological treatment. After data collection, modeling is used to estimate each patient's risk of symptom relapse, after which data modelers and clinicians assess this information alongside relevant patient details and similarities to other patients, thereby stratifying cohorts by risk levels of adverse outcomes. This helps to better understand health problems and needs in different patient populations. However, patients are not stratified a priori into risk groups, but rather through an iterative, interactive process. This process can be aided by visualizing modeling outputs stratified by attributes of interest within cohorts. Therefore, there is increasing interest in visual analytics tools that support cohort analysis for post-treatment hypothesis generation.

Data challenges. Visual analytics looks for the tight integration of visualization, computational analysis, and interaction for exploratory analysis of complex data, which can be heterogeneous, large-scale, multivariate, and temporal. Such data is extracted from patient records, where many attributes are often associated with each other, are collected at different

stages within patient monitoring protocols, and may contain missing values. Visual analytics can overcome many data challenges, for example, through compact but effective encodings for cohort summarizations or custom encodings that highlight associations and patterns in cohort attributes. More importantly, visual analytics can support interactive methods for inspecting different levels of detail within patient cohorts, by visualizing the cohort overview and detail, the patient of interest within the cohort, and iteratively stratifying the cohort. Thus, through carefully designed visualizations that integrate multiple data facets, clinical researchers can better understand patient health distributions and improve post-treatment decision-making.

Modeling challenges. Modeling cohorts by risk, such as predicting the outcome of a treatment plan, frequently employs machine learning and statistical methods. However, modeling results need to be actionable in a clinical setting. Thus, post-treatment care planning is a multidisciplinary field where clinicians collaborate with data modelers. Visual analytics can support the interpretability of modeling outputs in a clinical setting through mixed-initiative workflows that enable human analysis of machine-derived (modeled) results. At the same time, visual analytics techniques need to consider differences in the mental models of the users; e.g., clinicians are more interested in the clinical, actionable interpretation of the modeling outcomes, which can be applied when treating new patients, while data modelers are interested in the modeling activity and in tools that help them refine modeling approaches. As a result, visual design choices need to account for these aspects and align with domain-specific user activities in multidisciplinary collaborations.

In this dissertation, I claim that visual analytics can offer effective ways to improve risk detection in post-treatment care and to understand outcomes in patient cohorts. To this end, I use an Activity-Centered Design (ACD) [129] methodology to collaborate with domain experts in clinical research and propose four visual analysis systems. ACD is used in the building of these systems to gather tasks, design, and develop software with regular feedback. The systems are evaluated with the collaborators using a pair analytics-based

approach, through case studies and demonstrations, with thematic analysis applied on the feedback to extract lessons learned for future research. This body of work addresses several challenges in visual analytics for cohort analysis in post-treatment research, specifically:

Q1. *How can visualization support cohort analysis?*

Q2. *How to visually represent cohorts and their characteristics, and what interactions to support?*

Q3. *What system implementations work for post-treatment decision-making?*

Q4. *What makes a visual analytics system valuable to biomedical users?*

1.2 Contributions

In this section, I present the contributions of this dissertation, which focus on visual analytics methods that facilitate the analysis of heterogeneous patient cohorts and support knowledge discovery, hypothesis generation, and decision-making for post-treatment care. Taking into account the official IEEE area model for VIS, the contributions of this work are:

C1. **Domain Characterization** (Q1) First, I describe the domain characterization for two medical application domains, symptoms in head and neck oncology and digital behavioral biomarkers in neuroscience. This step identifies the requirements for visual analysis and cohort modeling design.

C2. **Representations and Interactions** (Q2) After that, I present the design and development of several visual analytics systems for post-treatment research. I introduce novel visual encodings to summarize heterogeneous cohorts that exhibit unconventional characteristics, such as multi-stage time series, temporal networks with temporal nodes, weighted associations across temporal and non-temporal attributes, and multivariate spatial and non-spatial temporal patient attributes.

C3. **Integrated Workflows** (Q3) The proposed visual analytics systems integrate visualization with both supervised and unsupervised cohort modeling to leverage model

explanation and evaluation, i.e., human-machine workflows. These modeling methods, together with interactive visual encodings, support pattern, clustering, association, sequence, prediction, and risk analyses in patient cohorts.

C4. **Evaluation** (Q4) Finally, I evaluate the utility of these visualization systems through case studies and feedback from domain experts, code the feedback into key research dimensions, and present the lessons learned from these multidisciplinary collaborations.

The rest of this document is structured as follows:

In Section 1.3 (Background), I present the relevant research background to this thesis.

In Section 1.4 (Related Work), I present the related work relevant to my proposed work.

In Section 1.5 (Methods), I present the methodology for designing, developing, and evaluating the proposed work.

In Chapter 2 I introduce a visual analytics system, OpenDBM [59] (Fig. 1.1), for cohorts and individual patient analyses in digital biomarker research. The chapter describes the domain characterization of digital biomarker modeling for behavioral research and proposes a front-end application to support the understanding of disease outcomes. The project uses custom visual encodings for the analysis of large-scale (>50GB), multimedia patient cohorts. Furthermore, it enables the analysis of set formations based on user-defined attributes. It supports the evaluation and understanding of modeling outputs for large audiences (e.g., academics, clinicians, technical and non-technical researchers). This work was evaluated by academics, industry researchers, and clinical researchers. However, because the system is released as open-source and intended for broad use, and because it addresses a relatively new neurological domain without established standards of care, it prioritizes general cohort analytical tasks and is better suited to data discovery and hypothesis generation rather than to clinical decision-making.

In Chapter 3 I move to the domain application of cohort analysis in head and neck cancer symptom research. This research shifts the focus to clinician-data modeler collaborations for hypothesis-making and patient care decision-making. After presenting the domain

characterization for oncological post-treatment research based on patient-reported symptom measurements, the chapter presents a visual analytics system, THALIS [60] (Fig. 1.1), that introduces new encodings to stratify multivariate, multi-stage cohorts by symptom risk. In particular, this system uses data visualization to evaluate the applicability of association rule mining as a modeling approach for symptom clustering research and to identify risk patterns and associations across different treatment stages. This work was developed for and evaluated by data modelers and clinicians, but it does not account for the differences in analytical tasks between clinicians and data modelers. Furthermore, THALIS supports risk rule modeling on a limited set of results for the entire cohort, without considering key attributes, such as treatment plans.

In Chapter 4, due to the collaborators' interest in the rule mining-based modeling, I extend the work from Chapter 3 through a data visualization system, Roses [61] (Fig. 1.1), that supports the evaluation and understanding of an upgraded risk modeling approach, using sequential rule mining and hierarchical rule clustering. This project uses custom visualizations to predict, explain, and find patterns in outcome risk for cohorts stratified by treatment plans, this time focusing on post-treatment risk prediction. In particular, this work supports configurable workflows on the front end to better accommodate the varying interests in modeling results among data modelers and clinicians. The system was evaluated by both data modelers and clinicians, but due to its configurable front-end design, the amount of analytical workflows was at times overwhelming.

In Chapter 5 I continue with the same domain application in head and neck cancer research, and propose a visual analytics system, L-VISP (Fig. 1.1), that aims to separate the workflows and front-ends for clinicians and data modelers, inspired by the design from OpenDBM, trying to provide more analytical flexibility than THALIS, and trying to offer more workflow structure than Roses. In L-VISP I focus on visual analytics for human-machine analysis workflows centered on data modeler needs, to support more complex and black-box modeling for cohort risk, while also considering the evaluation of the modeling

results in a clinical context, by clinicians. L-VISP evaluates various Long Short-Term Memory (LSTM) methods for symptom risk prediction on user-specified patient sets, stratifying the patient cohort using selected attributes and machine-derived patient clusters/sets. This enables a better understanding of post-treatment outcomes for a target cohort. Custom visualizations assist data modelers and oncologists in developing hypotheses about existing patients and improving healthcare for future patients.

In Chapter 6 I discuss the results of this body of work. For readability purposes, I refer to the visual analytics systems presented in Chapters 2-4 using the following acronyms: **OpenDBM** for the system in Chapter 2, **THALIS** for the one presented in Chapter 3, and **Roses** for the system in Chapter 4 and **L-VISP** for the system in Chapter 5.

Next, I will present a background overview of post-treatment research workflows and the building blocks of this dissertation: the ACD method for developing the proposed systems and the evaluation methodology for these systems.

1.3 Background

In this section, I present an overview of the patient treatment care research workflow, how it connects to post-treatment care, and briefly describe the two medical domains used as case studies in this dissertation. Specifically, I will discuss applications in oncology (a medical branch specializing in the treatment and prevention of cancer, such as head and neck cancer) and neurology (a medical branch dealing with the treatment of disorders and diseases of the nervous system, from conditions such as Parkinson's disease, epilepsy, autism, to depression, and post-traumatic stress disorder). By symptoms in these medical domains, we refer to negative physical or mental features that affect patients' quality of life or cause a health dysfunction. These symptoms are usually a consequence of either the disease of the patient or the prescribed treatment (i.e., treatments including antidepressant medications for patients with depression can cause fatigue and drowsiness, affecting the patient's daily life).

1.3.1 Post-Treatment Care

In many therapeutic areas of medicine, such as oncology, once a patient is diagnosed, a treatment plan is determined. These treatments influence, to a large extent, how the patient's overall health status evolves Fig. 1.1. Usually, during cancer treatments, clinicians closely monitor whether the treatment cures the disease and whether it contributes to a decline in the patient's quality of life [40]. Naturally, this patient-monitoring stage leads to longitudinal studies, which, when collected from hundreds or thousands of patients, can yield rich datasets for predicting more effective treatment plans for new patients. These studies are also used for post-treatment care [117,186], which is mainly due to current therapeutic standards that do not monitor patients as often after treatment completion; as a result, this stage is far less documented and is supported by sparse datasets. Together with treatment outcomes, treatment can provide strong indicators for predicting how a patient's health will evolve post-treatment, an understudied domain in many medical applications. As a result, post-treatment care is highly dependent on the studies conducted during treatment [43].

After treatment, care planning is a complex modeling process due to challenges arising from patient data that is complex, with many attribute types and extracted from different sources. Many medical applications incorporate information on individual differences, in addition to cohort data, to deliver personalized care. In the therapeutic domains of oncology and neurology, identifying individuals at risk of adverse outcomes is fundamental to creating safer therapies and improving patients' quality of life. However, traditional outcome stratification models are not always accurate and need to be evaluated alongside different attributes within patient data [145, 191]. In addition, most patient cohorts are not stratified a priori by risk level, making it more difficult to understand how outcome risk is connected to different cohort characteristics [186]. As a consequence, risk modeling remains an understudied domain in many medical applications, and analytical tools that overcome its challenges and support its advancement are needed.

1.3.2 Symptoms in Oncology

In recent decades, advancements in oncology have led to a greater variety of personalized cancer treatments for head and neck cancer (HNC) patients, with more varied treatment plans and better survival rates [42]. Personalized HNC treatment is a complex, longitudinal process that uses a variety of therapies, including surgery, radiation therapy, induction chemotherapy, or a combination of treatments. For example, a patient can be prescribed with radiation therapy and then with chemotherapy as well, or with both in parallel. Unfortunately, the type of prescribed treatment can cause symptoms during treatment (or acute symptoms) and post-treatment (or late, long-term symptoms), or even permanent health problems which can affect the patient's quality of life [56]. As an example, radiation-based treatments in head and neck cancer can cause dry mouth [153] due to the radiation damage to the salivary glands. Similarly, some patients might forever experience swallowing dysfunctions after treatment completion due to the organ damage produced by radiation [94, 149].

The MD Anderson Cancer Center at the University of Texas documents and quantifies these symptoms through a standardized symptom and quality-of-life monitoring program. This program uses questionnaires collected weekly at the time of the treatment appointment and at longer intervals post-treatment. The questionnaires contain the MDASI (MD Anderson Symptom Inventory) [40] items, a 28-symptom patient-reported symptom outcome used in clinical research. MDASI has thirteen core items, which include symptoms common in multiple types of cancers. MDASI-HN inventory [165] includes nine HNC-specific symptoms as well (e.g., swallowing dysfunction), and six daily life-interference symptoms (e.g., mood). These patient-reported data need a hybrid analytical approach that connects treatments with symptomatic side effects and the patient's health history.

The risk of appearance and the level of severity of these symptoms depend on a variety of factors. In some cases, symptoms can persist or develop even after treatment has been completed, affecting patients' well-being [26]. Monitoring symptom burden post-treatment delivery is more difficult due to less frequent patient visits to the clinic. Thus, there is growing

interest in understanding the risk of patients developing symptoms, the relationship between symptoms and treatment decisions, and the likelihood of symptoms occurring during and/or after treatment, to identify long-term symptoms that limit patients' quality of life.

Since cancer patients can experience a multitude of symptoms that can co-occur or can cause other symptoms, oncology experts are interested in modeling clusters of frequently co-occurring symptoms and in modeling how symptoms are correlated with the prescribed treatment. Symptom-clustering research in oncology is strongly dependent on the patient's diagnosis and treatment and focuses on symptom severity [7, 50, 57]. 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.

1.3.3 Digital Biomarkers in Neurology

One consequence of the rapidly aging population and the rapid evolution of society is an increase in neurological diseases. Thus, there is increasing interest in neuroscience that delves into the underlying mechanisms of various neurological diseases, with a focus on the risk of disease onset and progression. Digital biomarkers (DBMs) are objective, quantifiable, physiological, and behavioral medical measurements, collected using digital devices (e.g., smartphones, smartwatches) from patients [157], which can be used in neuroscience to diagnose patients, to predict disease risk, or to monitor the longitudinal response to treatment.

OpenDBM is an open-source feature extraction toolkit that extracts behavioral DBMs from patient videos. This can help not only to better understand patient behavior during/post-treatment, but also to assess the risk of negative behaviors using the modeled DBMS. However, these DBM measurements, extracted with OpenDBM, generate large-scale, longitudinal datasets with hundreds of attributes and data files and thousands of time points that are difficult to analyze without more appropriate analytical tools. Moreover, because DBMs are a relatively new component in neurology, there are no golden standards regarding what biomarkers are representative of each disorder risk and to what extent; thus, research in this domain has yet to combine different categories of DBMs together for comprehensive patient

risk investigations.

1.4 Related Work

In this section, I present the relevant related work for the four proposed systems offered in the following four chapters. I start with the medical cohort (Section 1.4.2) and XAI (Section 1.4.2) visualization, which is relevant to all four systems, then I continue with more related work relevant to THALIS, Roses, and L-VISP, which provide contributions in cluster visualization Section 1.4.3, symptom clustering Section 1.4.4, and rule visualization Section 1.4.5. Fig. 1.2 provides some examples of existing visual analytics (VA) systems in cohort and medical visualization with respect to some key features to which OpenDBM, THALIS, Roses, and L-VISP contribute.

1.4.1 Medical Cohort Visualization

Visual analysis of patient cohorts often relies on finding connections between different patient attributes from medical records. Electronic Medical Records (EMRs) store longitudinal patient information, often in the form of time series. In general, time-series visualization has utilized point graphs, circle graphs, line graphs [87], parallel coordinate plots [92], or stacked bar charts and their variations [6] to encode nominal, ordinal or quantitative time-oriented data, including in cancer research [146,174,202]. For EMR data, Plaisant et al. have introduced personal patient summary visualization using timelines [158,184,185], or matrix-based representations [52]. Loorak et al. [120] proposed a stacked bar graph approach to explore patient treatment processes, while Baumgartl et al. [14] explored EMR storyline visualizations to detect pathogen outbreaks. Rogers et al. [162] showed outcome trajectories of different patient procedures using line charts. However, most of these approaches are not scalable to large EMR datasets with hundreds of items/patients and multivariate attributes spanning tens of data points. Wong et al. have employed summarization techniques to overcome scale issues via tree-based encodings [196] and Sankey-based representations [86, 195]. At the same time, Karpefors' tendril plot [102] introduced a clustered timeline view

	Scalability	Temporal data	Multivariate data	Multi-stage data	Attribute association	XAI & human - machine analysis	Actionable results	Context & details (patient vs. cohort)	Multiple modeling outcomes or data facets	Clinician - modeler collaboration
EventFlow [123]	x	x	x					x		
TimeSpan [102]		x	x	x			x	x		
CarePre [84]		x	x		x	x	x			
Eventthread [67, 68]		x	x	x		x	x	x	x	
CAVA [172]	x	x	x				x		x	
COCO [109]	x	x	x		x		x			x
RetainVis [98]	x	x	x			x		x	x	
ThreadStates [154]	x	x	x	x	x	x	x	x		
RuleMatrix [122]		x	x		x	x	x			
GUCCI [117]		x	x		x	x	x	x	x	x
Oncothreads [74]	x	x	x		x	x	x			
Precision risk [111]			x			x	x		x	x
TSSIM [162]			x			x	x	x	x	x
DASS [161]	x	x	x		x	x	x		x	x
OpenDBM	x	x	x		x	x	x	x	x	x
THALIS		x	x	x	x	x	x	x	x	x
Roses		x	x	x	x	x	x	x	x	x
L-VISP		x	x	x	x	x	x	x	x	x

Figure 1.2: Visual analytics selection from the related work and their comparison to OpenDBM, THALIS, Roses, and L-VISP. Most of these tackle medical cohort visualization Section 1.4.1 and medical XAI visualization Section 1.4.2. Some discuss rule mining and clustering visualizations Section 1.4.5, Section 1.4.3

of outliers and trends for dense clinical trial data. These works do not contribute to multi-stage timelines, as we do in THALIS and Roses, nor to patient timelines extracted from multimedia sources, with thousands of time points per attribute, as we do in OpenDBM, or to associated timelines, as in L-VISP.

For large scale medical records data Fig. 1.2, such as records that store hundreds or thousands data points with tens of attributes collected over many time points (n >10), patient clinical histories are often visualized using clinical pathway summaries for individual patients [28], or cohort temporal summaries when dealing with larger cohorts (n >700) [196]. In OpenDBM, we had hundreds of attributes collected frame-by-frame from multimedia sources, resulting in thousands of time points. Furthermore, we had to integrate all these attributes and provide a temporal overview, regardless of the data extraction rate (note that video and

audio are extracted at different rates). Related work in medical cohort visualization does not contribute to this. Visual abstractions for temporal cohort data have mostly used matrix-based representations [52], flow-based representations [73, 195], or timelines [14, 76, 77, 158]. Variations in tree-based representations have been used for summarization, ordering, and statistics of event sequences in temporal and clinical data [128, 144, 185, 196]. Other systems for time-dependent cohort data have used PCPs or flow-based representations that use line bundling [14, 136]. Unlike the visual representations proposed in these works, THALIS and Roses offer visual abstractions for timelines with unconventional characteristics, namely multi-stage timelines, or associated timelines where the level of association matters (i.e., weighted associations), as in L-VISP.

Most of the work in medical visual analytics focuses primarily on chronic conditions such as cancer [20], stroke [120], diabetes [47], or infectious disease control due to the COVID-19 pandemic [14,172], and less on neurological disorders. Our work in OpenDBM adopts a new approach to promote individual patient data exploration while building on prior approaches for cohort data exploration. There is work in visual analytics for facial activity and head movement and separately for voice acoustics and speech measurements [37,134,177,178,198], however, some of it does not use video data and none accounts for all four measurement categories together.

Based on the 2021 survey of Guo et al. [79], general analytical tasks in medical applications are cohort summarization and comparison, outcome analysis, and prognosis analysis. We contribute to all except the latter area and address some of the ongoing visualization challenges they list, namely scalability, data heterogeneity, multivariate event sequence visualization, interpretability of machine learning, and associative analysis. In particular, our focus across all proposed visual analytics systems is on developing custom, novel visualizations for cohort stratification and risk analysis to support post-treatment care research. In addition to the data heterogeneity and complexity challenges, this proposal presents solutions for multi-stage cohorts, with temporal, multivariate, and associated attributes. Given

these data characteristics, we propose novel encodings for multi-stage timelines (THALIS, Roses), temporal networks with temporal nodes (Roses), temporal items with weighted associations (L-VISP), and multivariate attributes with spatial and non-spatial characteristics (OpenDBM).

1.4.2 Medical XAI Visualization

In explainable AI (XAI) medical applications Fig. 1.2, cohort analysis tackles clinical statistics from patient records [88, 195], cohort history comparison [20, 39, 204], cohort medical image attribute comparison [108, 132, 161, 190], or cohort stratification for risk analysis [130, 182, 191]. Visual encodings vary widely between these applications, from custom histograms [14], time series plots [73, 86, 97], matrices [52, 77, 97, 128], radial charts [75], etc. Similarly, we use histograms, matrix-based, and time-based encodings to summarize cohort characteristics in all proposed projects, but we propose novel encodings when the data shows less common characteristics in medical cohorts, e.g., multi-stage timelines (THALIS, Roses), a large number of attributes (n >100) (OpenDBM). For XAI cohort analytics, we explore cohort stratifications based on user-selected attributes (THALIS, OpenDBM) and comparison of machine-derived cohorts (clusters) (Roses). When working with large cohorts (>700 patients) where the focus is on finding outlier patients and understanding why they show unexpected clinical characteristics, scatterplot projections are a common way to interpret cohort clusters [55, 135, 136, 192]. We use scatterplot projections in all projects to support outlier detection and patient vs. cohort analysis, which are important considerations for post-treatment research, but in THALIS and Roses, we try to incorporate other cohort attributes in these projections using different visual marks (size, color, shape, etc.), including temporal attributes (Roses) in the scatterplot glyphs. In L-VISP, we use scatterplot projections arranged in a matrix-like representation to provide a better sense of the size and distinct characteristics of a selected cohort vs. the rest of the population.

There is a high demand in medical XAI for applications that can enhance the human interpretation of machine-generated outputs in patient cohorts, as well as for collaborations

between data scientists and medical experts [4,19,29,97,119,146,174,183]. Related work has proposed a variety of applications that incorporate the output of the ML model for patient prognosis, survival prediction, prediction of treatment outcome [73,77,79,97,99,111,114,195], and patient treatment recommendation [52]. Our work across all four projects supports multidisciplinary collaborations between data modelers and clinicians in cohort analyses of patient outcomes and risk. However, in THALIS, Roses, and L-VISP, we use visual analysis to help data modelers understand the model's behavior and outputs for patient cohorts, alongside relevant clinical attributes and treatment. We generally support designs that can be used and evaluated by clinicians with modeling experience for clinical validation, but in L-VISP we separate the clinician and modeler front-ends to support more complex modeling for data modelers. We do not use artificial agents in any of the proposed projects, nor do we explore the collaborations between artificial agents and human agents in human-machine analysis [143].

Visual analytics for black-box cohort modeling is a challenging domain due to the inherently opaque nature of these models. Some tools have been proposed to leverage the collaboration between a clinician and an AI model [105,114,189]. However, many times, clinicians collaborate with data modelers to interpret cohort modeling [67, 121, 174]. In L-VISP, we focus on the data modeler tasks, where the clinician is a secondary user. We use visual analysis to help data modelers understand black-box models and support the collaborative clinical validation of the model predictions with clinicians. We explore a more guided analytical workflow than in our previous work, visually separating user activities across multiple front-end panels. Another notable difference from THALIS and Roses, rather than identifying temporal [61] and non-temporal associations [60] between items, we uncover weighted associations.

1.4.3 Cluster Visualization

Cohort analysis uses various unsupervised learning methods such as factor analysis (e.g., PCA), partitional (e.g., K-means), or hierarchical (e.g., agglomerative) clustering. Cluster

analysis is traditionally visualized using methods such as scatterplots [135], matrices [160], radar charts [130], dendrograms [56], and heatmaps [2]. Temporal clustering is an open problem in symptom research due to the issue of missing data [9, 127]. Furthermore, cancer patient clustering takes into account clinical variables that can include disease stage, treatment plans, medication, toxicity of treatment, etc. [53, 126]. For head and neck cancer (HNC) patients, Gunn et al. [74]. Rosenthal et al. [166] have studied specifically symptom burden for HNC patients by clustering patients based on reported symptom ratings and clinical covariates to find similarities between symptoms and HNC patients using heatmaps and cluster heatmaps. Still, they either do not consider temporal data or analyze patients that underwent specific treatments, respectively, like we do in THALIS, Roses, or L-VISP.

Cohort analysis often relies on domain expert interaction to help support human-machine integrated workflows. For general clustering, several interfaces have provided user interaction for iterative re-clustering and visualization of unstructured cluster data [33, 34], although these rely on generic, abstract encodings such as scatterplots. Other tools have been built to support model building for biostatisticians [48], although these tools do not consider spatial or temporal outcomes and are aimed at statisticians rather than clinicians. In contrast, we focus on both the needs of statisticians (data modelers/scientists) and clinicians in all our projects. Other applications have integrated interactive interfaces with application-specific visual encodings with linked views [8, 66, 190, 192] to support active collaboration between data modelers and clinical researchers. However, none of these approaches consider temporal changes in outcome data, as we do in THALIS, Roses, and L-VISP, or nuanced quality-of-life outcomes (all four projects), and none account for missing data, as we do in THALIS and Roses.

1.4.4 Symptom Clustering

Cancer patients experience multiple co-occurring symptoms often related to each other and to the therapy applied; however, much of the symptom clustering research focuses on single symptoms. In contrast, the term "symptom cluster" (SC) refers to two or more in-

terrelated symptoms that develop together and may or may not be caused by a single underlying mechanism. Several studies have identified symptom clusters in cancer patients [7, 50, 57], though symptom cluster research is still an emerging field. The two most common methods used to determine SC are factor analysis (e.g., principal component analysis, that is, PCA) [104, 107, 170] and cluster analysis (e.g., hierarchical agglomerative clustering) [58, 81, 91, 138]. However, these approaches have not addressed changes in symptoms over time, as we do in THALIS, Roses, and L-VISP, which remains an elusive goal.

Association Rule Mining (ARM), introduced by Agrawal and Srikant in 1994 [5], is an unsupervised data mining method for identifying interesting relationships in data. ARM has been applied to risk management and marketing [80, 103], and more recently in clinical settings [110], although not in symptom clustering. We use ARM in THALIS to identify common symptom clusters across two patient monitoring stages: during and post-treatment. Furthermore, a popular method for clustering time series focused on clinical event sequences in the visualization domain is sequential pattern mining [32, 47, 182]. 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 in Roses uses sequential rule mining (SRM), which accounts for the likelihood that a temporal pattern will be followed. We use SRM to cluster temporal symptom measurements and to find temporal prediction patterns in cancer cohorts.

1.4.5 Rule Visualization

Rule-based modeling is a common approach to create explainable models [116, 199]. In XAI, rule-based explanations are often used to interpret black-box models such as neural networks [141], support vector machines [133], and latent factor models [156]. Rules have also been adopted in the visualization of medical data, with applications in clinical risk prognosis [10, 118] and disease or treatment toxicity prediction [60, 141, 176].

Association rules have been visualized via scatterplots, matrix views, node-link representations, mosaic plots, and parallel coordinates plots, as indicated by two surveys [27,90,95,207],

and also as grouped matrices [83]. Elmqvist and Tsigas [150] used differently sized colored shapes to indicate the information flow in systems of interacting processes, with color indicating the influence of different methods. In biological modeling, both RuleBender [171] and the Kappa environment [62] propose interactive node-link visual representations of rule-based intracellular biochemistry. Visual causal vectors have been used to indicate causality between data elements [188], and animated causal overlays have been used to highlight causal flows and to demonstrate the relative strength of the causal effect [13]. However, when applying rule mining to patient measurements, the results must be evaluated in conjunction with cohort attributes to better support the generated hypothesis. In THALIS and Roses, we try to connect all the dots by linking the common patterns identified by the rules to diagnostic and demographic attributes.

Alongside rule set items, visualization systems also have to integrate relevant rule metrics, such as the support and confidence, to denote the relevance of rules. Yuan et al. [201] found that feature alignment and predicate encoding are influential visual factors for representing rules, arguing that different rule structures strongly influence interpretability and decision-making. Applications that support rule itemsets and the explanation of rule metrics in disease progression have used matrix-based representations accompanied by barcharts and tree-based circular glyphs [10,141]. In contrast, others used node-links to represent temporal rules from diagnosis codes [148]. Similarly, to explain rule results, in THALIS, we use visual marks to highlight important rule metrics that must be accounted for to understand a pattern's impact across a cohort. At the same time, in Roses, we propose a new way to summarize overlapping rule sets using rule clustering and a projection to show associations between rule items.

More generally, rule-based modeling and visualization are common in domains that seek to understand causality. Although our work aims to identify temporal relationships among data based on association rules, these relationships are not necessarily causal and differ from those in biochemical pathways. Although we visualize individual rules in THALIS, we depart

from that in Roses and propose a 2D clustering projection approach to analyze temporal rule clusters using temporal itemsets.

1.4.6 LSTM Visualization

Long Short-Term Memory (LSTM) models are a deep learning method that can deal with complex time series while showing excellent results for a variety of applications, spanning from finance and economics to epidemiology and other biomedical applications [85, 115, 142, 205]. Past work in XAI visual analytics for explaining LSTM prediction models has supported understanding of model performance by exposing models' hidden-state dynamics, evaluating performance metrics, and comparing modeling outputs with ground-truth data [30,36,85,175,197]. These works have experimented with matrix-based visualizations, as well as with timeline and flow-based visual representations of the predicted results [85, 175]. We use similar visualization techniques, but, unlike previous work that focuses on single tasks, we use them to combine these analytical tasks. Specifically, in L-VISP, we expose the model's hidden-state dynamics, evaluate its performance, and compare its predictions against ground-truth data. Another difference from previous work is that we aim to visually combine the outputs of two distinct and complementary LSTM models (i.e., the Bi-LSTM and IMV-LSTM). Finally, unlike some of the previously mentioned work, our work aims to evaluate and compare model results across patient cohorts. Namely, it supports model evaluation on cohorts that are either user-defined (e.g., female patients under 50) or derived from cluster modeling (e.g., a patient cluster with severe symptom ratings).

Our collaborators have previously experimented with unsupervised rule mining, clustering [21, 61], and LSTM-based modeling [186, 187] to predict symptom risk and identify associations in multivariate patient cohorts. In L-VISP, we do this with supervised LSTM methods, which can help to predict symptom risk in new patients at the beginning of their treatment, without necessitating their temporal records during treatment. Unlike previous LSTM-based approaches to symptom prediction, the methods in this work use Bi-LSTM [186] to capture bidirectional temporal dependencies for improved predictive performance, while

incorporating interpretability mechanisms to enhance transparency.

LSTM-based models are a deep learning method that can deal with complex time series data while showing excellent results for a variety of applications, spanning from finance and economics to epidemiology and other biomedical applications [85,115,142,205]. Past work in XAI visual analytics for explaining LSTM prediction models has supported understanding of model performance by exposing models' hidden-state dynamics, evaluating performance metrics, and comparing modeling outputs with ground-truth data [30,36,85,175,197]. These works have experimented with matrix and tabular-based visualizations, as well as with timeline and flow-based visual representations of the predicted results [85, 175]. In addition to exposing the model's behavior by analyzing the features of hidden states, we combine, in L-VISP, the exploration of the model's performance across predicted items (i.e., symptoms) with performance metrics and compare outcomes against ground-truth data. Unlike unsupervised rule mining in THALIS and Roses, in L-VISP, we use LSTM-based modeling to predict symptom risk and identify associations in multivariate patient cohorts. We combine outputs from multiple LSTM-based methods using visualization for symptom modeling and evaluate these methods on patient cohorts that are either model-derived (clusters) or user-specified attribute sets.

1.5 Methods

1.5.1 Activity-Centered Design

The activity-centered design (ACD) approach [129], is a design model that emphasizes user activities, under the principle that *"people's activities around the world tend to be similar, and because people are quite willing to learn things that appear to be essential to the activity, activity should be allowed to define the product and its structure"* [49]. Visual analytics can use this paradigm; however, in many scientific domains, characterizing the application domain poses significant challenges to visual analytics designers and domain experts, which happens because the domain's problems are exploratory and the analyzed data is heterogeneous. As a result, the ACD method is suitable for applications in scientific research because

of the scarcity of trained domain experts and because it supports slow thinking [100]. The ACD method puts value of the tool depending on the user activity, not the number of users of the tool (for example, an application that serves only a couple of researchers in cancer care is not less valuable than an application used by thousands of people for interior design ideas) [131]. In this proposal, I use an ACD-based approach for the domain characterization, design, and development of all the proposed visualization systems. The workflow used for the four projects is, in short, as follows Fig. 1.3:

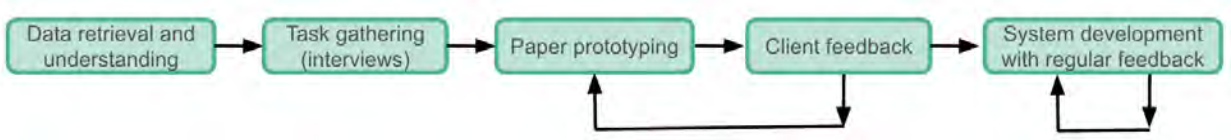

Figure 1.3: Activity-Centered Design simplified workflow

1. Cohort samples are retrieved and analyzed.
2. Visualization designers meet with domain experts, namely, the tool's clients, to gather and understand the user activities and tasks.
3. Designers propose several paper prototypes based on these tasks and narrow down the final design with the clients.
4. During the development phase, visualization designers meet periodically with clients to gather feedback and refine the software prototype.

We used ACD in all the projects presented in this dissertation.

1.5.2 Evaluation

The utility of the visualization systems was evaluated using qualitative methods. The evaluations are based on pair analytics [11], where the visual analytics designer was the navigator of the visual analytics tool, and the tool clients were the drivers of the tool. Although pair analytics requires two participants per session, we organized evaluation sessions with all clients due to limited availability and because we observed that group sessions helped

to generate more hypotheses and feedback. However, these sessions usually had two main drivers, namely a data modeler and a clinical practitioner. The evaluation sessions were conducted online, through screen sharing, starting with demonstrations of the tool and then walking through case studies. The drivers (evaluators) were encouraged to think aloud and make hypotheses while the navigator was driving the interface, and a navigator helper took notes or recorded interactions, feedback, and hypotheses from the domain experts. The evaluators' feedback was extracted from these notes and recordings, and occasionally, from written feedback.

Unlike related work tools that usually employ experiments, case studies followed by surveys, or case studies with datasets from different scientific disciplines to evaluate visual analytics systems, we had certain constraints. We faced limitations in the number and availability of domain experts for these tools (e.g., clinicians who would use these systems should have data modeling experience), as well as data sample limitations. Pair analytics helps researchers with different expertise and from different work environments to naturally interact and share hypotheses. This methodology reduces tacit knowledge that is not verbalized and elicits collaborative analyses while also not constraining domain experts to be too verbose, such as in the case of the think-aloud method. We also adopted a group setting to evaluate how beneficial the end results of these projects were in real-life use cases.

We performed a reflexive thematic analysis [25] on the evaluation feedback gathered from our collaborators from all projects and coded it into three main dimensions of this research:

Actionability. This helped us to understand whether domain experts think the systems are fit to be used in practice.

Perceived Usefulness. This helped us to understand whether the domain experts think they would benefit from the proposed systems.

Trust. This helped us to understand if domain experts trust the system enough to actually use it and consult it during decision-making.

These dimensions are explained in more detail in the discussion of this research in Chap-

ter 6 Section 6.2.

Chapter 2

Opening Access to Visual Exploration of Audiovisual Digital Biomarkers: an OpenDBM Analytics Tool

2.1 Introduction

In this chapter, I present the design, development, and evaluation of a visual analytics system for digital biomarker knowledge discovery and hypothesis-making, which can be applied to cohort post-treatment analysis. The system is open-source and was designed for large audiences (e.g., technical and non-technical researchers, academics, clinicians, and industry researchers). The aim is to facilitate the understanding of disease outcome risk through the visualization of hundreds of derived (mean) and raw patient biomarker measurements that were modeled and extracted from patient videos. This work uses visualization for model understanding and evaluation by determining patient adverse behavioral risk through combinations of patient attributes (i.e., digital biomarkers). For example, in patients treated for depression, behavioral indicators suggesting the persistence of depressive symptoms post-treatment, detectable through digital biomarkers, include a consistently low gaze (e.g., reduced upper-lid raiser activity), lowered head posture (e.g., low-pitch head movements and elevated brow-lowerer activation values), and reduced vocal intensity (e.g., low audio amplitude). This system enables an iterative analysis of patient risk set formations and highlights risk attributes that are correlated with the said sets. To better understand individual patient health status, a separate panel visualizes raw individual patient biomarker measurements. This project introduces a novel visual encoding that connects multiple risk components to anatomical locations. The encoding summarizes spatial and non-spatial multivariate, temporal attributes, and it illustrates the importance of providing context and humanizing the

data when analyzing patient cohort datasets. The evaluation of this system on a 95-video set proves the potential of risk analysis using digital biomarkers in clinical research.

The contents of this chapter were presented at the IEEE VIS 2022 Visualization in the Biomedical AI workshop [59] and the draft is posted on arXiv.

2.2 Motivation

The global market value of digital biomarkers (DBMs) is projected to exceed $7 billion by 2026 [1]. DBMs are objective, quantifiable physiological and behavioral data collected and measured using digital devices, such as smartphones and smartwatches. Like traditional biomarkers, DBMs have clinical value, such as diagnosing disease and predicting disease outcomes. For example, lowered gaze, slowed movements, and sad facial expressions in a patient's behavior can serve as predictors of depressive disorders. However, DBMs introduce additional benefits that exceed traditional biomarkers' constraints, such as capturing longitudinal and continuous measurements that generate large, rich, and complex datasets [12], with hundreds of variables and files, and thousands of time points. Providing clinical researchers with practical tools to derive and interpret DBMs increases their ability to assess changes in health status relevant to healthcare applications [180]. In addition, current DBM research is represented by numerous studies with DBMs that are not validated properly [72] or are duplicates of existing DBMs. Open-source DBM tools are necessary to broaden the validation of DBMs, reduce duplication, and expedite innovation. To support the growing demand for the adoption of DBMs by clinical researchers, more practical tools are required to better inform non-technical biomedical researchers on how to use and identify DBMs [44,82].

In particular, the growing role of DBMs in the therapeutic domains of neurological disorders has sparked renewed interest among clinical researchers in exploring measurable audiovisual changes to better understand how patients feel and behave [17, 35, 96, 112, 140, 157, 159, 169, 181]. Growing open source software projects, such as OpenDBM, are lowering the barrier for non-technical clinical researchers to apply quantitative models, including

machine learning models, to extract audiovisual features in human speech, voice acoustics, head movement, and facial expressions [68, 69]. However, despite open source tools accessible to extract audiovisual features, clinical investigators are burdened with interpreting large (N >100) and complex quantitative datasets [82], with data points that have hundreds of variables collected over thousands of time points.

Given the novelty of DBMs and their still-growing taxonomy and use [41], there is interest among behavioral and biomedical researchers in finding practical tools that can facilitate exploratory analysis for the generation of data-informed hypotheses. This work aims to improve researchers' understanding of the breadth and scope of the hundreds of audiovisual DBMs available for investigatory adoption. We propose a visual analytics interface for the OpenDBM software[1]. Our proposed interface reveals patterns and outliers in facial, head movement, acoustics, and speech DBMs extracted from videos. To our knowledge, this work presents the first audiovisual DBM interactive visualization tool extracted from and made available through open source software.

2.3 Setup and Requirements

The design process followed an Activity-Centered Design approach [129]. Our team held remote meetings for nine weeks with five research groups in DBM therapeutic areas, collectively representing academia, clinics, and industry. Although most of the collaborators were principal investigators with faculty positions conducting behavioral or biomedical research, all were familiar with the OpenDBM software. Throughout this process, the team iteratively gained insight into user approaches to explore mappings between DBMs and conditions and disorders of interest (e.g., major depression and schizophrenia), gathered functional specifications for a DBM interface, and prototyped and evaluated the interface. Due to the wide variety of patient behavior for these disorders, we collected many specific requirements. However, we focused on the following subset of high-level activities to serve all our collaborators and the open source community:

[1] `https://aicure.com/opendbm/`

A1. Show details for any subset of DBM variables available through the OpenDBM pipeline. For instance, for the early detection of Parkinson's disease, head movement measurements are of greater importance than other DBMs, such as voice acoustics. Adaptability to different workflows is an essential factor in open source. Additionally, analyzing hundreds of variables can be highly challenging, and sometimes researchers don't know where to start their analyses. Thus, having the means and freedom to choose what to explore visually is very important.

A2. Support interactive visualizations for both raw and derived data. Visualizing derived mean variables is important to get an effective cohort overview and context for individual patients, while visualizing raw temporal variables supports in-depth analysis of individual patients. This is critical for data quality checking. For example, researchers might want to exclude from their analyses videos where the audio or the patient's face was not captured.

A3. Emphasize trends and outliers in the DBM data. For instance, patients are expected to exhibit negative emotions when discussing unpleasant or uncomfortable subjects. Domain experts should be able to readily observe patterns across patients, which can provide valuable insights for future studies. Furthermore, highlighting correlations between biomarkers is fundamental to improving the understanding of these conditions.

2.4 Visualization Design

The visual system is open source and can be operated through the OpenDBM Github project. It operates independently from the DBM extraction pipeline and serves as a complementary application for visualizing the extracted DBM outputs. We used React with D3.js for the front-end of the visualization system, and Python for the back-end. The interface has two interactive panels: the Cohort Panel and the Individual Panel. These panels are composed of multiple coordinated views that support brushing and linking operations.

2.4.1 Data

Vocal and facial expressions convey emotion and communication behavior and are one of the most researched topics in psychology and related disciplines; as a result, audiovisual DBMs extend from these basic and applied science measurement tools [70]. When a video is processed through OpenDBM, several vocal and facial feature extraction toolkits combine to present hundreds of unique variable categories relevant to four different audiovisual DBM domains: speech, acoustics, facial expression, and head movement. Each audiovisual DBM domain provides two sets of quantitative variables: raw, captured as a frame-by-frame time sequence measurement, and derived, capturing summary statistics on the total collection of frames. These raw and derived variables provide a wide range of objective behavioral cues, such as transcription and lexical richness for speech, jitter and shimmer for acoustics, eye blink and facial tremor for head movement, and facial action units and facial asymmetry for facial expressions. The proposed interface uses these raw and derived variables to display relevant details and statistics about video cohorts and individual videos using two panels: the Cohort and the Individual Panels. The official documentation provides the complete list of DBM variables extracted by OpenDBM.

2.4.2 Cohort Panel

The Cohort Panel (Fig. 2.1) has three main views and functions: to provide a cohort overview based on a selected set of variables, observe variable distributions, and find correlations between variables.

Two query subpanels are available for variable and video ID selection, with the variable query subpanel (Fig. 2.1.A) having three alternative components for each of the three main views (Fig. 2.1.B, D, E). In the video ID query subpanel (Fig. 2.1.C), selected IDs are highlighted in the other views, while unselected videos can be hidden from the rest. All views have accompanying print buttons to generate plot images that can be used in further studies.

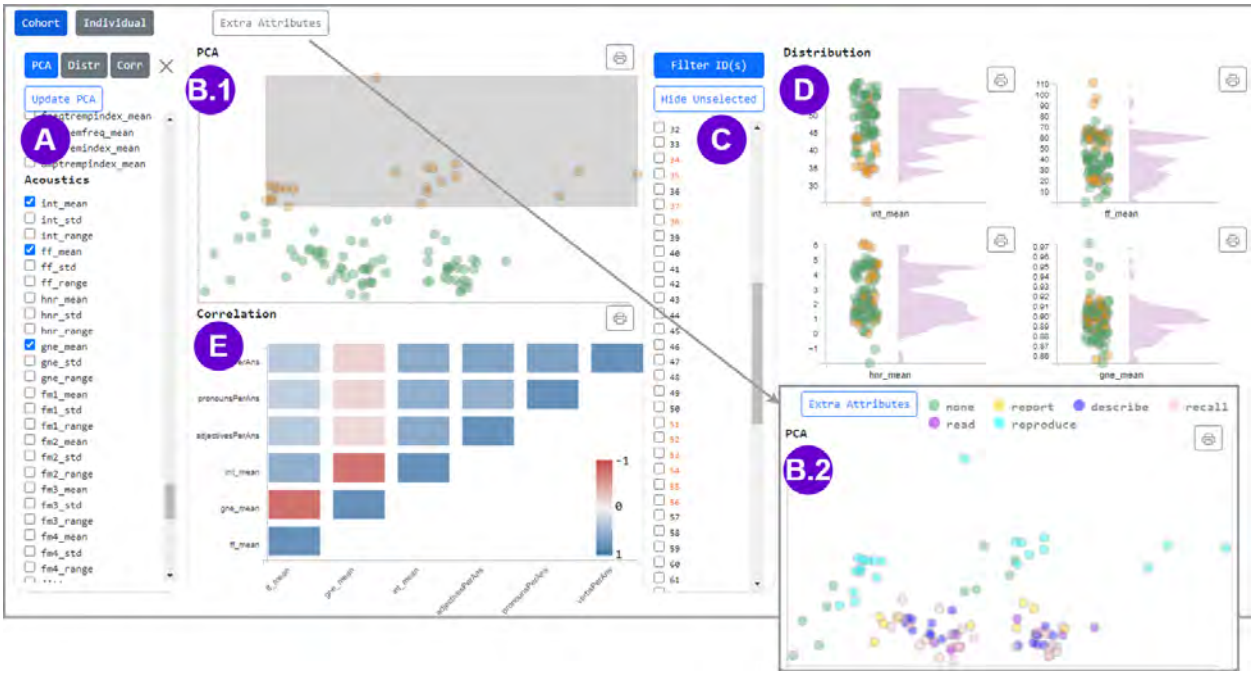

Figure 2.1: Cohort Panel. A) Query Subpanel with three alternative components for biomarker variable selection for views B, D, E. B.1) PCA View that uses a scatterplot to display in 2D video data based on the variable selections in A. B.2) The PCA scatterplot is color-coded based on an extra attribute, namely, the task that was performed during each video. C) ID Query Subpanel, where IDs can be selected to be highlighted in views B and D, and unselected IDs can be hidden. D) Distribution View showing cohort distributions for four selected variables. E) Correlation View displaying pairwise Pearson correlation coefficients for six selected biomarker variables.

PCA View. This view (Fig. 2.1.B.1) uses a scatterplot for a cohort overview by arranging videos in 2D based on a selected set of biomarker variables (A1, A2, A3). The axes correspond to the first two components computed by Principal Component Analysis (PCA) [194]. We employed factor analysis (i.e., PCA) to help researchers get a better sense of the underlying structures in the high-dimensional video data while retaining patterns. The view shows trends and outliers, while brushing interactions highlight selected elements in the Distribution View and ID query subpanel.

Distribution View. This view (Fig. 2.1.D) displays distribution charts for a selected set of biomarker variables (A1, A2, A3). Split into two components, each distribution chart shows one variable distribution throughout the cohort. The left side uses a scatterplot for easier detection of individual videos, while the right side uses a density plot for a concise cohort overview. Hovering on the scatterplot will highlight corresponding elements in the PCA Scatterplot and ID query subpanel, and tooltips will display video IDs and variable

values.

Many times, domain experts study biomarker data while trying to find patterns between different cohorts. During the collection of system specifications, a frequent request was to distinguish between sub-cohorts with varying plans of treatment, health conditions, age range, etc. Thus, if such data are available, the system will color videos (Fig. 2.1.B.2) based on that list of extra attributes in both the PCA and Distribution Views (A3).

Correlation View. This view (Fig. 2.1.E) contains a correlation matrix to emphasize the interrelationships between a selected set of biomarker variables (A3). The matrix shows the coefficients computed using Pearson's Correlation [16] method. The accompanying tooltips display the coefficient values for each pair of variables.

2.4.3 Individual Panel

This panel (Fig. 2.2) has five coordinated views, showing facial, movement, and acoustic temporal variables with some derived variables, and correlations between raw variables.

The panel features an ID query subpanel (Fig. 2.2.E), where one video can be chosen to display its DBM data. An interactive timeline (Fig. 2.2.B.1) splits the raw data into 20 time frames and highlights the corresponding time interval in the other views upon change. Similar to the cohort panel, each view features print buttons.

Figure 2.2: Individual Panel. A) Head Sketch View, which uses custom colored masks to display facial asymmetry, pain expressivity, overall expressivity (A.3), action unit (AU) intensity (A.1), and head movement (A.3) biomarkers. In A.2, the AUs' numbers are displayed, and the AUs involved in anger expressivity (bottom selection) are marked with purple highlights. B.1) Timeline used to split the data into 20 time intervals. Upon change, the timeline will update the mean values displayed in A and highlight the selected interval in views C, D, and F, while the corresponding frame intervals will be shown in B.2. C, D, F) Facial, Movement, and Acoustics Views that display temporal data distributions for selected biomarker variables. E) ID Query Subpanel, where one video can be chosen to display its DBM data. G) Correlation View that shows a pairwise Pearson correlation coefficient for a set of variables.

Head Sketch View. This head sketch (Fig. 2.2.A.1,2,3) supports derived value abstractions for facial activity and head pose biomarker data (A2, A3). Four alternative masks that use custom heatmaps can be applied to support facial biomarker visualization. The Asym mask uses colored glyphs to highlight facial features' asymmetry values. The Pain mask uses a colored glyph to highlight the pain expressivity values. The Expr mask uses colored glyphs to highlight the overall, upper, and lower face expressivity values (Fig. 2.2.A.3). The AUs mask (Fig. 2.2.A.1) highlights action units' intensity values using arrows pointing to the direction of the corresponding facial features' movements [54]. As emotions are expressed through a combination of action units [163], the AUs mask has a complementary layer (Fig. 2.2.A.2) that indicates the set of AUs that get activated for each of the seven available emotions. Upon hovering, the action units' numbers are visible on the facial sketch. In addition to facial activity masks, a head pose (Mov) mask (Fig. 2.2.A.3) is also available and indicates head movements of yaw, roll, and pitch using colored pairs of arrows for each action. By default, the head sketch masks show derived values. However, this view will show the mean values for the selected time frame when the timeline is updated.

Facial, Movement, and Acoustics Views. The Individual Panel features one view for temporal data for each of the three biomarker categories, namely, facial activity (Fig. 2.2.C), head movement (Fig. 2.2.D), and voice acoustics (Fig. 2.2.F) (A1, A2, A3). Each view uses histograms to display temporal distributions for selected variables in the accompanying query subpanels, with the X axis representing time. The placement of these views facilitates the discovery of patterns among biomarkers. When the timeline is updated (Fig. 2.2.B1), the corresponding time interval is highlighted in red on all histograms, while the frame intervals are visible for each biomarker category (Fig. 2.2.B.2).

Correlation View. This view (Fig. 2.2.G) uses a correlation matrix to support the same functionality as the Correlation View in the Cohort Panel (A3). However, here, the Pearson correlation is computed using individual longitudinal data.

DBMs, as an interdisciplinary tool in AI and medicine, have a far-reaching potential in

basic and applied sciences that will continue to drive qualitative domain experts' interest in quantitative DBMs. Therefore, when we started this project, we realized the importance of making the interface accessible to a broad audience, including qualitatively trained domain experts. As a result, we used conventional visual encodings suitable for various visual literacy environments, such as scatterplots, histograms, density plots, and matrices. Additionally, we experimented with custom visualizations, such as the facial and head pose masks, taking advantage of the power of visual mapping techniques and trying to make better sense of the behavioral measurements from facial activity and head movement. We chose to first experiment with these two DBM categories because they were of interest in many use cases during our design requirements interviews.

2.5 Evaluation

We evaluated our system using two case studies that involved three domain experts and a video-simulated actor dataset. Using simulated actor datasets is a well-accepted practice for capturing the prototypical representation of the multiple and complex emotional states of psychiatric and neurodegenerative disorders [15,31,71,155]. A video-simulated actor dataset was generated and used for our case studies. It included 95 videos, all under two minutes, of one adult male actor, instructed to perform five categories of tasks while simulating major depressive disorder for some videos. The five tasks included: describe a picture, describe a memory, describe your day, read a passage, and reproduce a vowel sound or a basic facial emotional expression. The evaluation was conducted through video conferencing using screen sharing and the think-aloud method.

2.5.1 Case Study I: Cohort Analysis

For this case study, the domain experts were first interested in discarding videos with no relevant information from the cohort (i.e., videos where the actor does not perform any tasks). The investigation started with the PCA View (Fig. 2.1.A, B.1), where audio intensity, fundamental frequency, and glottal to noise excitation ratio, all derived audio biomarker

variables, were chosen as parameters to display the videos in 2D. Most videos were grouped towards the lower-central part of the view. Next, the previous set of parameters was used for the Distribution View (Fig. 2.1.D). After brushing the upper part of the PCA scatterplot, the distribution charts revealed that the selected videos had the lowest audio intensity values of the cohort. When brushing the left or right outliers, the distribution charts showed that the selections belonged to opposite cohort extremities for fundamental frequency and glottal to noise excitation (GNE) ratio, which aligned with previous research [139]. When using the option to color the videos by task category (Fig. 2.1.B.2), the scatterplot showed that the upper and left sides of the view were made up of videos with no audio or videos where the actor reproduced facial expressions or vowels. Lastly, the domain experts wanted to check the correlations between these parameters and speech DBM variables, such as the number of pronouns, verbs, or adjectives used per task (Fig. 2.1.E). The Correlation View revealed strong positive correlations between the DBM speech variables and the audio intensity. This case study showed the interface's ability to display relevant trends and outliers in subsets of derived DBM variables of interest (A1, A2, A3) and helped the evaluators detect videos without acoustic DBM data.

2.5.2 Case Study II: Individual Analysis

This case study started with the exploration of raw data for a video in which the actor performs a picture description task (Fig. 2.2.E). The facial emotion expressions were chosen as parameters for the Facial View (Fig. 2.2.C). The histograms showed high spikes for most negative emotions, such as distress, anger, fear, and sadness, implying that the task entailed describing a negative impact image. Hence, the domain experts were interested in observing facial action unit changes over time, since they are connected to facial expressions, so they used the Head Sketch View (Fig. 2.2.A.1) and the timeline for this task (Fig. 2.2.B.1). After applying the AUs facial mask for action units and the Mov mask for head pose changes (Fig. 2.2.A.3), they investigated different time frames using the timeline, and observed that, indeed, at most times, the action units for negative emotions were active (Fig. 2.2.A.2). Next,

the evaluators checked correlations between head poses and emotions using the Correlation View (Fig. 2.2.G), and observed that, surprisingly, roll head movements were negatively correlated with most emotions, while the other head poses did not show any particularly strong correlations to emotions. Curiously, the evaluators watched the actual video, which showed the actor describing a picture of a building on fire. This case study showcases the interface's ability to show patterns in selected raw DBM variables for one video (A1, A2, A3) and helped the evaluators detect a video where negative emotional impact was present.

2.5.3 Expert feedback

We received positive feedback for the interface, considering that previously, domain experts were limited to manual and laborious means of inspecting video data: *"Very cool, so much better to use for the analysis we did last year, huge time saver"*. When asked what they found most useful for their own research studies, most people pointed out the interface's ability to delineate subcohorts using different colors (Fig. 2.1.B.2) *"very excited about the color coding"*, as well as the histograms (Fig. 2.2.C, D, F) *"I had someone looking away from the camera, this is actually picking up their data."* However, one evaluator indicated that it would be helpful to *"input our own data [in the interface] to work with the [bio]markers,"* as it could speed up symptom research for different disorders.

2.6 Discussion

The proposed interface successfully supports all main activities (A1-3). Both the Cohort and Individual panels provide flexibility in choosing desired DBM variables for evaluation (A1). The Cohort Panel supports the analysis of raw DBM variables, while the Individual Panel supports the analysis of derived DBMs (A2). Finally, all visual encodings show cohort trends in the DBM data (A3), such as the PCA view, which shows clusters and outlier patients with respect to selected DBMs, and the distribution view, which shows cohort statistics for DBMs. In contrast, the facial, voice, and acoustics views from the Individual Panel show temporal trends for each DBM for a given patient.

Overall, this system introduces visualization approaches to domain experts in the therapeutic fields of DBM. Some of the limitations of this system are that the Cohort Panel suffers from scalability issues (i.e., the scatterplots) when it comes to large cohorts of hundreds of videos, while several DBM normative ranges are currently being researched, and could better inform domain experts on abnormal patterns and help generate more accurate hypotheses about an individual's health status.

This visualization approach begins to address the need for transparency behind data quality control and quality assurance of these integrated open-source toolkits. The open-source space is a natural fit to drive domain expert users to test and ratify best practices. Future work can build upon our visualization approach by further improving visualizations of raw data to check data quality from new and unvalidated toolkits.

2.6.1 Research Questions

Q1. *How can visualization support cohort analysis?* Using an ACD methodology, we interviewed domain experts in behavioral biomarker research from industry, academia, and clinics. This helped us to collect design requirements for this domain and key activities that all researchers shared. Data visualization provided a more accessible way to traditional methods of analyzing large datasets extracted from patient videos (such as looking at the video in parallel to analyzing one file at a time). These datasets contain hundreds of variables, hundreds of files, and thousands of time points (frame-wise measurements) per video. Analyzing this amount of data would be overwhelming for a single patient; needless to say, it would be even more so for a cohort.

Q2. *How to visually represent cohorts and their characteristics, and what interactions to support?* This project proposed a separation between the analysis of cohort data and individual patient data. This design option seemed natural since the derived cohort variables are extracted separately from individual frame-by-frame variables. We employed a coordinated multiple-view design in the Cohort and Individual Panels to connect different categories of DBMs (facial, movement, voice). We proposed a novel visual encoding to humanize these

measurements when possible, connecting facial activity data, which were spatial, with head movement data over time.

Q3. *What system implementations work for post-treatment decision-making?* The data for this project was already modeled and extracted using the OpenDBM feature extraction toolkit. The scope of this work was to provide a visualization interface to enhance the analysis of the resulting datasets. This interface could be used both during and post-treatment, as patients could be recoded during both stages. For post-treatment care, the proposed visualization system could highlight behavioral treatment outcomes (machine-derived biomarker values), so that clinicians (humans) would make better-informed decisions about treating current and future patients (visualization for ML). We ensured that the OpenDBM visualization interface supports user activities by evaluating incremental prototypes with our clients throughout the development phase.

Q4. *What makes a visual analytics system valuable to biomedical users?* The OpenDBM interface supported research collaborations for the OpenDBM toolkit model understanding and evaluation. We evaluated this project with research groups in academia and industry, the latter being a multidisciplinary group, with clinical researchers and data modelers. The visualization system was able to help biomedical researchers to evaluate how well the feature extraction toolkit extracts the DBM variables (if what the patient videos correlated to the behavioral measurements extracted by the toolkit), and to better understand behaviors in patients through machine-derived measurements (human-machine analysis and workflows).

Takeaways. One of the main findings about the OpenDBM visualization system was that designing a cohort visual analytics system for open source was challenging. That was because researchers from different institutions had different interests in the data, depending on the disease they investigated. At some point during the prototyping phase, we had to draw a line and narrow down the capabilities of the visualization system so that it satisfied most of the common user activities, as well as the project timeline. Conducting the domain characterizations, we observed that behavioral DBM research was in its infancy stages, and

still is. As a result, research in post-treatment decision-making was challenged by the missing standards of the DBM values to detect disease-specific symptoms. As a result, in this project, DBM visualization served for hypothesis-making more than decision-making.

2.7 Conclusion

OpenDBM introduced a visual analytics system for cohort analysis in neuroscience, in the context of digital biomarker research. This work presented the domain characterization for an application that is suitable for post-treatment care, which supports the understanding of cohort risk stratification and cohort characteristics when dealing with large-scale, multivariate, multimodal, unstratified patient data. The system introduced two panels to visually analyze both cohorts and individual patients. In addition, the system was used to visualize model outputs (i.e., digital biomarker measurements modeled by the OpenDBM toolkit) for improved understanding and evaluation. It was designed to engage a broad audience audience, including academics, clinicians, and industry researchers.

In the next chapter, I will introduce a visual analytics system for a different medical domain, namely oncology, that focuses on user activities in the context of cohort risk stratification and patient risk assignment during and after the implementation of treatment. This next work will shift from large audiences (e.g., industry, academia, clinics, technical and non-technical users) to a more targeted audience, namely the collaborations between clinicians and data modelers, and as a consequence, to more fine-grained tasks and application-specific cohort visualizations.

Chapter 3

THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy

3.1 Introduction

This chapter introduces the design, development, and evaluation of a visual analytics system, THALIS, for patient cohort risk stratification in the domain application of head and neck cancer symptom research. For example, a patient prescribed radiation therapy for a neck tumor may develop swallowing dysfunction after completing treatment due to radiation-induced toxicity affecting the neck area surrounding the tumor. As a result, clinicians will look at existing cohorts to find similar patients to their current patient to discover the cohort attributes that correspond to swallowing problems after treatment completion. In other words, the clinician will categorize the cohort into high and low risk of swallowing problems. Patient attributes that might be relevant to this risk could be older age, smoking status, and swallowing difficulty during treatment. To help with this research, THALIS applies association rule mining for risk modeling and targets model understanding and evaluation through data visualization. Specifically, the system shows how user-specified attributes determine cohort stratification into different categories of risk (i.e., burden and progression of symptoms). Furthermore, THALIS ensures the domain sense and actionability of the association rule modeling results, explains the composition of outcome risk by linking outcome components (e.g., progression of symptoms and patient characteristics), and shows the association between the said components. This work introduces custom, scalable encodings for multivariate, temporal, multi-stage cohorts, targeting multidisciplinary collaborations between data modelers and clinical practitioners with modeling experience. A novel encoding,

namely the filament plot, supports cohort symptom pattern and outlier detection by summarizing multi-stage time series. The system was evaluated with domain experts in symptom modeling and oncology in a cohort of 699 patients with head and neck cancer.

The contents of this chapter were presented at IEEE VIS 2021 in the whole paper track, in the Applications area [60]. During the preliminary steps of the design and implementation phase, this work was presented as a poster at the IEEE VIS 2020 Poster Session.

3.2 Motivation

Thanks to advances in therapeutic care, nowadays cancer patients survive for years post-treatment. However, they are plagued with long-lasting or permanent residual sequelae, whose severity, rate of development, and resolution post-treatment vary greatly between survivors [38,192,193]. At the same time, patient questionnaires and electronic health records storing patient-reported responses are leading to larger than ever oncological symptom data collections, with hundreds or reports that store tens of multivariate attributes, collected over tens of time points. These symptom data collected from cohorts of patients [165] offer essential information that can improve clinical decision-making and individual care delivery both during and post-treatment [130, 147], and could be critical for the efficient detection and resolution of longitudinal symptoms. These factors have led to demands from healthcare providers to better understand symptom development and prevention based on cohort data.

However, meaningful interpretation at the individual patient level of symptom repositories is plagued by data and analysis issues that have prevented their practical use in clinical care. These issues include the wide range of symptoms, their partial co-occurrence, their variability between patients and over time, and, in the case of head and neck cancers (HNC), and other cancers that employ radiation therapy, further symptom dependency on the anatomical location of the tumors and the course of treatment prescribed. Furthermore, symptom research analyzes either individual symptom evolution or symptom clusters. Symptom cluster research aims to identify co-occurring symptoms and to understand the

underlying mechanisms that drive these clusters using machine learning [137, 170]. At the same time, preliminary HNC analyses based on factor analysis (e.g., PCA) have not always been replicable on patient datasets [21] with hundreds of data points. Consequently, there is growing interest in alternative machine learning approaches for this type of longitudinal, multivariate data. Furthermore, these approaches need to make sense in an applied healthcare setting and need to be actionable by clinicians and radiation oncologists. Therefore, there is a growing demand for mixed human-machine analysis and a need to facilitate and balance computational and human effort for symptom data analysis.

In this chapter, we present an interactive data mining environment to support the clustering, exploration, and analysis of longitudinal symptoms collected from cohorts of cancer patients. Our approach intertwines association rule and factor analysis unsupervised models with custom visual statistical encodings and visual analysis, in order to estimate the longitudinal symptom evolution of an individual patient, in the context of cancer therapies and similar patients. This visual analysis methodology was successfully developed through an interdisciplinary, remote, multi-site collaboration.

The contributions of this work are: 1) a description of the application domain data and tasks, with an emphasis on the multidisciplinary development of clustering tools for symptom data in cancer therapy; 2) the design of a novel blend of data mining and visual encodings to predict and explain longitudinal symptom development, based on an existing cohort of patients; 3) the description of custom interactive encodings: interactive association rule diagrams, filaments, and percentile heatmaps; 4) an implementation of this approach in a visual symptom explorer named THALIS: THerapy Analysis of LongItudinal Symptoms (Fig. 3.1); 5) an evaluation by domain experts of the resulting mixed workflows and encodings over an existing head and neck symptom repository; 6) a description of the design process and of the lessons learned from this successful collaboration.

3.3 Design

3.3.1 Setting

Our system was developed through a remote collaboration between three different research groups over the course of one and a half years. During this collaboration, our visual computing research group worked closely with oncology and data mining experts. The core team includes three radiation oncology experts with clinical and research experience, a senior data mining expert, a data mining graduate student, and a team of visual computing researchers with varying expertise. Our team met weekly to produce informative mixed machine-human analyses of longitudinal symptom data collected from head and neck cancer patients who were undergoing treatment at the MD Anderson Cancer Center in Houston, TX. This work is part of a longer, six-year collaboration between the lead investigators who had been working together on a series of related projects using oncology patient data.

Due to the long-term and remote nature of our collaboration, which spanned three sites, but within the same time zone, we employed team science principles [130]. Our design process blended an agile design process based on regular team meetings along with an Activity-Centered Design (ACD) approach to the design of the visualization system [129]. The ACD paradigm is an extension of human-centered design, with emphasis on user activities and workflow.

Through a series of iterations, the research team met to define functional specifications, prototype the interface, evaluate prototypes, and decide on changes to the specifications. Furthermore, because this approach was designed to develop interfaces that can be shared and designed remotely, our approach proved to be an effective alternative to methods that rely on in-person group meetings during the COVID-19 pandemic. Additionally, because the ACD paradigm is focused on supporting the collaborators' activities, our collaborators stayed motivated to continue to attend meetings even during circumstances that required remote meetings and exceptional working conditions for clinical practitioners [152].

3.3.2 Activity and Task Analysis

THALIS serves oncologists who have experience in symptom research. Our collaborators also had extensive experience using basic unsupervised machine learning methods such as factor analysis through principal component analysis (PCA) [194], which they had used to determine that symptom burden varies over time and over patient populations. However, PCA results obtained on smaller datasets did not generalize on datasets with hundreds of data points with tens of attributes, so over the course of the project, the group's interests shifted from PCA to alternative approaches. Furthermore, predicting the symptom trajectory of an individual patient in the clinic based on the population data in the repository was not possible due to a lack of an appropriate computational approach. In addition, oncologists expressed frustration due to repeated failures of patients to follow instructions to reduce symptom burden, such as following a prescribed regimen of swallowing exercises or taking the prescribed pain medication. The physicians felt that having the means to explain to patients a predicted symptom trajectory, in the context of other patients, could be beneficial in terms of adherence to therapy.

Taking into account evolving requirements and specifications, we summarize the project activities and their corresponding visual analysis tasks as follows:

A1. Analyze alternative symptom clustering approaches and apply them to an existing symptom dataset

- T1.1. For each approach, show similar patients, based on symptom severity at a specific time point
- T1.2. For each approach, detect symptom correlations during and post-treatment
- T1.3. For each approach, detect patient outliers and trends

A2. Analyze longitudinal symptom progression in the dataset, with particular emphasis on the acute versus late stage of symptoms, and different therapy options

- T2.1. Analyze patient symptom trajectories as a whole, by therapy, and by stage

- T2.2. Compare symptom trajectories between patients, by therapy type
- T2.3. Summarize symptom ratings for the entire cohort, by stage

A3. Map an individual patient to its relevant cohort, and explain their longitudinal symptom trajectory in the context of the cohort in an actionable manner

- T3.1. Show an individual patient in the context of the cohort
- T3.2. Display demographic and diagnostic data, and indicate patients with similar diagnostic attributes
- T3.3. Display which anatomical locations are affected by each symptom
- T3.4. Filter a patient's symptoms by association rule

Our evaluation describes example workflows centered on these activities. Non-functional requirements included a request for the A3 data to be displayed in a manner amenable to audiences with low visual literacy, awareness of variability in symptom ratings across patients, and awareness of missing data.

3.3.3 Data

According to the ACD paradigm for data visualization [129], the project requirements were based on a starter dataset, which was then expanded during the duration of the project, as more data became available. Patients who had completed fewer than two questionnaires were not included in the analysis. The final dataset included 699 HNC patients.

For each patient, two types of information were recorded: 1) patient demographics and diagnostic data, which covered three attribute types: quantitative data (e.g., age, weight, or the total radiation dose); ordinal data (disease stage), and nominal data (e.g., therapeutic combination); and 2) longitudinal symptom data, as time-series attributes with quantitative values (ratings for 28 symptoms) over a maximum of 12 time points. In this longitudinal assessment, 28 symptoms are considered, 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 issues). The symptoms are rated on a 0-to-10 scale ranging from *"not present"* (0) to *"as bad as you can imagine"* (10) for the specific elements of the core and the HNC, and from *"did not interfere"* (0) to *"interfered completely"* (10) for the interference elements. Each patient rated all 28 symptoms during a questionnaire completion (time point).

The dataset included a total of 12 time points. Due to the desired longitudinal aspect of the analysis, we separated these points into three categories: baseline (week 0), acute stage (on-treatment period), and late stage (>= 6 weeks post-treatment). For acute time points during treatment, data were collected every week (at most 7 weeks), while post-treatment time points data were collected at lower granularity, at 6 weeks, and 6-, 12-, or 18-month post-treatment. Previous time point values were substituted for missing values; missing baseline values (i.e., for the first time point) were marked with 0. Patients with no symptoms recorded during the acute or late phases were not included in the analysis for that time frame.

3.3.4 Front-end Design

The design of THALIS followed a parallel prototyping approach [51], a method proven to lead to better design results by opening up the visual encoding and interaction space, which in turn generates more detailed and constructive feedback than in serial prototyping. THALIS was implemented in Python and JavaScript with the D3.js library [23]. The design is based on coordinated multiple views of the data, to support both layering and separation of information and workflow components, and the ability to integrate visually heterogeneous data. A main clustering panel allows the analysis of patient groupings based on similarity (Fig. 3.2), respectively, the analysis of symptom groups via association rule mining (Fig. 3.1.A). A second main panel supports the longitudinal analysis of patient symptoms (Fig. 3.1.B), in coordination with the other panels. A third main panel explicitly supports the context anal-

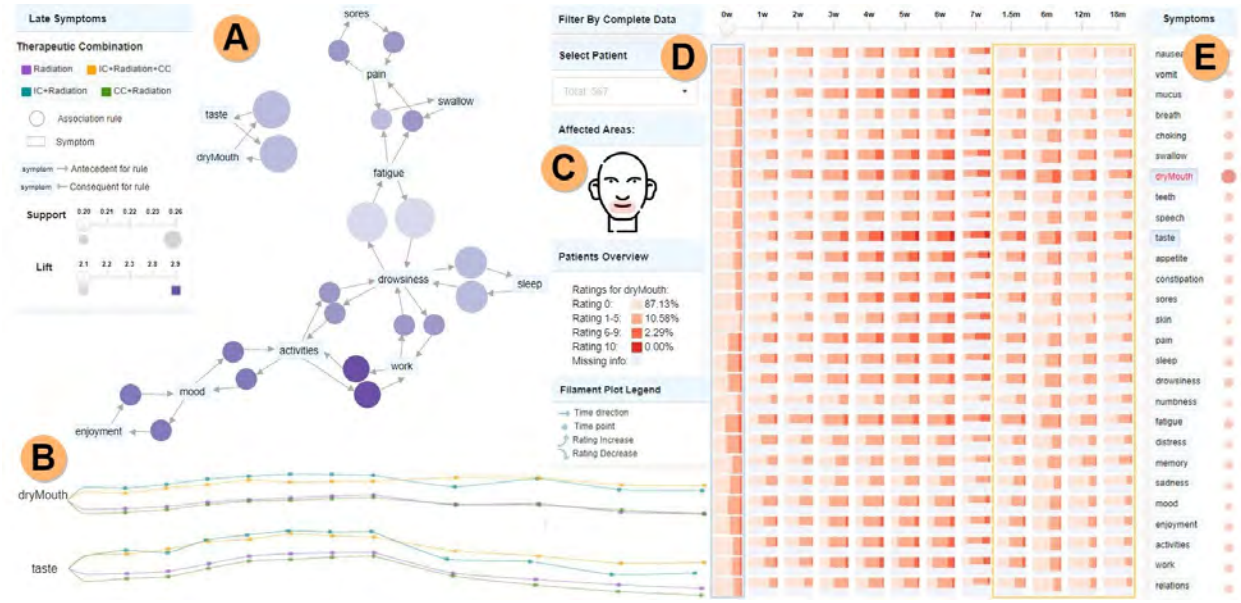

Figure 3.1: THALIS analysis of longitudinal symptom data. A) Association Rule Diagram panel, showing here association-rule-mining (ARM) relationships among the most frequent late-stage symptoms; rules are represented using bubbles, with size and color encoding the support and lift metrics. B) Symptom trajectory panel—filament plots encode the mean rating values per therapeutic combination, with more frequent observations in the acute stage (left-end) than in the late stage (right-end). C) Sketch of areas affected by the selected symptoms (dry mouth and taste). D) Cohort symptom panel showing, via summarization with shade and height, the percentile rating distribution. E) Correlation matrix showing associations with the selected symptom.

ysis of cohort symptom data (Fig. 3.1.C, D, E). The panels are connected through brushing and linking, and through explicit filtering operations.

Clustering Panel

Due to the interest of the experts in activities A1 and A3, the clustering panel shows a therapy cluster view of patients (Fig. 3.2). Alternatively, the panel shows an association graph view of related symptoms (Fig. 3.1.A), illustrating the two main clustering approaches of this project (A1). These views are coupled with computational modules for clustering.

Therapy Cluster View. In prior research, the clinicians had analyzed a subset of the patient data using factor analysis. They had identified different groups of patients with high, medium, and low symptom burden, depending on the therapeutic combination, which they had illustrated via heatmaps and dendrograms. However, they were also aware that the heatmap representation did not represent the outliers well in the patient dataset, nor did it support the individual patient analysis well, and they were also not confident in the

Figure 3.2: Custom scatterplot of patients at a specific time point, for a selected rating severity. The left position is associated with a lower symptom burden, calculated based on the symptoms selected in the list. Shape, size, and color encode demographic, diagnostic, and therapy features (see legend). In this example, highlighted patients correspond to the high rating severity group, indicating that the three symptoms selected (mood, enjoyment, and walk) severely affect the vast majority of patients across all therapies, genders, and tumor sizes. Outliers are easily noted.

therapeutic distinction between these groups. We agreed that a scatterplot view, color-mapped to the different therapies, would serve activities A1 and A3 better, by capturing more clearly individual patients and cohort patterns in the data.

We first organized the symptom ratings into a patient-symptom matrix for the selected time point, where each element (i, j) corresponds to the score given to symptom j by patient i at that time point. Previous research in HNC symptom clustering [74] had applied hierarchical clustering using Ward's method [98] with Euclidean distance on the patient-symptom matrix to group patients based on their raw symptom ratings. After alternative clustering with complete and average linkages, we found that Ward's method generated more informative groups of high symptom patients, which made sense to the clinicians. We identified two patient groups with high and low symptom burden (T1.1). This two-group clustering was preferred by clinicians, who found it easier to compare two groups instead of more. The axes of the scatterplot correspond to the first two components obtained by applying PCA to the patient-symptom matrix. Clusters for a specific time point are extracted and displayed, while clusters for different time points can be investigated via the time slider, which will update the scatterplot.

The scatterplot was customized to separately capture acute and late symptom burden distribution as identified by the symptom clusters, and to reflect via marker color, shape, and size the therapeutic combination administered to each patient, their gender, and their disease stage (T3.2) (Fig. 3.2). The data can be filtered by attributes, and filtering operations update the other views. A filtering control panel serves double duty, providing the plot legend. This customized scatterplot encoding effectively captured the distribution of symptoms in the patient population, patient outliers, and the therapeutic distribution in the data (T1.1, T1.3, T2.3).

To assess the impact of symptoms on clustering, we also provide an option to dynamically recalculate the clusters based on user-selected subsets of symptoms (Fig. 3.2) and update the scatterplot accordingly.

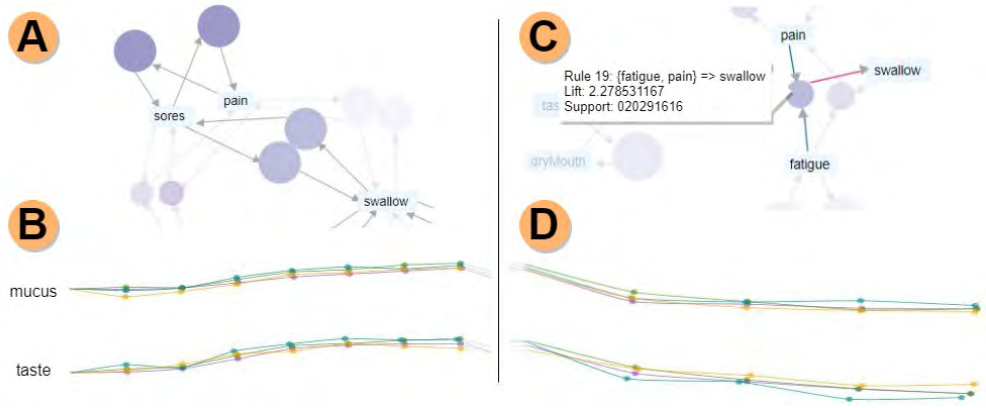

Figure 3.3: Acute vs. late phase analysis. A) Association rule diagram for the acute phase. Rules are filtered based on support (frequency) and lift (dependency between symptoms); other rules are faded in the background. B) Mean rating value filament plots for all therapies, with the acute phase highlighted. All therapies follow similar trajectories for both mucus and taste, and towards the end of the acute phase, taste has a considerable increase in ratings for all therapies. C) Association rule diagram for the late phase, showing the antecedents (fatigue, pain) and consequents (swallow) for rule 14.

D) Mean rating value filament plots, showing a slightly different trajectory for IC+Radiation.

Association Rule Diagram View. Driven by the limitations of the factor analysis discussed above, this project pursued the Association Rule Mining (ARM) as an alternative and novel approach to symptom cluster analysis (A1). ARM is an unsupervised data mining

technique for identifying relationships within the data [5]. In marketing applications, an association rule in the form $X \rightarrow Y$ indicates the pattern that if a customer purchases X, they will also buy Y, where the patterns are extracted from relational data expressed as transactions. Similarly to the strong positive correlations found between items in a supermarket basket, relationships within clinical data can help identify disease comorbidities [89,106,110].

Table 3.1: Example of 3 transactions containing four symptoms: fatigue, drowsiness, pain, and swallow.

tid	items
001	fatigue, drowsiness
002	pain, drowsiness
003	fatigue, pain, swallow

In this project, we extended the potential of ARM to symptom clustering applications. To this end, we adapted the most common ARM method to our problem: the Apriori algorithm [5], for frequent item set mining and association rule learning. In our approach, the symptoms experienced at each time point by each patient are treated as a transaction. The algorithm first identifies frequent symptoms to determine sets of symptoms that co-occur with high certainty and then extends to larger symptom sets (n >20). Table 3.1 contains an example of three "transactions" from our data. Transactions were extracted from existing questionnaires. The lack of ratings for a symptom in a questionnaire implied that the symptom was not included in the transaction. If a patient was missing an entire questionnaire, no transaction was generated for that patient. The ARM was performed using all available data, and no data imputation was performed.

We followed Agrawal and Srikant's proposed association rule [5] in the form:

$$X \rightarrow Y$$

which indicates that if a patient suffers from symptom X (the antecedent), they will also be affected by symptom Y (the consequent). Based on the first transaction in Table 3.1, such a rule can be:

$$\{\text{fatigue}\} \rightarrow \{\text{drowsiness}\}$$

where {fatigue} is the rule antecedent and {drowsiness} is the consequent. For itemsets larger

than this pairwise example (e.g., the last transaction in Table 1), either the antecedent or the consequent could contain multiple items.

Two standard measures, *support* and *lift*, are tuned to filter the association rules by a minimum value. *Support* is the measure of how often the transactions contain both X and Y, in our case, how frequently sets of symptoms X and Y occur together. The *support* of a subset of symptoms S is defined by:

$$\sigma(S) = \frac{|S|}{|T|}$$

where $|S|$ is the number of transactions that contain all the symptoms in set S and $|T|$ is the total number of transactions in the dataset. In Table 3.1, $\sigma(\{\text{fatigue}, \text{drowsiness}\}) = \frac{1}{3}$ as both symptoms appear together in 1 out of 3 transactions.

Lift is the measure of the importance, or strength of the rule, and it shows how more frequently than we would expect by random chance do X and Y appear together. *Lift* is defined as:

$$\lambda(X, Y) = \frac{\sigma(X \cup Y)}{\sigma(X) \times \sigma(Y)}$$

where $(X \cup Y)$ refers to transactions that contain both X and Y. For example:

$$\lambda(\{\text{fatigue}\}, \{\text{drowsiness}\}) = \frac{\sigma(\{\text{fatigue}, \text{drowsiness}\})}{\sigma(\{\text{fatigue}\}) \times \sigma(\{\text{drowsiness}\})}$$

We applied ARM to each of the acute stage and the late stage (T1.2, T3.4), and empirically chose to illustrate the top 20 rules yielded by this approach, because only a small number of rules were of clinical interest. We chose minimum values for the *support* and *lift* metrics that were suitable for frequent and interdependent symptoms.

Symptom Trajectory Panel

Designing an appropriate encoding for the symptom longitudinal data (A2) turned out to be particularly challenging, primarily due to the nature and richness of the temporal data,

the acknowledged variability in ratings between patients, and the missing or uneven time points, which were expected in this context. The design process explored a wide range of possible temporal encodings, many of which suffered from scalability issues, and, after several sessions, focused on a promising and novel encoding called a "tendril plot" [102]. A tendril plot is a visual summary of the incidence, significance, and temporal aspects of adverse events in clinical trials, in which individual temporal threads, one per each patient, emanate from a common root and shoot upward and curl either to the left or to the right depending on whether the next event in the timeline was adverse or an improvement. For clinical trial data, tendrils were shown to create beautiful, compact, naturally clustering pathlines that illustrate the positive or negative evolution of each group of patients. The clinicians had also seen this representation and thought it could work (T1.2, T2.2). Although promising on paper, unfortunately, the tendril implementation did not yield similarly clean illustrations for the symptom data, due to the much smaller number of time points, the variability in therapeutic sequences, and the variability in patient outcomes, which are not typical of clinical trials.

Numerous design variations yielded a new custom temporal encoding, which we call a filament plot (Fig. 3.4.D). Filament plots also emanate from a common root and then proceed in a left-to-right direction aligned with the time sequence. Wider timesteps, typical for late stage, are therefore more widely spaced. Each filament represents the whole observation period for a specific patient, with dots along the filament to indicate time stamps. To account for inter-patient rating variability, the curvature degree for the filament at each time step encodes the relative change from the previous rating, where upward rotation indicates worsening symptoms (rating increase). In contrast, downward rotation shows symptom amelioration (rating decrease).

To calculate the rotation, if patient p is located at position (x_t, y_t) at timestep t for a symptom with rating r, we compute the next position (x_{t+1}, y_{t+1}) at timestep $t+1$ by first

calculating the horizontal rotation angle as:

$$\theta = \frac{\theta_{max} \cdot \Delta r_{t+1}}{2 \cdot \Delta r_{max}}$$

where θmax is the total maximum rotation allowed, whose value is set to $\frac{3\pi}{4}$; Δr_{t+1} is the rating difference between $t+1$ and t:

$$\Delta r_{t+1} = r(t+1) - r(t);$$

and Δr_{max} is the maximum difference between two rating values, which is 10 in our case. Negative differences between ratings (i.e., rating decreases) produce negative angle values for θ.

Next, we want to rotate θ degrees relative to the horizontal line P_1P_2 defined by the points $P_1 = (x_t, y_t)$ and $P_2 = (x_t + l, y_t)$ where l quantifies the time elapsed between $t+1$ and t. A higher l indicates that more time passed between $t+1$ and t (that is, late vs. acute). Finally, we rotate P_2 around P_1 by θ degrees.

For missing data during the observation period, the associated points are not represented, and we consider no rating change from the previous time points; the surveillance period is described in each filament until the last recorded time point for each patient. We account for the time ratio between the acute (1 week) and late (months) stages, so the distances illustrated for the acute time points are smaller than those for the late time points. Hovering over a filament grays out all the other filaments in the plot. This interaction helps to compare the symptom trajectories for the same patient and, by brushing and linking with the different views, to highlight the additional patient data (T3.1).

This compact representation helps in the analysis of symptom evolution trends by clearly indicating the overall symptom burden (low/high). The representation also helps identify outlier trajectories that should be further evaluated and facilitates the discovery of steady vs. variable progression of symptoms. The panel includes two such filament plots, which support the side-by-side comparison of different symptoms for selected patient groups. To further enhance visual support, during the evaluation of the acute period in the entire THALIS

environment, the acute time periods are highlighted in the filament plots, and vice versa for the late period (Fig. 3.3.B, D).

To better support activities A1 and A2, an additional option uses the same filament encoding, this time with the color assigned to the therapy type, to capture the mean trajectory for each therapeutic combination (Fig. 3.1.B). Since in the therapy case, the mean symptom ratings across the population have meaning, the filaments are spread out according to the mean ratings per therapy (T2.1). This therapy analysis option helps estimate what treatment plans are less symptomatic, or, in contrast, lead to high symptom burden. In addition, to satisfy activity A3, the current patient's filament is highlighted in black in each plot (Fig. 3.4.D). Whereas reliable automated symptom prediction is an unsolved problem in symptom research, THALIS supports human-machine analysis via trajectory views of similar patients.

Cohort Symptom Panel

The last panel explicitly supports activities A1 and A3, and provides an abstract summary of the entire temporal symptom data. As in other fields [122], and as indicated by our activity analysis, this summary provides context for a specific datapoint, but does not lead the investigation. The panel comprises a percentile heatmap, a correlation matrix, and an anatomical sketch (Fig. 3.1.D, E, C).

The percentile heatmap (Fig. 3.1.D) is a custom representation showing the rating distribution of individual symptoms over time, for the entire patient cohort (T2.3). We arrived at this representation after exploring a variety of alternatives, such as stacked line plots, parallel coordinates plots, and radar charts, guided by feedback from collaborators. We settled on a matrix-based layout due to its compactness and its ability to support small multiple plots. Each row corresponds to a symptom, with rows grouped by symptom category, and each column corresponds to a time point. Each cell in this matrix is a horizontal bar graph showing through the shade the percentage of patients reporting within a specific range (0,

1-5, 6-9, or 10) for that symptom, at that time point. The height of the bar maps the percentage of individuals in the entire cohort who reported the symptom ratings at that time point. The current patient is indicated in this heatmap by cross markers (Fig. 3.4.C) (T3.1). This encoding proved to be an intuitive way to show what symptoms produce a higher burden on patients, and when, as well as to indicate how many patients were affected by these symptoms from the entire cohort (T1.2, T2.3).

To support exploration driven by a specific patient (A3), a dropdown selection box is also provided (Fig. 3.4.B). A selection in this box highlights the patient data across panels (Fig. 3.4). A timeline selector also allows for the choice of a particular time point in the data(Fig. 3.4.C), and further interface elements will enable the selection and analysis of sets of similar patients. Additionally, a compact correlation matrix (Fig. 3.1.E), along with the percentile heatmap, supports T1.2, showing the strength of the correlation between a selected symptom and all other symptoms, with circles encoding the Spearman coefficient via color and size. Finally, because a discussion of task T3.3 revealed that patients tend to point to the location of their symptoms, an anatomical sketch (Fig. 3.1.C) supports visual anchoring based on anatomy. The regions of the head and neck affected by the selected symptoms are highlighted in this sketch.

3.4 Evaluation

Because no design approach is failproof, although ACD has higher success rates than HCD [129], we evaluated THALIS through a combination of multiple demonstrations and case studies involving domain experts, namely a senior data mining specialist and three senior clinical radiation oncology experts. Two case studies were completed during separate, dedicated sessions, in addition to regular feedback sessions. As the designers and evaluators were in different locations, and due to COVID-19 constraints, these sessions were conducted remotely using screen sharing and note-taking. The oncology experts directed the exploration using the think-aloud method, while the first author drove the interface according to their instruc-

tions. Both case studies analyze a set of 699 HNC patients, which was significantly larger than prior clinician analyses, and span all activities, A1-A3. Qualitative feedback was also provided during weekly design-driven sessions and was used to improve the overall design of THALIS.

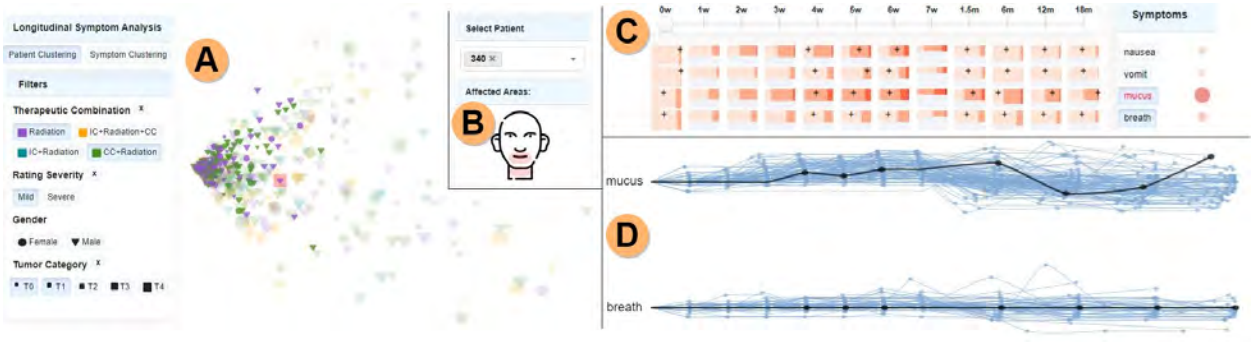

Figure 3.4: Symptom burden analysis. A) Patients in the mild symptom burden cluster, having tumor categories T0 and T1 (current patient, 340, is highlighted in red), with all other patients faded. B) The anatomical sketch shows that the mouth and neck areas are affected by the selected symptoms (mucus, breath) for the current patient. C) The patient's ratings are shown by black marks. In this case, the patient had a low rating for mucus at the first assessment (0 weeks), while at the end of the observation period (18 months post-treatment), the rating increased. D) Filament plots encoding symptom trajectories for the selected symptoms, for the patients filtered in the scatterplot. One filament per patient shows the temporal development for that symptom; black filaments mark the current patient, confirming the mucus rating increase in the late stage.

3.4.1 Case Study I: Symptom Burden Analysis in Radiotherapy

The study sought to assess the impact of therapy on symptom burden in this set and took place before we developed the associative rule model. Oncologists originally hoped to replicate the published analysis results obtained in significantly smaller cohorts of 80 to 270 patients [56, 101, 166]. Using the system over the course of several sessions showed, however, that those clustering results were not generalizable to the larger cohort (n >700). So the investigation shifted focus to discovering and analyzing outliers in terms of patient characteristics and symptom trajectories. The study workflow started directly with the therapy scatterplot panel (Fig. 3.4.A) (T1.1, T3.2). At first glance, most of the patients were visibly grouped in the center-right part of the plot, suggesting a substantial similarity. Filtering patients (T3.1) based on their rating severity revealed that this group corresponded to a mild-rating severity cluster. Further filtering by therapy and tumor category, the experts noted that most of these patients were treated with radiation with or without concurrent

chemotherapy (CC) and, not surprisingly, presented a small tumor size and a low symptom burden at the end of the observation period. They concluded that for this group, the therapy plan did not have a practical impact on quality of life. Next, the oncologists examined whether a smaller set of symptoms, as in their previous studies, would correlate with patient groupings (T1.1, T1.3). To do this, they filtered the data by daily interference symptoms, including, for example, mood, enjoyment, and work (Fig. 3.2). This time, they found that almost a third of the patients suffered from high symptom burden in this symptom group.

Encouraged by this finding, the analysis quickly moved to the filament plots (Fig. 3.4.D), to examine the symptom trajectories (T2.2). The plots captured a general trend in most symptom trajectories, namely, a rating decrease post-treatment, except for numbness, memory, breath. In addition, these three symptoms, along with nausea and vomiting, exhibited a steady symptom development, with fewer patient outliers or drastic rating changes over time (T1.2). In fact, there was no correlation between the temporal outliers in the filament plots and the therapy scatterplot outliers. This finding indicated that patients experienced steady ratings for these five symptoms over time, regardless of overall symptom burden or therapy treatment. This observation was of notable interest, so the analysis moved on to examine the cohort context (T2.3). Using the percentile heatmap (Fig. 3.1.D) and the correlation matrix, our collaborators observed that groups of symptoms such as swallow and dry mouth, or taste, appetite, constipation, and sores showed higher ratings over time, suggesting possible interrelationship or causal factors between these symptoms. For example, when selecting dry mouth, the panel indicated strong correlations between dry mouth and mucus, choking, and swallow, but also with taste, drowsiness, and fatigue as well. Finally, the anatomical sketch layout (Fig. 3.1.C) emphasized which head and neck locations are affected by the selected symptoms (T3.3). In this case, we observed that both dry mouth and taste affected the area of the mouth. The oncologists are planning studies to verify this set of symptom cluster hypotheses.

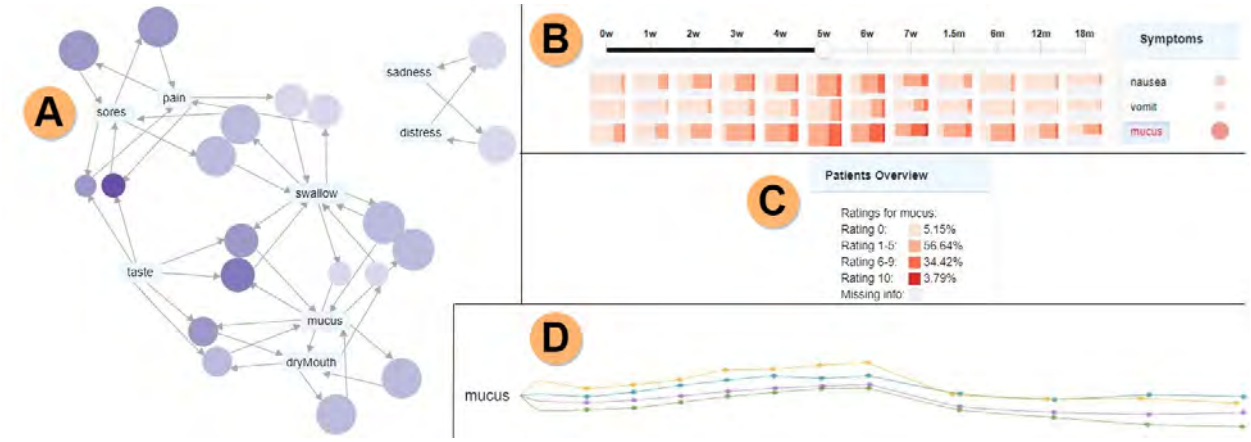

Figure 3.5: Symptom cluster diversity analysis. A) Symptom association graph for the acute phase showing mucus and swallow correlate with many symptoms. Note that the network layout is fixed, and that by construction it places centrally nodes with high degree. B) The percentile heatmap shows a spread of high ratings for mucus along the whole observation period. C) Summary panel for mucus showing that among patients who reported ratings for week 5 during treatment, more than 95% noted mucus as a present symptom. D) Mean rating filament plot emphasizing rising ratings at the end of the acute phase, especially for the IC+Radiation+CC treatment.

3.4.2 Case Study II: Symptom Cluster Diversity

This study aimed primarily to explore the value of associative rule mining in longitudinal symptom analysis (T1.2). Examining the association diagrams, the oncologists were stunned to find surprising symptom clusters during and post-treatment; in particular, eight common symptoms for the acute stage (Fig. 3.5.A), with two strongly coupled subgroups: distress, sadness, and swallow, pain, sores, taste, mucus; and, respectively, 12 frequent symptoms during the late part of the treatment (Fig. 3.1.A), showing symptom clusters such as taste, dry mouth, and sores, pain. The experts were impressed to see that the sores, pain cluster is strongly associated with taste in the acute phase, while in the late phase, there is a connection between drowsiness, sleep (Fig. 3.6.A), which is known to be a factor in dangerous muscle-mass loss. The taste, dry mouth cluster in the late phase supported our collaborators' previous findings. However, the connection between fatigue, drowsiness in the late phase and the centrality of mucus (Fig. 3.5.A), as well as the taste, sores connection within the acute graph, was unexpected. *"In our group, we have established this arc from taste to dry mouth in the late stage, but we haven't thought of the taste to sores link in the acute phase. That is striking."*

The ability to highlight a particular symptom (T3.4) or rule and filter rules based on their

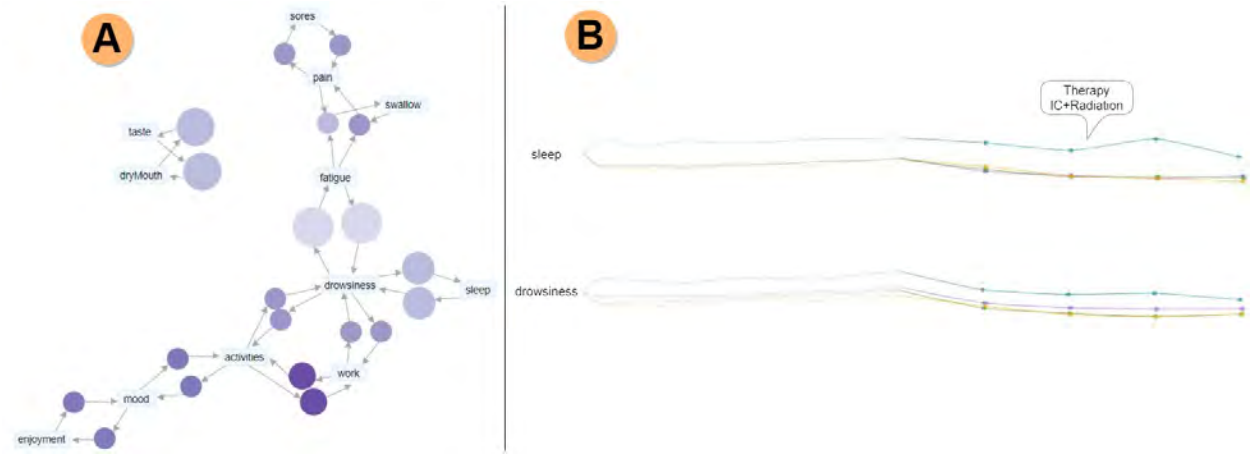

Figure 3.6: Late symptom cluster analysis. A) Symptom association graph showing drowsiness as a central symptom for the late phase. The connection between sleep and drowsiness is expected, as these two symptoms are known to be a factor in dangerous muscle-mass loss. B) The filament plots show mean rating values, with the late phase highlighted, and the acute phase faded. Notably, in the case of sleep and drowsiness, IC+Radiation is the therapy associated with higher symptom ratings, and it is noticeably different from the other therapy plans.

support and lift (Fig. 3.3.A, C) was found to be essential during the exploration, helping us to determine which symptoms were more persistent or more dependent on each other. For instance, fatigue, drowsiness were the most common symptoms (based on their support) and activities, work the most dependent on each other (based on their lift) in the late phase (Fig. 3.1.A). The insights observed from the symptoms association graphs were further extended using the percentile heatmap (Fig. 3.1.D), revealing the spread of high ratings for taste and fatigue over the whole patient supervision period (T2.3). Furthermore, because mucus was usually perceived as an acute symptom, the experts found it remarkable that a large number of patients experienced mucus during the late period (Fig. 3.6). The mean value filament plots were used to show the mean ratings per time point for each therapy while highlighting the treatment phase of interest (acute/late) (Fig. 3.3.B, D) (T2.1). The plots showed that the trends were remarkably conserved over time between therapies, although their magnitudes could differ. To achieve a better understanding, the option of separate filaments according to the starting mean rating (baseline) was used (Fig. 3.1.B, Fig. 3.5.D), which showed a difference in the symptom burden between therapies for the ARM-identified symptom groups. For example, in the case of taste and mucus, in both acute and late phases, the highest rated treatments were IC+Radiation+CC (induced chemotherapy, radiation, and

concurrent chemotherapy) and IC+Radiation. In contrast, CC+Radiation and Radiation alone were rated lower compared to the other two treatments. Notably, in the case of drowsiness, sleep, IC+Radiation was remarkably separated from the other treatment plans (Fig. 3.6.B). The oncologists concluded this case study, and the associative approach was a gold mine for their symptom research, highlighting the diversity of symptom clusters over time.

3.4.3 Expert Feedback

THALIS received excellent feedback from the oncology team, often indicating a change in thinking about their work. Below is some sample feedback in relation to our activity analysis A1 - A3:

(A1, A3) Quote from the most senior oncologist: *"I gotta be honest, every time I meet with you guys and we see these visualizations, I get so much material for future research. In general, to be fair, my focus in clinical practice [and in helping patients] tends to be on dry mouth and swallowing. I say – we're going to talk about dry mouth and swallowing, cause these two are really bad– and then there's all the other stuff. And then I see this [the ARM and heatmap and filaments], and here's this other stuff, that is usually at my periphery, but I don't focus on, although patients do mention it. If I were sitting with a patient and I'd look at this interface and ARMs—I get it, hey, there's actually a LOT of moving parts here [beyond dry mouth and swallowing], and they're related, and they have different time sources. It's sobering."*

(A1, A2) Both case studies had the team exclaim, on multiple occasions, about being *"blown away"*, *"that [symptom] spread over time just jumps out at you"*, *"This entire ARM approach is so different from [the approach we've followed in our previous research on symptom clusters]. I want to stick a flag in the ground with the ARM work, and look at dose to organs and use ARM to see dose-to-swallowing correlation, based on this spatial structure underneath"*, *"This interface and the ARM provide great preliminary data for so many grants [projects] right off the bat!"*, *"Really impressed"*, and *"This [relationship] is not intuitive, so*

it's very interesting. And I wouldn't have thought about it. But now, it makes perfect sense. Duh!", *"The [filament view] is such a great asset for the interface."*

(A3) The oncologists: *"[THALIS's] ability to go from patient to population is fantastic, I really love it, it's exactly what I need"*, *"I like that when a patient is with [oncologist], they want percentages, e.g., 66% of patients have normal appetite after 12 months, and [THALIS] shows that"*, *"When I see a patient, this [taste-dry mouth] association in the late phase is the default picture I have in my mind. But here I see that also fatigue connects to drowsiness, and that these symptoms show up in the acute phase as well, and that I really need to discuss these issues with my patients." "I can share [this view] with my patients, to explain that pain and swallowing and fatigue are really tightly related—we don't know if it's causation, but they definitely show up together, so could you please, please, take your pain and anti-inflammatory meds, and could you please do the swallowing exercises we've talked about?"*

3.5 Discussion

The case studies and the domain expert feedback demonstrate THALIS's value in bridging the gap between machine and human analysis, and its ability to help generate novel insights. Our integrated approach is capable of capturing longitudinal differences between acute and late stages while detecting outliers and trends in the symptom and therapy data. More importantly, our approach supports individual patient analysis while handling a large cohort (n >700) both computationally and visually. Through an ACD approach, and as indicated by expert feedback, THALIS successfully serves the core interests of its audience. In conjunction with the clustering panel, the symptom association rule view, the filament plots, and the cohort symptom panel enabled discovering interesting relationships in the data, and in several cases led to unexpected but insightful results. Furthermore, THALIS couples multiple customized novel visual encodings with symptom clustering algorithms in the background, enabling the domain experts to explore various scenarios and test their hypotheses in real-time. Its use of a multi-view paradigm supports flexible analytical workflows

that enhance computational power and human expert knowledge.

Through close collaboration with domain experts, our solution introduces compact, customized visual encodings for the symptom data: a filament encoding and a percentile heatmap. The percentile heatmap scales well with the number of subjects, by design, at the cost of summarization. Although the inherent scalability of filaments with the number of items shown is limited, these encodings successfully abstract the cohort data with the help of similarity-based filtering operations, which are appropriate in this context; for hundreds of dense observations, as common in other problems, tendrils [102] offer a better solution. In terms of scalability, the ARM graph can provide rules for any number of time points in the late and acute time periods. However, the graph representation for association rules is suitable for a smaller number of rules (less than 100 [83]). The scatterplot and correlation matrix are time point specific, so any number of plots could be generated. On the other hand, some views are prone to clutter. Some of these encodings may have limited generalizability beyond this application domain. In the case of filaments, they work in this application because there is a significant correlation between similar patients' trajectories and because our application emphasizes relative trajectory changes as opposed to absolute values. This type of correlation and relativity may not be true across application domains. However, our custom encodings can be repurposed for other longitudinal problems that feature missing data, such as in astronomy or biology [84, 124, 125, 171]. Future work includes longitudinal clustering of patients and symptoms, applying the ARM approach to sequential data, and interactively changing the metrics of the ARM and the number of rules.

3.5.1 Research Questions

Q1. *How can visualization support cohort analysis?* Through the ACD method, we interviewed both clinicians and data scientists in head and neck cancer research, which helped to understand the needs of our clients and the primary user activities for cancer treatment research. In this project, data visualization combined cohort data extracted from different sources and with various types of attributes, some collected over time, providing a common

ground for clinicians and data modelers to collaborate and generate hypotheses.

Q2. *How to visually represent cohorts and their characteristics, and what interactions to support?* We proposed a multiple coordinated view design to combine temporal symptom measurements with clinical variables within the cohort. We supported the selection of a desired patient to compare it to its cohort to better relate a future patient to existing ones. We emphasized the two-stage patient monitoring process with custom encodings, which highlighted the connection between treatment stages – how, during treatment, data impacts post-treatment longitudinal outcomes. For this purpose, we proposed several encodings, notably the filament plot, which showed the cohort overview of temporal symptom measurements and enabled inter-symptom comparisons.

Q3. *What system implementations work for post-treatment decision-making?* This project involved modeling symptom data to better associate patient studies during treatment with post-treatment outcomes. THALIS supported human-machine workflows by visualizing rule-mining-modeled symptoms, which were clinically validated and evaluated by clinicians. We presented incremental designs to our clients throughout the system prototyping phase, which helped with an incremental evaluation of the modeled results. This eventually helped build trust in the results and in the upcoming new cohort modeling research avenues.

Q4. *What makes a visual analytics system valuable to biomedical users?* THALIS supported clinician-modeler collaborations by assisting clinicians' evaluation of rule mining modeling results. The rule-mining results were explained in conjunction with relevant clinical patient attributes (visualization of different data facets for model understanding). THALIS demonstrated the importance of relating model outcomes to other sources of data to make these outcomes actionable in clinical practice.

Takeaways. The main lesson learned from THALIS was the importance of documenting the domain characterization when designing an application-specific visual analytics system. This is related to the importance of supporting clinician-data modeler collaborations when designing visualizations for cohort XAI and using data visualization to emphasize the domain

sense and actionability of the modeling results. It was relatively easy to gain trust in our results because the modeling methodologies, such as association rule mining, were more transparent (unsupervised modeling) than supervised, black-box models. Another notable finding was that the ACD methodology helped with the incremental, more reliable design of the front-end, by gaining trust in the visualizations due to the incremental use of said visualizations during regular update meetings, and by introducing unconventional visual encodings. The need for novel encodings seemed to be a consequence of unusual cohort characteristics, in this case, the two-stage temporal data. The two-stage patient supervision protocol was a key consideration in the front-end design to highlight how health trajectories during treatment influence post-treatment outcomes.

Considerations for future work (Chapter 4) include: 1) to expand on the association rule mining approach to find temporal symptom associations/clusters between the acute and late treatment stages, and as a result, 2) to delve into visual analytics for longitudinal risk modeling using rule mining, 3) to model treatment-induced risk as opposed to THALIS that extracted rules from the entire cohort data without considering the influence of treatment on symptom risk, 4) to find designs that can visualize more rule results, as opposed to the limited number presented by THALIS (e.g. 20 per stage), and 5) to explore designs that take into consideration the difference in analytical tasks between clinicians and data modelers.

3.6 Conclusion

THALIS presented an example of a visual analytics system that aims to support cohort analysis and modeling for multivariate temporal patient data. This work introduced domain characterization for outcome modeling with symptom measurements in head and neck cancer, which focuses on a two-stage patient monitoring protocol, i.e., during and post-treatment implementation. The presented system focused on custom visual encodings to incorporate rule mining modeling results in conjunction with multivariate longitudinal patient attributes, with the aim of stratifying patient cohorts by outcome. In particular, the

proposed visualizations helped to summarize cohort characteristics and make the results of the association rule actionable for risk analysis in clinical practice. Both clinicians and data modelers evaluated this work.

In the following chapter, I will present an extension of this domain application, this time using visual analytics for outcome risk prediction and risk stratification through more configurable analytical workflows, which consider the differences in the mental models of the clients (i.e., data modelers vs. clinicians).

Chapter 4

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

4.1 Introduction

This chapter presents the design, development, and evaluation of a visual analytics system that tackles configurable workflows in multidisciplinary collaborations through a highly customizable front-end. The proposed system provides flexibility to both data modelers and clinicians to evaluate cohort modeling results, considering both clinical interpretation and prediction results evaluation. Unlike previous work, this project focuses more on determining longitudinal risk after treatment completion. In this project, the starting premise is that the treatment type and symptoms during treatment can predict symptom risk post-treatment. The proposed system uses custom visual encodings to explain how different categories of outcome risk occur for other treatment plans, comparing these risks across the cohort stratified by treatment type. This work expands the previous risk modeling approach and introduces a modeling method for longitudinal treatment outcome risk, using sequential rule mining and rule clustering. The system uses visual analytics to interpret and evaluate rule mining predictions and introduces a custom visual encoding to summarize multi-stage, temporal attributes (symptoms), namely the rose glyph. Additionally, a rose glyph projection summarizes multi-stage networks with temporal nodes (symptoms), which are the clustering result of large-scale (n >20), overlapping sequential rule sets. The rose glyph supports various tasks, such as explaining and comparing risk results for different sets of patients alongside relevant patient attributes. Both modelers and oncologists evaluated the system on a cohort of 766 head and neck cancer patients.

The contents of this chapter were presented at IEEE VIS 2023 in the full paper track, under the area of Applications [61].

4.2 Motivation

In recent decades, 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"*), for many patients, treatment leads to side effects that can significantly affect quality of life even after 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, in stratifying patients into high-risk and low-risk cohorts, and in studying the relationship between symptoms and treatment decisions, with an effort to identify long-term symptoms that affect the patient's quality of life.

In HNC, identifying the risk of symptoms is particularly challenging due to the effects of specific treatments and various clinical factors [154]. Furthermore, the temporal nature of the progression of symptoms during and post-treatment requires special consideration when making predictions for a given patient. In contrast, some symptoms are often correlated with other symptoms, either due to direct influence or by shared root causes. These factors make predicting treatment outcomes difficult and hamper the decision-making and delivery of personalized care. Because these challenges complicate the interpretation of treatment outcomes, there is a need for alternative human-machine analysis tools that can enhance computational and human effort to help modelers better understand HNC symptoms.

Current computational symptom research is focused on symptom clustering [81, 104]; however, there is little work [60] that correlates symptom patterns over time or compares the results of different treatment modalities. Sequential rule mining (SRM) is a promising unsupervised learning approach for discovering common temporal patterns in symptom data. Still, 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. However, the model results also have to make sense in a clinical setting, and so they need to be interpreted by domain experts with real patient data. Beyond helping modelers, visual analysis can help further with model interpretation in the context of clinical patient data.

Visual computing with temporal symptoms has several challenges. First, the large size of the patient cohort (>700 patients), number of symptoms (>20 symptoms), and time points (>10 time points) requires scalable encodings that are readable by users, 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 for flexible comparison between symptom patterns for sub-cohorts. Fourth, since the interpretation of causal structures requires both data mining and clinical expertise, the systems need to allow for multiple workflows and levels of detail to analyze both symptoms and patient sub-cohorts. Finally, concluding high-dimensional cohort data requires the use of interpretable algorithms to help extract patterns that are both useful and simple enough to be understood by users, for which we propose combining rule mining and clustering to yield simple but flexible results.

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 between 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.

4.3 Design

4.3.1 Setting

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

Our design process followed an Activity-Centered Design (ACD) approach [129], focusing on user activities and workflows. This paradigm has shown higher success than traditional human-centered design for scientific, interdisciplinary collaborations. In this project, 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 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 [129].

4.3.2 Activity and Task Analysis

Our system serves model builders in the research of cancer symptoms. Our collaborators have experience in ML approaches for predicting patient outcomes and symptom analysis, but were interested in alternative methods for temporal symptom analysis that focus on exploring the differences between patients receiving different treatment modalities. There was also a need to efficiently present and interpret the results of the proposed model to

our clinician collaborators. In addition, it was imperative to compare the toxicities found by the model between different treatment groups to find symptoms that depended on the treatment modality. 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 visualization tasks:

A1. 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 symptom severity across treatments
- T2.4. Facilitate the comparison of clinical patient data for multiple treatments

Our evaluation describes examples of preferred workflows focused on these activities, while experts in the oncology domain clinically validate the results. Non-functional requirements included clarity in the model results, visual explanation, scalable visualizations that can display symptom and patient statistics, and intuitive visual abstractions that effectively guide the user during their data assessment.

4.3.3 Data

The data used to build the proposed work is from a cohort of 823 HNC patients who underwent treatment at the MD Anderson Cancer Center in Houston, TX. Demographic and

diagnostic information was recorded for this cohort, covering 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). This dataset is a more comprehensive dataset compared to the one used in Chapter 3

Data on self-reported longitudinal symptoms were extracted from patient questionnaires [40] at 12 time points: before starting treatment, weekly for 7 weeks during treatment, 6 weeks post-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 issues). The 12 time points were divided into two categories: the acute stage (once before the start date of treatment, or week 0, and all 7 weeks throughout the treatment) and the late stage (the remaining four post-treatment assessment dates). Not all features were available for all patients. 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.

4.3.4 SRM Modeling for Medical Data

Association Rule Mining (ARM) [5] 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 black-box models used in various applications, it has also caught the attention of medical research [10, 148, 176]. We applied ARM in our previous work [21, 60] 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 did not help us predict *late* symptoms based on symptoms during treatment.

An interesting extension of association mining for temporal data is sequential rule mining (SRM) [46]. 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:

$$R1 : \{taste, nausea\} \rightarrow \{dryMouth\} \tag{4.1}$$

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 a large number of rules (n >20) may typically be required to make knowledge actionable. Moreover, the 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 who have taste and nausea problems during treatment, followed by

dryMouth problems post-treatment:

$$sup(R1) = \frac{|\{(taste, nausea) \cup (dryMouth)\}|}{|S|} \tag{4.2}$$

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:

$$conf(R1) = \frac{sup(R1)}{sup(\{taste, nausea\})} \tag{4.3}$$

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:

$$lift(R1) = \frac{sup(R1)}{sup(\{taste, nausea\}) \times sup(\{dryMouth\})} \tag{4.4}$$

A lift value ≤ 1 indicates that the rule cannot predict the consequence with more accuracy than could be expected by chance.

As noted above, 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 to obtain valuable, easy-to-interpret, and meaningful information regarding the patterns within the data.

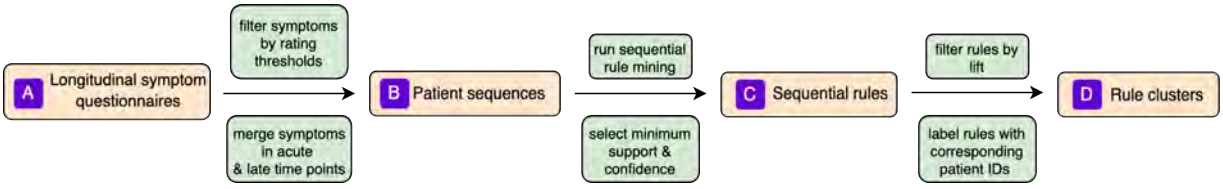

Figure 4.1: SRM Modeling. A) Patient-reported symptom ratings are recorded as longitudinal records. B) Records are processed into patient symptom sequences. C) Patient sequences are provided as input to the SRM algorithm. D) The sequential rules are filtered and clustered into rule clusters based on their corresponding patient IDs.

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 post-treatment (the *late* stage) (Figure 4.1.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 [63] to compute the sequential rules, which is an adaptation of the Deogun et al. algorithm [46] for multiple sequences of events. We followed the documentation of the open source data mining library called SPMF [64] 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 data preprocessing, we computed sequences from patient timelines (Figure 4.1.B). Each sequence corresponds to the temporal ratings of one patient across both the *acute* stage (baseline and during treatment) and the *late* stage (post-treatment). Consequently, we abstracted the sequences into two-stage patterns, *acute* and *late* (Sec. 1.3). In the *acute* pattern, we include a symptom only if the patient provided a rating above a given severity threshold (e.g., $\geq$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. $\geq$3). Clinically, a rating $\geq$5 is considered a moderate-to-high severity, while three is regarded as mild severity. The same threshold is not enforced 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 minimize patient variability and individual symptom severity ratings.

Next, the SRM algorithm was applied to these sequences to identify sequential rules (Figure 4.1.C). Similar 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 a 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 one 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 expected, the extracted rules within each treatment cohort showed many similarities in terms of symptom patterns (often differing in only one or two symptoms) and in the set of patients supporting the rules (more than 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 that support the rule. Next, we computed the similarity between the rules based on their common patient IDs using the Jaccard index [93]. 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 [45] 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 because we have found it produces highly interpretable results through the use of dendrograms [123], which allows us to manually adjust the clusters and identify outliers. We decided on the number of clusters after inspecting all treatment results. We created rule clusters (Fig. 4.1.D) by merging the antecedent symptoms and the consequent symptoms of 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).

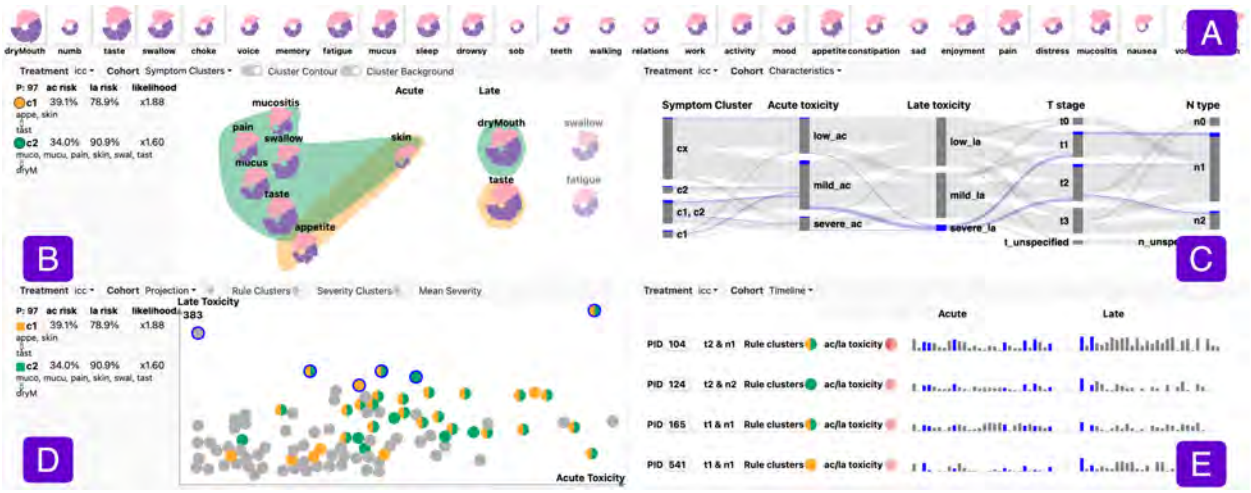

Figure 4.2: Longitudinal symptom analysis and prediction for head and neck patients (ICC treatment). A) Overall severity over time for each symptom, across treatments. B) Sequential mining component, showing two clusters that use *acute* symptoms (left) to predict *late* symptoms (right). Lower opacity 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 symptom burden.

4.3.5 Front-end Design

The proposed visual system was built using Python for the back-end and React with D3.js for the front-end. The design is based on coordinated multiple views that support diverse analysis workflows. The interface is split into five panels that support six types of visual components. The top panel (Fig. 4.2.A) is the only panel that cannot be configured by the modeler and shows the stratified overall symptom severity for the entire patient cohort. The rest of the interface is split into quadrants that can be configured with any of the following five visual components: the symptom clustering component (Fig. 4.2.B) - which

denotes temporal symptom clusters for one treatment; the patient clustering component (Fig. 4.2.D) - which shows patient cohort symptomatology attributes for one treatment; the cohort characteristics component (Fig. 4.2.C) - which correlated diagnostic data to symptom clusters and symptom overall severity over time; the cohort timeline component (Fig. 4.2.E) - which displays an in-depth view of each patient's longitudinal and diagnostic features; and the symptom query component (Fig. 4.7.D) - which provides overall statistics regarding the appearance of symptoms during (acute) and post-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 at their top-left to facilitate workflow configurations.

Figure 4.3: Rose glyph. Color-coded petals aggregate the mean severity for patients for the 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.

Our design relies on the use of custom Rose Glyphs (Fig. 4.3) to encode the trajectory of a single symptom severity within the entire cohort (Fig. 4.2.A), or subgroups in the data (Fig. 4.2.B). For the selected subgroup, the mean symptom rating at each time point is encoded using variable-radius slices (petals). Based on feedback from collaborators, the symptom trajectory starts with the baseline ratings at 9 o'clock. It progresses clockwise in order of increasing time points, showing rating details for each of the twelve time points of the patient observation period. Pink petals encode acute treatment time points, while purple petals encode late time points. Late time points are wider to denote that they represent longer intervals than the acute time points (e.g., months vs. weeks). We took inspiration from Florence Nightingale's Rose Diagram [24] for this glyph. Still, instead of focusing on comparing events within a time frame, we mainly concentrate on the temporal

trajectory and comparing trajectories across symptoms. We used the radial glyph design because it provided a compact way to display the severity of symptoms across cohorts, while highlighting which of them represent a greater burden on patients, and an efficient way to compare the progression of severity between symptoms.

Overall Symptom Severity

This component (Fig. 4.2.A) shows 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 list of symptoms starts with dryMouth, which is one of the most severe symptoms throughout the observation period, and is the most persistent symptom post-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 the relative frequency of symptom occurrence, as there can be a significant variation in self-reported symptoms between items that may not correlate with their impact on quality of life. Symptoms predicted by SRM in at least one of the existing treatments are highlighted with a shadowed border (i.e., taste). This encoding provides a compact way of showing the overall trajectory of symptoms for the entire cohort, and it serves as an entry point for further analysis, as well as giving a reference point when evaluating treatment-specific symptom patterns.

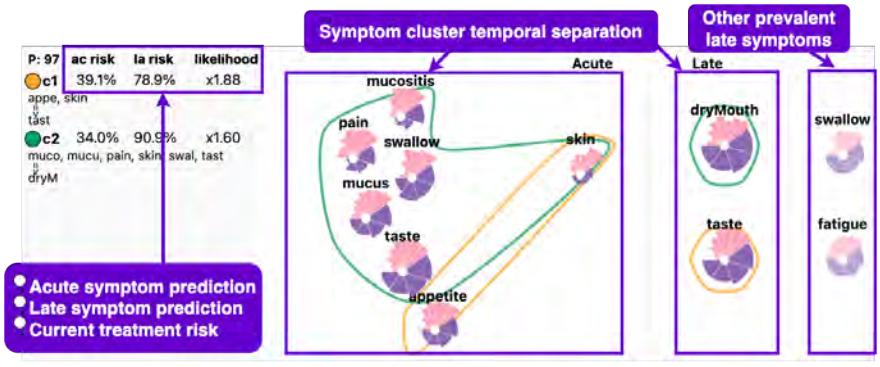

Figure 4.4: Symptom clustering for treatment ICC. The clusters in orange and green predict in the *late* stage dryMouth and taste problems. Cluster 1 (orange) shows a higher risk of developing these toxicities for ICC rather than the other existing treatment modalities (i.e., 1.88 times more likely)

Symptom Clustering

This component (Fig. 4.4) provides a visual abstraction for the symptom clusters found in Section 4.3.4 through a 2D projection of the corresponding symptoms using rose glyphs (T1.1, T2.1). The view is split into two halves to facilitate temporal separation between acute and late stages. The X and Y axes in the acute half correspond to the first two principal components after applying PCA to the Jaccard similarity between symptoms, based on the common patient IDs they share. Because many of these symptoms have an underlying association, we used PCA, as opposed to other projections, because it works better for correlated attributes. We use a force-directed layout to ensure that the symptom glyphs do not overlap in the projection. Symptoms are represented using rose glyphs to show the mean severity distribution over time among patients who correspond to clusters. This also improves temporal symptom severity comparisons between a selected treatment and the overall cohort or another treatment.

In the latter half (Fig. 4.4), 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. Furthermore, we list 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 with low opacity for the rose glyphs as opposed to predicted symptoms.

The left legend of the component shows the details for each symptom cluster (Fig. 4.4): 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 (Fig. 4.7.B, C) categorically color-coded. The envelopes' background can be turned on (Fig. 4.2.B), which can better emphasize the correspondence of symptoms 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 highlight the complementary 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 [60]. 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 (n >20). 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, which did not capture well the temporal nature of the rulesets. As a result, we detached from displaying actual rules. We 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 the interpretation of symptoms, to enhance the comparison of the trajectory between symptoms.

Cohort Symptom Query

This component (Fig. 4.7.D) 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 patients who have reported symptoms above the given thresholds at least once during and post-treatment for each symp-

tom. 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 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 to explain the prevalence of symptoms.

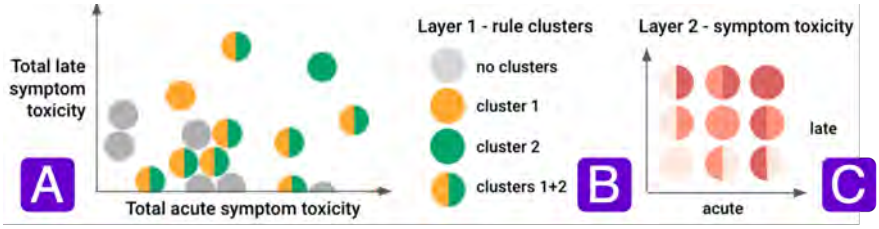

Figure 4.5: Patient Clustering View. A) Scatterplot showing patient glyphs. Two options for patient encodings in the scatterplot: B) encodes cluster memberships, and C) encodes temporal symptom burdens.

Patient Clustering

This component (Fig. 4.2.D) provides a custom 2D scatterplot projection of the patient cohort, with axes corresponding to the total severity scores of symptoms for acute time points (X axis) and the late time points (Y axis) (Fig. 4.5A). 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.

This component has two interchangeable layers: the first layer (Fig. 4.5B) 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 to which it belongs, where each section is colored to match the cluster colors from the symptom clustering component. The second layer (Fig. Roses 5.C) divides the points into two sections, representing, from left to right, the acute and late treatment periods, respectively. The color of each section is assigned to the overall symptom severity for its corresponding period, with lighter red encoding low severity and

dark red encoding high severity. Furthermore, this layer can be applied by selecting a subset of symptoms from the top rose glyph row (Fig. 4.2.A), and it will be updated to show the acute/late severities of the selected symptoms. Brushing operations are available on this view, which will highlight or filter information in the rest of the views based on patient selection (T1.4).

Alternative designs experimented with other projection methods and glyph encodings. However, we found most projection methods like PCA [60] 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 the collaborators. Using symptom severity made it easier for collaborators to identify patients with increases or decreases in treatment severity between the acute and late stages. For the scatterplot glyph design, we considered alternative shapes instead of circles for different clusters. Still, we found that it was challenging to capture an arbitrary number of cluster memberships across treatment modalities using shapes. 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 the cohort information to actual patients for the given treatment.

Cohort Timeline

This component (Fig. 4.2.E) functions as a detailed view of the attributes of each patient (T1.4), using timelines and small multiples to show the 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 (Fig. 4.8.C). The right half uses a barchart timeline, split by acute and late stages, showing mean ratings

for each of the 28 symptoms (Fig. 4.8.C). The symptom bars are ordered after the top interface row symptom order. They are colored blue when they represent symptoms that are present in at least one of the rule clusters for the selected treatment, or gray, vice versa. 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 the flows in the cohort characteristics component.

Oncology experts are often interested in analyzing a single patient and comparing them with 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 time points from the symptom data. This resulted in cluttered components that took most of the screen space due to the large number of time points (n >10), 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).

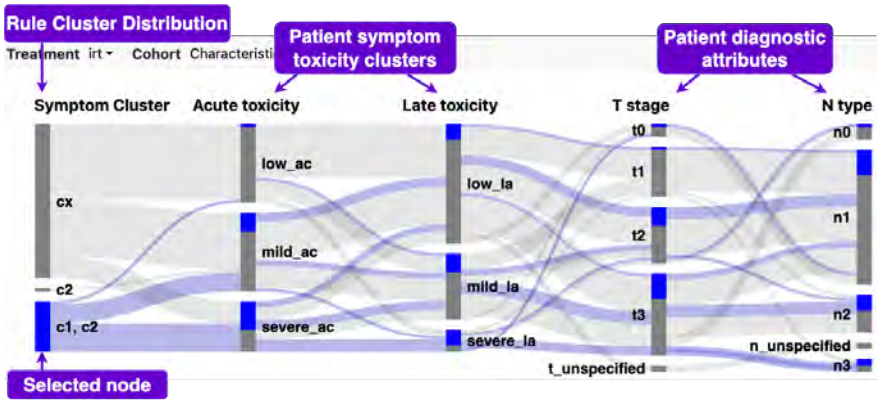

Figure 4.6: Sankey Diagram for IRT treatment. Node $c1, c2$ is selected, showing that a tiny part of the patient cohort with this cluster combination is linked to low symptom severity in the *acute* stage.

Cohort Characteristics

This component (F Fig. 4.6) 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). In addition to showcasing possible combinations of symptom groups, we stratify patients into low, medium, or high symptom burden for acute and late stages using K Means clustering on the total symptom toxicity scores for both stages. This further emphasizes how acute symptom burden is translated into late symptom burden. The nodes from the diagram can be selected, and the corresponding nodes and flows are highlighted in blue (F Fig. 4.6), while filtering options in the other views highlight in 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 categorical and temporal characteristics in previous work [195]. Our collaborators easily adopted this design and became a key component of their analyses. The diagram's ordinal axes are ordered from top to bottom (i.e., T/N stage, acute/late toxicity), in accordance with the request 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 across a treatment modality and how it is connected to the distributions of the other cohort attributes.

Flexible Workflow Support

Due to the variation of the requirements that would support the analysis at both the patient cohort level and 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 also sought common

symptoms between treatments that may develop independently of the treatment strategy. In addition, they were interested in predicting what a new patient should expect given a selected treatment to better assist future treatment decisions. To support these multiple workflows, the afferent components can be flexibly swapped.

4.4 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. Only the model builders were part of the entire design process for our interface and model building, while 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. 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 who presented five treatment modalities: RT, IRT, CC, ICC, and Surgery_and_other. We show in abbreviated form these case studies.

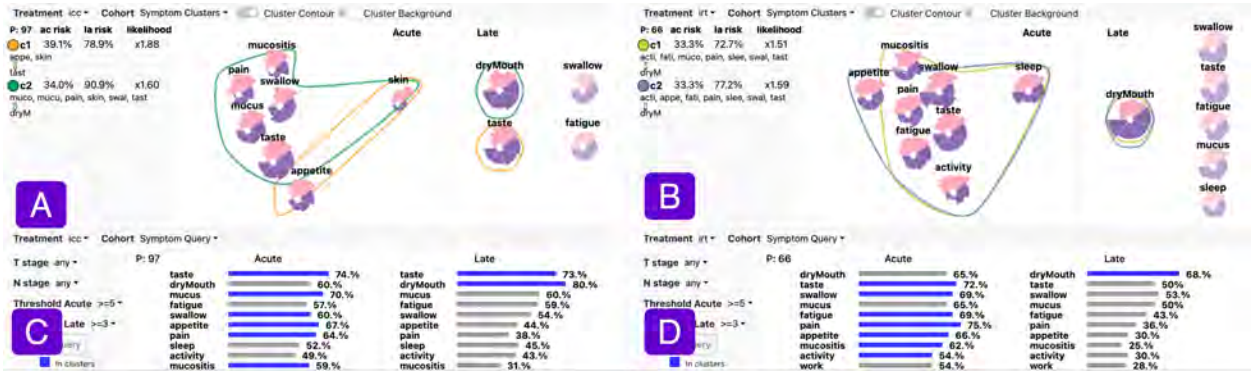

Figure 4.7: Treatment comparison. A) Overall cohort toxicities for all time points. B) and C) Symptom clusters for treatments ICC and IRT. Both treatments show two clusters, with similar *acute* symptoms, but ICC presents taste as a *late* symptom (B), as opposed to IRT(C). Although the rose glyphs are 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).

Case Study I: Multi-treatment Analysis

The model builders wanted to find temporal symptom patterns across multiple treatment modalities and compare the results. Oncologists hoped they would discover specific symptoms highly correlated with particular treatment strategies. After examining the top row of the interface (Fig. 4.7.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 a decrease post-treatment, as 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). The interface was then used to show the symptom clusters for ICC (Fig. 4.7.B) and IRT (Fig. 4.7.C) in conjunction with the symptom queries (Fig. 4.7.D, E). 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 identical overall symptom profiles (T2.1). Although the symptom queries showed many prevalent late toxicities (Fig. 4.7.C, E), 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 (Fig. 4.7.B, C). Taste was predicted as a late toxicity for ICC, correlated with loss of appetite, and, surprisingly, with skin problems (Fig. 4.7.B). DryMouth showed apparent severe toxicity in late when compared to the whole cohort (Fig. 4.7.A), more so for IRT (Fig. 4.7.C) (T2.1,3). The evaluators appreciated how the rose glyph projection kept symptoms with similar trajectories together. For example, in the ICC symptom clusters, pain and mucositis showed strikingly similar trajectories (Fig. 4.7.B). They hypothesized that this might be a sign that pain, being such a general symptom, was highly correlated with mucositis problems in this cohort. The evaluators also showed particular interest in the outliers of the acute projections, namely issues with sleep in IRT and skin in ICC.

Replacing the symptom query components with the Sankey diagrams for the two treatments (Fig. 4.2.C, F Fig. 4.6), the evaluators observed that IRT showed N3 stage (advanced) for node lymphs. At the same time, ICC did not present such a high attribute value (T2.4). In terms of the predicted values of the model, 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 that show an increased risk of developing these patterns within the given treatment (T2.2). For example, although only 97 patients were given ICC (Fig. 4.7.B), the model showed a higher likelihood (i.e., almost twice as 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.

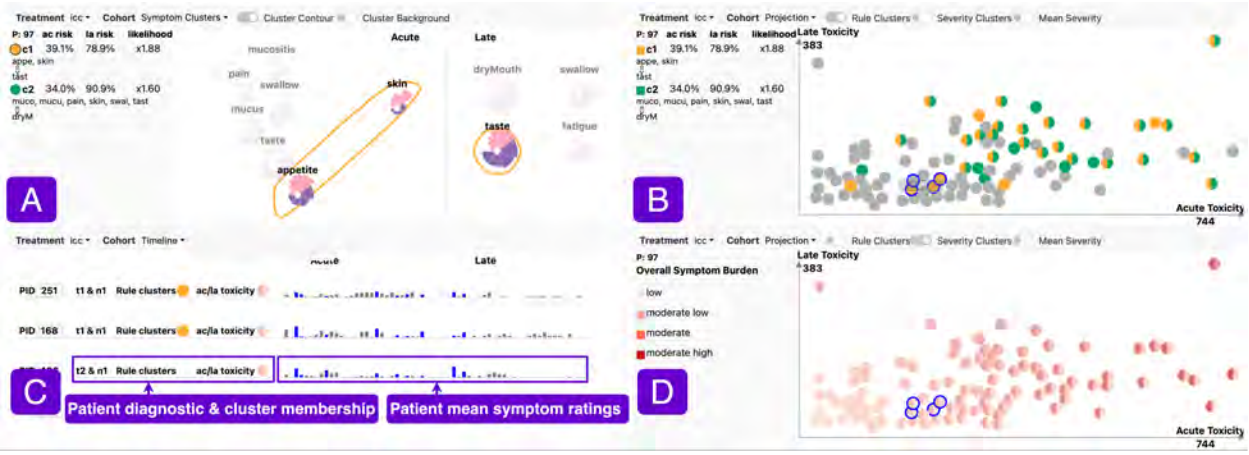

Figure 4.8: Single-treatment analysis. A) ICC treatment symptom clusters with cluster 1 (orange) selected. B) ICC patient projection with the cluster label layer. The cluster 1 outlier patients from the lower-left side are selected and highlighted in blue in the scatterplot filtered in the other views. 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*.

4.4.1 Case Study II: Single Treatment Analysis

For the second study, the cancer co-modelers wanted 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 (Fig. 4.8.A), patient projection

component using the symptom cluster layer (Fig. 4.8.B), the patient timeline component (Fig. 4.8.C), and the cohort characteristics component (Fig. 4.2.C). At first glance, the patients who usually suffered from moderate-to-high symptom burden overall showed patterns among the two existing symptom clusters (Fig. 4.8.A, B) (T1.1). Patients usually showed problems in both clusters, with lower burden patients sharing mostly cluster 1 (appetite, skin $\rightarrow$ taste) (Fig. 4.8.B). Selecting the previously mentioned sub-group of patients with cluster 1 from the scatterplot (T1.4), the evaluators looked at their timelines (Fig. 4.8.C). They 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 problems below the T3 stage (T1.3). While evaluating the cohort characteristics component, the oncology co-modelers commented that they expected severe symptom burden in late stages to be correlated with higher T stage (i.e., T3) (Fig. 4.8.C). Still, this view proved that it was not the case.

Next, the evaluators wished to understand the overall temporal toxicity among the patients. The cohort characteristics component was changed to show the patient projection with the overall temporal severity layer applied (Fig. 4.8.D) (T1.2). In this way, they could better understand the relationship between the 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.

4.4.2 Expert Feedback

The visual system and model results received extremely positive feedback. One of the senior model builders affirmed: *"The interface is extremely useful for navigating through the 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"*. One oncologist thought that the system would help explain late symptoms to patients and expressed interest in deploying the system to other clinicians to assess whether it can help them when dealing with patients.

In terms of visual encodings, 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". 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".* The diagram 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 helpful to observe symptom burden temporal trends at the cohort level while detecting outliers. The other components served as applicable complements to the model analysis.

4.5 Discussion

This work was developed as a collaborative project with 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 flexible and easily comprehensible explanation of common co-occurring symptoms and predicting late-stage symptoms for different treatment groups. Although 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.

Since 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 >700). 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. At the same time, the timelines can show any number of patients using scrolling operations. 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 homogeneous cohorts of symptom data (>700 patients with >20 attributes over >10 time points), we felt that it was more important to provide a highly configurable interface, supporting several workflows, at the cost of some scalability issues. The user can set any sub-cohort with any visual component in any part of the interface, which enhances the user's analysis process.

Notably, some of the patients used in the model building were still in the observation period and, as a result, they were missing symptom ratings for many post-treatment time points. This impacted the results of the model's predictions. Future work includes refinements in the SRM model using the interface once the data set is complete.

4.5.1 Research Questions

Q1. *How can visualization support cohort analysis?* Through the ACD method and considering the experience from THALIS, we interviewed data modelers in head and neck cancer research for task-gathering because this project aimed to experiment with new modeling approaches for patient cohorts. The modelers presented specific design considerations to better visualize the modeling results for the clinician collaborators. In this project, treatment-induced toxicity was one of the primary research considerations. As a result, we stratified the patient populations into cohorts by treatment plans and used data visualization to analyze in-depth a given cohort, as well as to compare different cohorts to understand how treatment type influences patient outcomes and post-treatment quality of life.

Q2. *How to visually represent cohorts and their characteristics, and what interactions to support?* We proposed a multiple coordinated view design by splitting the front-end into quadrants. The users could pick a visual encoding and a treatment type for each quadrant. This design principle enabled the configuration of different workflows, such as clinician

workflows, modeler workflows, in-depth treatment analysis, or cohort comparison. For the visualization of cohort characteristics, we faced some modeling and data challenges, such as the need to visualize the overlapping temporal rules that represented outcome predictions for each cohort. These results were abstracted into temporal networks with temporal nodes. We proposed the rose glyph encoding for several tasks within the system, such as providing temporal trajectories for all symptoms or grouping and comparing temporal rules.

Q3. *What system implementations work for post-treatment decision-making?* As opposed to THALIS, we focused on the different analytical needs for clinicians and data modelers. Roses supported the modeling activity by summarizing the rule mining results for modelers. At the same time, visualizations for different cohort data facets helped to link all the pieces together for clinicians (e.g., connecting the rule clusters to clinical statistics within a cohort, or to the overall symptom burden within a cohort).

Q4. *What makes a visual analytics system valuable to biomedical users?* We experimented with a highly configurable front-end to explore different workflows in clinicians and data modelers and to comply with differences in mental models of our collaborators. The evaluation sessions showed how Roses can elicit hypotheses about the toxicity of treatment-induced symptoms in different cohorts. Design considerations for clinical researchers ensured that the results of the rule mining modeling were accessible for clinical interpretation and actionable in practice. The configurable front-end supported more comprehensive analyses due to accounting for different levels of detail.

Takeaways. Unlike THALIS, where the domain characterization process was essential to understand how data visualization can help cancer post-treatment research, in Roses, we had the benefit of having all that information at hand. As a result, we focused on expanding modeling approaches from THALIS and exploring a highly configurable front-end that takes into account multiple levels of cohort details to better accommodate the differences in user workflows and different insights. The highly configurable interface 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 modeler was more interested in identifying the rules with the highest prediction metrics for each treatment, and thus benefited from a layout that was more focused on showing multiple different panels. In contrast, some clinicians were interested in assessing the value of the rules when explaining results to patients, and thus benefited more from configuring the layout to allow for side-by-side comparisons between panels. On the other hand, the amount of possible front-end configurations was overwhelming for some users at the beginning. Another finding in this project was that due to unconventional cohort characteristics, we had to come up with a novel encoding to interpret the cohort results. Besides the two-stage characteristic, we dealt with temporal networks that contained temporal nodes, which were the results of the sequential rule cluster modeling. Notably, this visual encoding was the key component of our interface, supporting several analytical tasks. We proposed the rose glyph to ensure actionability and transparency in the modeling results.

Considerations for future work (Chapter 5) include: 1) better separation for clinician vs. modeler activities and 2) better workflows structure since Roses's configurable front-end can become overwhelming due to too many layout options, 3) need for visual analytics support for more complex modeling activities (e.g. black-box model understanding and evaluation), and 4) visualize modeling results for user-selected cohort stratifications to better understand what are the cohort attributes associated with symptom risk (aka main risk components). Next, I will discuss the completed work and future steps for this research.

4.6 Conclusion

Roses introduced an example of a configurable visual analytics system for clinician-modeler collaborations in head and neck cancer cohort analysis, which accounted for different user interests and tasks, and consequently, different levels of cohort details. We presented the domain characterization for longitudinal outcome risk modeling after the completion of oncological treatment plans. The system supported model explanation, introduced a new method

to predict temporal risk post-treatment completion, and compared the modeling results to better understand how treatment plans influence adverse outcomes. Moreover, the system introduced the rose glyph to summarize multi-stage, temporal patient attributes (i.e., symptom measurements) and to better understand the risk stratification process.

In the following chapter, I will continue with the same application domain, but will not extend the work from THALIS and Roses. I will present an independent project that aims to explain and evaluate symptom risk predictions post-treatment, where I focus on data modeler activities and support clinician interpretation of the modeling results through separate front-ends, dedicated to the two types of domain experts.

Chapter 5

L-VISP: LSTM Visualization for Interpretable Symptom Prediction in Patient Cohorts

5.1 Introduction

This chapter presents the design, development, and evaluation of a visual analytics system, L-VISP, for analytical workflows centered on data modeler needs, and it supports cohort treatment risk modeling activity. L-VISP's design considers the evaluation of the results by clinicians as well, who play a secondary user role in this project. I continue my work on the head and neck cancer domain to assist the evaluation, interpretation, and actionability of supervised and unsupervised risk modeling. Specifically, this project adapts various LSTM methods for symptom risk prediction on user-defined patient cohorts or on machine-derived patient clusters. This helps to evaluate how the model predicts outcomes for a target cohort. L-VISP contributes to visualization for model understanding by attempting to explain the underlying mechanisms of LSTM black-box models. This is done through visualizations that expose the model's memory and the evaluation of its predictions alongside ground-truth data. The model explanation is assisted by custom visualizations and by separate clinician-modeler front-ends, which separate different users' needs and provide a more guided analysis. These front-ends show predefined user workflows, which bring together various data facets, such as prediction statistics for a specified cohort, for clinician analyses, or prediction performance metrics for patient clusters. Last but not least, L-VISP introduces an encoding that helps to interpret the decision of the LSTM symptom predictions, which highlights weighted associations between symptoms and the underlying connections between said symptoms/model variables. The system was evaluated on a 937-patient cohort with

data modelers and a clinician.

The contents of this chapter are currently under review at the *Computer Graphics Forum* journal.

5.2 Motivation

Personalized head and neck cancer (HNC) care focuses on creating treatments tailored to individual patients based on cohort characteristics from similar patients. Unfortunately, cancer treatment often results in numerous side effects, which differ between patient cohorts and can last for a long time post-treatment. As a result, clinicians are collaborating with data modelers to understand treatment-related symptoms that appear or persist after treatment, to predict adverse outcomes, and to stratify patients into high-risk and low-risk cohorts. One of the significant challenges in post-treatment research is posed by the scarcity of cohort data, imposed by the patient monitoring protocol [40]. Patients are closely monitored during treatment when they come to the clinic to receive the prescribed doses, as opposed to post-treatment, when they come for biannual follow-up checkups [56]. As a consequence, post-treatment patient data is collected less often, posing a challenge in outcome prediction. Long Short-Term Memory Network (LSTM) methods have demonstrated excellent results for temporal patient outcome prediction, surpassing traditional statistical and machine learning methods, and have also gained attention in HNC symptom risk prediction [186, 187].

Post-treatment symptom risk prediction is a multidisciplinary field where data modelers collaborate with clinicians to model patient outcome risk, but this modeling often suffers from low interpretability. This is especially true when supervised black-box models, such as LSTMs, are used. Visual analytics can support this research; however, it needs to consider the differences in the mental models of the users. For example, clinicians are more interested in the actionable interpretation of the modeling outcomes and in the accuracy of the methods, which can be applied when treating new patients. Data modelers, on the other hand, are also interested in understanding the mechanisms behind the model's decisions and tools

that help them refine and debug modeling approaches. Moreover, post-treatment symptoms can result from the cumulative effects of various factors [154, 179], such as treatment-related complications or patient-specific health and lifestyle changes following treatment. Symptoms can also be associated with each other, either due to direct influence or due to shared root causes. Consequently, there is a growing need for analytical tools that support collaborative cohort modeling through workflows that enable experts to interpret machine-derived (modeled) results with real patient data.

Although data visualization is a valuable tool for supporting analytical tasks, in the context of post-treatment symptom prediction, it must overcome several challenges. To effectively interpret LSTM model behavior, data visualization must integrate diverse data facets from heterogeneous cohorts and support data modelers' and clinicians' analytical tasks. Specifically, data visualization needs to: compare multiple cohorts of interest to understand the impact of the modeled risk; support cohort stratifications by levels of risk to better understand prediction results; and blend cohort characteristics with results from different models to gain a deeper understanding of risk categories. Notably, LSTM symptom prediction visualization needs to overcome the cognitive burden associated with the high information density of LSTM models.

To address these challenges, we introduce a visual analytics system, L-VISP, developed for and with data modelers and with clinicians as secondary users, which supports post-treatment risk modeling in HNC patients. This work's main contributions are: 1) the domain characterization, developed alongside domain experts, of an application that targets model explainability in LSTM predictive analysis for head and neck cancer cohorts in collaborative clinician - data modeler settings, with a description of the modeling problem and design requirements; 2) data modeling with unsupervised and supervised approaches to stratify patient cohorts by risk using temporal clustering, and to model symptom severity using LSTM methods; 3) a human-machine system that is a collaborative bridge for data modelers and clinicians in clinical research. The system blends data visualization with data modeling to

interpret model outputs on multivariate, temporal patient cohorts. The system uses custom visualizations to blend different facets of the ground-truth data and outcomes from multiple models, and to validate the models on machine-generated (clusters) or user-defined patient cohorts; and 4) the evaluation of the resulting system by modelers and a clinician, along with a thematic analysis of their feedback, and lessons learned from this multidisciplinary collaboration.

5.3 Project Setting

This work is part of an interdisciplinary collaboration between three research groups located at different sites: two data modelers with experience in symptom modeling, three visual computing researchers with modeling experience, and a radiation oncology expert with clinical and modeling experience. All collaborators are coauthors of this work. Characterizing the application domain is a challenging task, primarily due to the exploratory nature of the research questions and the heterogeneity of the data. As a result, in this work, we used an Activity-Centered-Design (ACD) [129] paradigm, which is particularly suitable for scientific research, primarily due to the scarcity of trained domain experts and the importance of slow thinking [100] for scientific research. ACD prioritizes the user activity over the number of users.

Through a series of iterations, the visual computing team and the data modelers held weekly meetings to define functional specifications and to prototype the interface. This was an iterative process that evaluated incremental designs, starting with paper prototypes, then narrowing down proposed designs alongside data modelers, and finally moving to digital prototypes. Before evaluating the final system, the team met with the clinician to demonstrate the system. The clinician was not part of the design process, but provided feedback during the domain characterization and was part of the evaluation of the final system. The data modelers were part of the domain characterization, design, and evaluation stages.

Table 5.1: Symptom ratings example for two patients (P1, P2) and for two symptoms (taste, swallow) over time (B for baseline, W0 for first week post-treatment, W6 for six weeks post-treatment, M6 for six months post-treatment, and M12 for twelve months post-treatment)

Patient	**Taste**					**Swallow**				
	B	**W0**	**W6**	**M6**	**M12**	**B**	**W0**	**W6**	**M6**	**M12**
P1	0	4	6	5	4	1	5	6	7	5
P2	2	3	4	6	5	0	0	4	3	3

5.3.1 Task Analysis

L-VISP was built to primarily serve data modelers in cancer symptom research. The system evaluates two LSTM variants for symptom risk in the context of patient cohort data. Our modeler collaborators have experience in ML approaches for predicting patient outcomes, but they treat the predictive model as a black-box. Thus, they needed an overview of the model's behavior and its sensitivity to input variation, and an explanation of the output. Through the ACD paradigm, the regular meetings revealed the modelers' process, which involved running multiple scripts with numerous parameters and verifying each output plot individually. However, this process made it difficult to assess multiple outcomes concurrently on a desired cohort. Furthermore, the group was interested in having clinicians validate the modeling results. As a result, part of L-VISP's front-end targets clinicians with modeling experience. These front-end components support the clinician's interpretation of symptom predictions on cohorts of interest. They are dedicated to one of the primary user activities, which include both clinician and modeler tasks, presented in detail in Section 5.4.

We identified several key tasks and, following the ACD framework, grouped them into two main activities: the first one for analyzing model performance and behavior for the patient population, which is stratified by symptom severity using temporal clustering, and the second one for analyzing the model performance for a target, user-specified patient cohort. Other works in LSTM modeling visualization focus on either finding patterns in prediction trajectories [85], comparing predicted data against ground-truth data to find

prediction errors [36], or visualizing model hidden states [175] to understand the model's decisions. Our activities, however, need to support a combination of these tasks. In addition, we compare modeling results for two patient cohorts and predictions between data items (i.e., symptoms). We present below the two activities, which are composed of several visualization tasks:

- **A1** Stratified model evaluation for patient clusters
  - **T1** Validate model predictions by comparing ground-truth symptoms to predicted symptoms, to test the model accuracy
  - **T2** Analyze the model behavior by exposing the model memory and evaluating results on different cohorts, to understand underlying mechanisms
  - **T3** Compare model results between cohorts by analyzing prediction trajectories and performance metrics, to find model deficiencies

- **A2** Targeted model evaluation for user-defined patient cohorts
  - **T4** Examine input-output relationships in the model by evaluating predictions under different cohort attributes, to test the model's robustness
  - **T5** Evaluate model performance for a given cohort by comparing predictions between desired cohorts against the rest of the patient population, to test the model accuracy and find cohorts for which the model predicts negative outcomes
  - **T6** Find model connections to symptom severity by evaluating predictions under different symptom severity thresholds, to test the model's accuracy and find symptoms linked to high risk of negative outcomes

Although these activities were extracted in accordance with data modelers' needs, A2 was documented to support clinician analysis as well. The clinician provided occasional feedback on the tasks as mentioned earlier during the domain characterization phase, which helped to define the activities. Our clinician collaborator was not interested in understanding the

mechanisms of the LSTM (A1), but in analyzing outcomes on groups of patients of interest (A2). Consequently, L-VISP becomes a collaborative tool that facilitates joint workflows. First, the data modeler uses A1 to debug and gain confidence in the LSTM model. Afterwards, the modeler and clinician can collaborate in A2 to validate the model's predictions against clinical intuition and explore outcomes for desired cohorts.

5.3.2 Data

The data was collected from a cohort of 937 head and neck cancer patients from the MD Anderson Cancer Center in Texas, treated with radiation therapy (RT). This dataset is a more comprehensive dataset compared to the ones used in Chapter 3 and Chapter 4. We refer to the whole dataset as the patient population, and to a subset of the patient population as a patient cohort. The data includes clinical and treatment attributes, and patient-reported symptom ratings. Clinical data includes demographics such as age (quantitative), gender, and smoking status (nominal); diagnostic attributes such as tumor size and lymph node stage (ordinal), tumor site (nominal); and additional treatments: induction therapy (IC), concurrent therapy (CC), and neck dissection surgery (ND) (nominal). The clinical data is visualized to support the analysis of the patient clustering and symptom prediction. We use symptom ratings and treatment attributes for symptom prediction, and symptom ratings for patient clustering.

The MD Anderson Cancer Center studies symptom severity in patients through a quality-of-life monitoring program. The program involves patient-reported outcome (PRO) questionnaires based on MD Anderson Symptom Inventory-Head and Neck Module [40], also known as MDASI-HN. This 28-symptom questionnaire is used for clinical and research purposes, in which patients are asked to rate symptoms using a 0-10 scale, from *"not present"* (0) to *"as bad as you can imagine"* (10). Symptoms are split into three categories: HNC-specific, general cancer, and daily interference symptoms. The PRO data are temporal and multivariate, collected before, during, and post-treatment throughout a total of 12 time points. Over half of this data collection happens during treatment, when a spike in symptom sever-

ity is expected due to treatment toxicity. In this work, we aim to identify patients who experience late symptoms well after treatment has concluded. Our data analysis evaluates modeling results based on several rating thresholds with clinical significance: $\geq 1 < 3$ (labeled as ≥ 1 in the front-end) for mild symptoms, $\geq 3 < 5$ for mild-to-moderate symptoms (labeled as ≥ 3 in the front-end), and ≥ 5, which clinically stand for moderate-to-severe symptoms. These thresholds were chosen based on our clinician collaborator feedback, previous clinical research on the MDASI-HN questionnaire [3, 164, 165], and our previous projects with the questionnaire [21, 60, 61, 186, 191].

We use the PRO symptom data before treatment, or baseline (B), at the end of treatment (W0), and during post-treatment at 6 weeks (W6), 6 months (M6), and 12 months (M12) (Table 5.1). Although we used all 28 symptoms to cluster patients by symptom severity, we performed LSTM modeling. We used visual analysis to evaluate and understand the model's decisions for the 9 HNC symptoms: swallow, speech, mucus, taste, constipation, teeth, mouth sores (mucositis), choking, and skin problems. We used all 28 symptoms in patient clustering to capture a comprehensive view of symptom burden and patients' variability. The selection of the 9 HNC symptoms for LSTM modeling followed, driven by clinical relevance and the need for focused analysis on symptom categories. Our collaborators showed particular interest in the HNC subset of symptoms before extending the analysis to the daily interference and general cancer categories. The modelers aimed to present results to clinicians for this symptom subset to inform planning for further patient cohort modeling projects. Visualization considerations for supporting the analysis of 9 symptoms include the front-end's limited real estate and the need to limit cognitive load due to high data density, aspects discussed in Section 5.6.

In addition to the ground-truth patient data, we visualize the results of our models (Fig. 5.1). The data modelers used the Bi-directional LSTM (Bi-LSTM) [186] for symptom severity prediction and the Interpretable Multi-Variable LSTM (IMV-LSTM) [78] for Bi-LSTM modeling understanding. As a result, we used visual analytics to support the evaluation

of the Bi-LSTM predictions, the Bi-LSTM performance metrics, and the IMV-LSTM features; all described in detail in Section 5.4.1. Additional cohort modeling is done with the Time2Feat time-series clustering method [22], which the modelers used to extract three patient clusters with different symptom severity thresholds. These results are integrated into our visualization system as well.

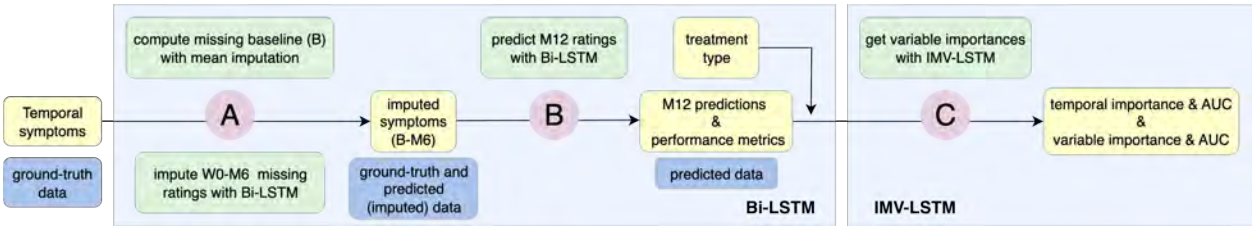

Figure 5.1: LSTM symptom modeling pipeline. A) Imputation of missing symptom ratings, for all time points up to the last one (M12) using Bi-LSTM; except for the B time point, which is imputed using mean B imputation. B) Bi-LSTM prediction for symptoms at the last time point, M12. C) Application of IMV-LSTM on the Bi-LSTM imputed values to extract temporal and variable feature importance for all symptoms.

5.4 System Design

This section presents the back-end modeling, namely the LSTM-based symptom prediction and the patient clustering method, and the front-end components that visualize the modeling results. L-VISP uses Python for the back-end and React with D3.js for the front-end. All modeling is computed offline before being loaded into the front-end.

5.4.1 LSTM Symptom Modeling

Our modeling approach uses a pair of variants from the LSTM family for symptom risk analysis (A1, A2). The first is a Bi-directional LSTM (Bi-LSTM) [168], which acts as our primary prediction model. By processing a patient's timeline in both forward and reverse, it gains a deeper context to forecast future symptoms. Specifically, Bi-LSTM contains two LSTMs that go in opposite directions, which allows them to capture upstream information and additional context at each time point. After running the LSTM in both directions, the hidden states are concatenated, i.e., the dimension of the hidden states, before generating the final output. This ensures that more information is gathered, which improves the final prediction results. The second LSTM model is an Interpretable Multi-Variate LSTM (IMV-

LSTM) [78], which serves as the explainer, and looks inside the Bi-LSTM to understand how it arrives at its predictions by exposing its memory/hidden states. We applied Bi-LSTM and IMV-LSTM on the nine symptoms of interest.

Our primary goal was to predict long-term symptoms at the 12-month mark (M12), which corresponds to long-term/chronic symptoms. To prepare the model for this task, we first trained the Bi-LSTM to impute missing ratings from earlier and subsequent time points, ensuring it had a complete patient history to learn from before making its final (M12) prediction (Fig. 5.1). Specifically, the modelers first imputed the missing baseline (B) ratings using the mean values of the entire cohort (Fig. 5.1.A). In their preliminary work [187], the modelers have experimented with other imputation methods, such as a K-nearest neighbors baseline imputation. Compared to the mean baseline imputation, the results have shown a similar AUC performance on the month 12 (M12) prediction. Given the comparable performance, they opted for a mean baseline imputation for the present project. They then trained the Bi-LSTM model recursively on the preceding time points and let it predict the current time point for all time points (W0-M6) before the last recorded one (M12) (Fig. 5.1.A). Using 3-fold cross-validation, our collaborators used the B-M6 imputed symptoms (Fig. 5.1.B) and trained the Bi-LSTM model over two training folds to predict M12 symptom values for all patients in the test fold. Model performance metrics were extracted, including the Area under the curve (AUC), F1 score (micro averaging), Precision, Recall, and Root Mean Square Error (RMSE) for each symptom.

To understand the reasoning behind the Bi-LSTM's predictions, we used the IMV-LSTM. This second model analyzed the complete patient history to determine the temporal importance (i.e., which symptoms across the B-M6 time points were most influential) and the variable importance (i.e., which other symptoms or treatments were most influential) for the final M12 prediction. Following the terminology in [78], we define the feature importance from the IMV-LSTM during the intermediate time points (B-M6) as temporal importance, and the final feature importance predicting M12 as the variable importance. IMV-LSTM

defines a hidden state matrix to monitor and obtain both the temporal importance (5.1) and the variable importance (5.2) from the LSTM hidden states. By using a mixture-attention mechanism, IMV-LSTM applies temporal attention/importance to the sequence of each variable's hidden states to obtain a summary of each variable's history. After that, variable attention/importance is computed from each variable's history-enriched hidden states. The mathematical definitions of the temporal importance and the variable importance for a given symptom are shown below:

$$A = \frac{1}{M} \sum_{m=1}^{M} A_m; A_m = [\alpha_{1,m}, ..., \alpha_{T,m}] \tag{5.1}$$

Where A is the temporal importance vector computed by taking the average of the attention weights α for all the data instances, M is the number of patients, and T is the number of time points preceding the last one (B-M6). In our context, the temporal importance of each symptom is the average of the attention-weight vectors over all the patients. To derive temporal importances for each of the 9 HNC symptoms, we trained nine separate interpretable multivariate (IMV)-LSTM models, with each model targeting one specific symptom. For each model, the predictors included three treatment conditions and all other HNC symptoms except the target symptom. This design allows the temporal importance scores to capture not only the contributions of preceding time points for the same symptom, but also the cross-symptom associations that influence the prediction of the target outcome. In a clinical context, given a symptom, such as pain, the temporal importance reflects how both within-symptom history and other symptom trajectories (e.g., swallow, taste, voice, etc) jointly contribute to predicting pain at M12. A high score means the model found a strong influence for a symptom at a time point preceding M12 to predict another symptom's M12.

$$B = \frac{1}{M} \sum_{m=1}^{M} B_m; B_m = [\beta_m^1, ..., \beta_m^n] \tag{5.2}$$

Where B is the variable importance computed by taking the average of the posterior probability β for all the data instances (patients) across all input variables, M is the number of

patients, and n is the number of input variables (9 symptoms and three treatment conditions). The resulting posterior probability is computed by a softmax layer of a neural network, whose input combines attention-weighted summary with the hidden state vectors of each variable (symptoms and treatment conditions). In a clinical context, for each predicted symptom (e.g., pain), the model calculated the importance score for all other symptoms (e.g., taste, voice, choke, etc) and treatments (RT, IRT, ICC). This score represents the influence of those other symptoms and treatments on the final M12 prediction.

The modelers conducted the same analysis for each symptom to obtain the temporal and variable importance. They first removed the symptom to predict from the training data to avoid the target symptom from dominating the variable/temporal significance. After extracting the temporal and variable importance, they obtained relations between each symptom and all the other symptoms. They used the IMV-LSTM's Area Under the Curve (AUC) score as a quantifier of the strength of the relationships. The higher the AUC, the better the model can distinguish between the positive and negative classes; thus, the more confident we can say that the temporal and variable importance patterns help make accurate predictions.

Both LSTM models were trained using the Mean Squared Error loss function with early stopping. Parameter tuning was performed using an 80/20 data split. The Bi-LSTM model used one recurrent layer with a size of 10 and trained with a learning rate of .0215 using Stochastic Gradient Descent. The IMV-LSTM used a hidden state size of 128 and was trained on a learning rate of .001 and a weight decay of .9 using the Adaptive Moment Estimation optimizer. The Bi-LSTM training on an NVIDIA RTX 4080 platform with a 3-fold cross-validation required, on average, 4.2 seconds per time point and around 17 seconds in total, while the prediction on the test set was completed in less than 0.1 seconds. These runtimes indicate that the modeling is computationally efficient and suitable for offline analysis.

5.4.2 Multivariate Temporal Patient Clustering

A key goal for our collaborators was to stratify the patient population based on the overall severity of their symptoms over time (A1). This was needed to find patients with high, medium, and low risk of symptom burden. This was a complex task, as each patient's trajectory involves multiple symptoms that evolve in unique ways. Our collaborators have experimented with several patient clustering methods before [60, 130], which could not capture the multivariate time series nature of our data well. Most temporal clustering methods that consider either univariate or multivariate time series (e.g., Dynamic Time Warping, K-Shape, CSPCA, MC2PCA) [22] suffer from poor explainability, and the original data dimensions are lost. As a result, in this work, our collaborators explored Time2Feat [22], which is specifically designed for complex time-series and aims to create understandable clusters. This method focuses on interpretable features extracted from time series and uses dimensionality reduction on subsets of features that retain the most information, providing highly interpretable results. The technique has demonstrated higher effectiveness, interpretability, efficiency, and robustness over several state-of-the-art multivariate time series clustering methods [22].

The modelers used the PRO symptom data, which they considered time series data (B-M12 time points) with 28 dimensions (symptoms), to cluster patients based on temporal symptom severity. In this project, we used the unsupervised mode of the Time2Feat method, which is fully automatic and uses Principal Component Analysis (PCA) to find the symptoms that best stratify the patient cohort by symptom severity. The modelers experimented with several clusters as input for this method, from two to seven clusters. In the end, they decided to further evaluate the results for three patient clusters, which represent patient groups with mild, medium, and severe symptoms. The three-cluster results showed a balanced stratification, with a 27/48/25% split. We evaluate the Bi-LSTM modeling for these patient clusters to better understand prediction patterns across cohorts.

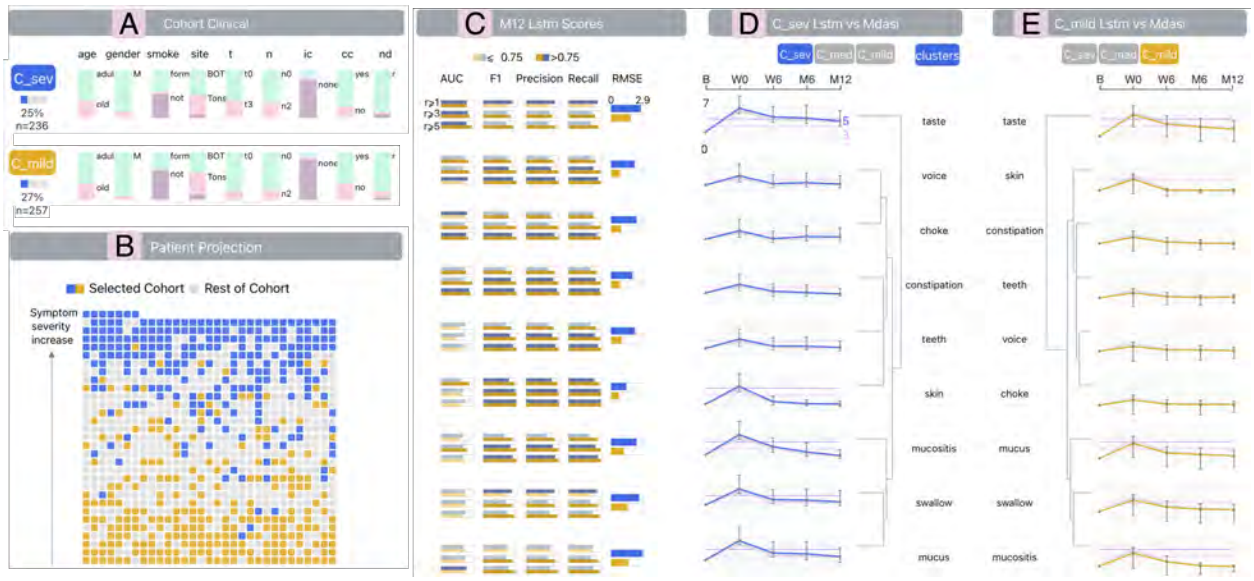

Figure 5.2: LSTM model performance on patient clusters (modeler activity). A) Clinical component showing two patient clusters, stratified by symptom severity. B) Patient projection (cluster with severe symptoms in blue and with mild symptoms in yellow). C) LSTM metrics, which are similar for the two clusters; taste has the highest AUC scores for all rating thresholds. Symptoms' vertical order corresponds to the dendrogram order in D). LSTM-predicted symptom trajectories for D) the severe symptom burden cluster (blue) and E) for the mild symptom burden cluster (yellow) against the ground-truth (gray area). The LSTM model shows mild underpredictions for the first cluster and overpredictions for the second. The dendrograms show similar trajectory groupings between the two cohorts.

5.4.3 Front-end Design

Our front-end design comprises several user panels with coordinated components, presented below, that support the tasks involved in the two main activities (A1, A2). Tooltips provide further details upon hovering over any element, and user-selected patients are highlighted in magenta across the front-end.

Cohort Attribute Distribution

The cohort attribute distribution component (Fig. 5.2.A) displays the distribution of clinical attributes, allowing for the selection of cohorts of interest for model evaluation (T1-6). Stacked bar charts display demographic and diagnostic characteristics for each patient cohort, with labels highlighting attribute values present in over 20% of the patient population. Smaller distribution values are visible upon hovering over a stacked bar. This component provides a clinical snapshot of each patient group. For the patient clusters evaluated in A1 (Fig. 5.2.A), buttons enable the selection of a cluster. For the custom, user-defined cohorts

evaluated in A2 (Fig. 5.6.A), this component helps to compare attributes between a cohort of interest and the rest of the population, and dropdowns accompany the attributes to support cohort queries. We chose this compact, horizontal layout because it handles a large number of clinical attributes and easily compares different cohorts at a glance.

Patient Projection

The patient projection component (Fig. 5.6.B) uses an interactive matrix where each cell is represented by an individual patient. It represents the patient population (e.g., the whole dataset of patients), and it shows how cohorts are clustered. The matrix is interactive, supporting brushing a single/group of patient(s) of interest. This component is used in both activities (A1, A2) by both modelers and clinicians to relate modeling results to the actual patients during each task (T1-T6).

Blue cells highlight the patients within the selected cohort, while gray cells represent the rest of the patient population. Patients chosen directly from the matrix are highlighted in magenta. For cohort comparison (Fig. 5.2.B) (T3), the second selected cohort is highlighted in yellow. Patients are sorted by overall temporal symptom severity in a list (T5). We then populate the matrix with the patient list from the bottom left corner. In this way, low-symptom-severity patients correspond to the bottom of the matrix, and high-symptom-severity patients correspond to the top Y positions. The left-to-right direction (X axis) corresponds to an increase in severity per row. We used this projection method as opposed to others (e.g., PCA, t-SNE, UMAP) as it showed the best visual stratification of the patient population, with fewer outliers in the low and high symptom severity clusters extracted in Section 5.4.2.

In our previous work with similar patient cohorts, we have experimented with scatterplot projections to visualize the patients, either by projecting the whole population with overlapping glyphs [60], or using groups of patient projections with no overlaps on larger cohorts [61]. Rather than using a traditional scatterplot, we represent the patient population

Figure 5.3: Bi-LSTM performance metrics for two symptoms at M12 time point.

with an interactive matrix. We drew inspiration from previous work that employs matrix representations for cohorts; however, these works use matrices for cohort summaries, not for representing individuals within a cohort [113, 200, 203, 206]. Our approach avoids the visual clutter and glyph overlap that often make scatterplots unreadable on large cohorts (n >900), and it aims to reduce cognitive load. Because patients are consistently ordered from top to bottom by decreasing symptom severity, users can easily switch between analyzing pre-defined clusters and their own custom cohorts. It also provides a scalable overview of the patient population, and it could be used on larger cohorts in exchange for reducing the size of glyphs/cells.

Performance Metrics

We use the performance metrics component (Fig. 5.2.C, Fig. 5.3) to evaluate the performance of the Bi-LSTM at M12 (T1). Bar plots display relevant metrics for each symptom for model performance under different rating thresholds: r≥1, ≥3, ≥5; which are clinically considered as mild, moderate, and severe symptoms, respectively (Fig. 5.3). The bar plots are rotated by 90°due to limited vertical space per symptom. We highlight good (> 0.75) performance metrics (e.g., AUC, F1 score, Precision, Recall) with dark blue, and the rest with light blue. The RMSE metric values are not reported for rating thresholds; therefore, they are represented using the cohort's standard color. Tooltips display the values for each metric/bar upon hovering, and the symptoms' order is given by the first symptom dendrogram/list of trajectories (Fig. 5.2.D). For cohort comparison, we depict the values of a second

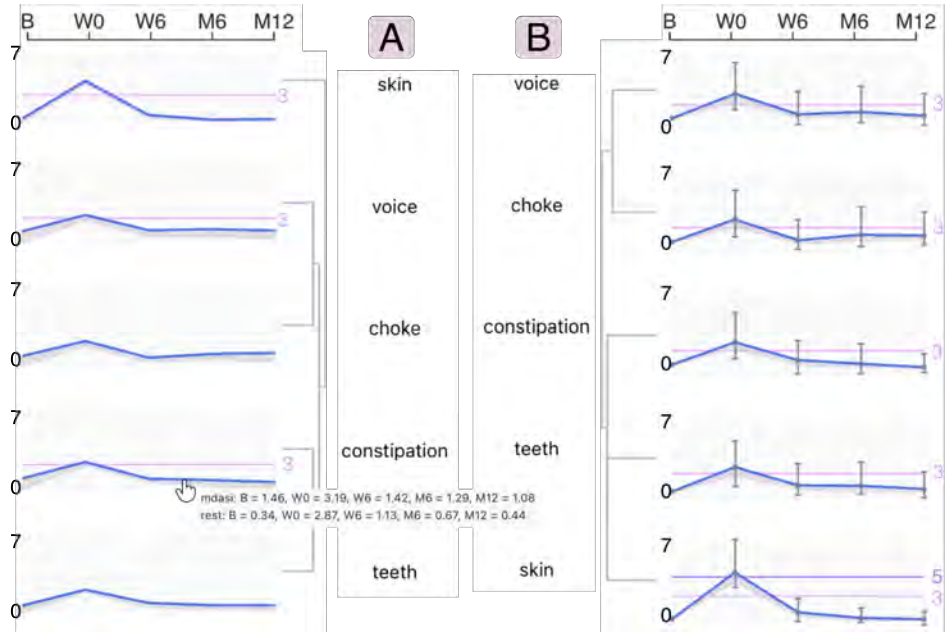

Figure 5.4: Ground-truth vs. predicted symptom trajectories for the severe symptom burden patient cohort. A) ground-truth symptom trajectories for the cohort (blue) against the population (gray) and B) predicted trajectories for the cohort (blue) against the population (gray) show similar symptom clusters in the dendrograms, with over-predictions at the end of treatment (W0) for all symptoms. Trajectory values span the [0,7] interval, and trajectory surpassing rating thresholds of interest (i.e., rating $\geq$3 for mild-to-medium severity, $\geq$5 for medium-to-severe symptoms) are highlighted with pink and purple threshold lines.

selected cohort using light and dark (score $>$ 0.75) yellow highlights. The grid-based display supports pattern and outlier detection in the metrics through the side-by-side positioning. While this view uses standard statistical charts, it displays these metrics side-by-side for different input thresholds, and the grid layout makes it easy to spot which symptoms and for what input conditions the model predicts well, and where it does not (T1).

Symptom Trajectory

The symptom trajectory component (Fig. 5.4.D, E) uses a lineplot to visualize how symptom severity changes over time. It can compare the model's predictions to the ground-truth data (T1) or contrast the symptom trajectories of a selected cohort with those of the rest of the patient population (T5). A blue line represents the predicted values for a given cohort. At the same time, the gray area highlights the difference to the ground-truth values (T1) (Fig. 5.4.D, E) or to the predictions of the rest of the patient population (T4) (Fig. 5.6.D, E). The distribution of Bi-LSTM mispredictions is represented using vertical gray bars (upward

direction for the count of overpredictions, and vice versa for underpredictions) at each time point. Brushed patients from the patient projection matrix are highlighted via magenta lineplots (T5) (Fig. 5.6.D, E). For cohort comparison (T3), the trajectories for a second cohort are depicted in yellow (Fig. 5.2.D, E). We chose this common encoding for its utility in pattern detection, as seen in other LSTM visualization work [85, 175]. Its key advantage is its versatility, which adapts to various analytical tasks for cohort time-series.

We clustered the symptom trajectories using hierarchical clustering (HC) to find consistent symptom grouping between predicted and ground-truth trajectories (T1, T4) across cohorts (T3) (Fig. 5.4). We used Euclidean distance with the Average metric for the symptom clustering. Still, we have previously experimented with other similarity search methods for time series, such as Dynamic Time Warping (DTW), Symbolic Aggregate Approximation (SAX), Cosine Similarity, and Pearson Correlation, as well as with other clustering linkage methods, including Complete and Ward. These methods showed either outlier sensitivity, did not have similar trajectories within-cluster, or showed significant variability in cluster formations across the patient clusters. Ultimately, we selected this method because it performed best for time series with comparable shapes and magnitudes, such as our symptom trajectories. It also did the best job of creating distinct and meaningful symptom clusters, but similar clusters across patient cohorts (T3, T5). We ordered the symptoms based on the HC results for each cohort. We used accompanying dendrograms, displayed through a mirroring technique, to represent the patterns between symptom clusters (T1, T4) for two patient cohorts (T3) (Fig. 5.2.D, E). The dendrogram dictates the vertical order of the symptoms in the other components. In the case of cohort comparison, the first dendrogram dictates the vertical symptom order in the other components (Fig. 5.2.D, E). In this way, the user can analyze a single symptom horizontally, across visual components. To reduce the cognitive load of identifying clusters and comparing trajectories across different rows in two cohorts, we offer an option to hide the symptom clusters and dendrograms. In this case, we list the symptoms in the same order across cohorts, which is based on all symptom

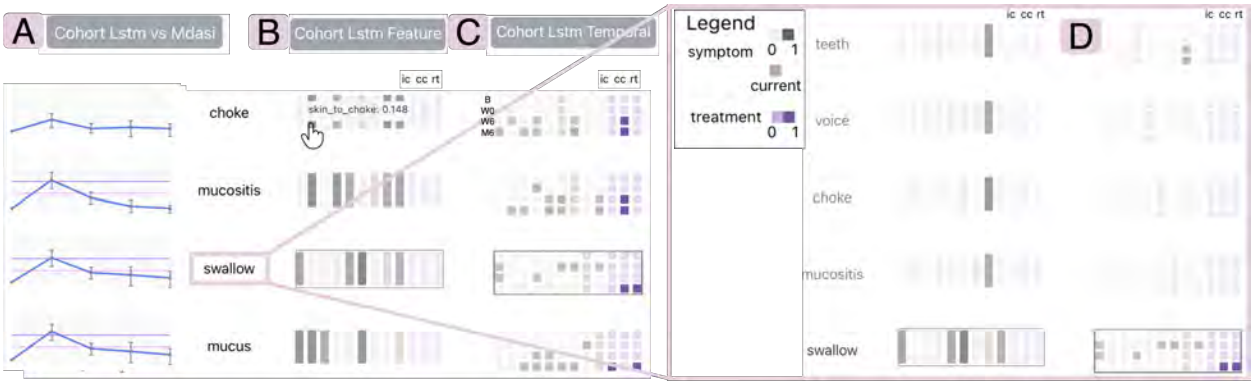

Figure 5.5: Model behavior for the medium symptom burden patient cohort. A) Predicted symptom trajectories with minimal mispredictions. B) Variable importance and C) Temporal importance. Hovering on swallow's row D), swallow's associations with mucositis, choke, voice, and teeth are shown from B), and high values in W6 and M6 time points (e.g., teeth) from C), meaning associations with M12 rating prediction.

clusters across cohorts. However, the dendrogram's goal is to help identify highly similar symptom behaviors. This is important, given that patients generally exhibit the same overall trend in symptoms, but with different severity thresholds, and a rise in severity at W0, as a consequence of the treatment's influence [40].

Since we visualize mainly mean values for a given cohort, we didn't specify the exact numerical differences between the ground-truth and the predicted values. We opted to juxtapose these differences (e.g., blue vs. gray or yellow vs. gray in Fig.5.2.D, E) or to visualize trajectories side by side to compare ground-truth and predicted symptoms (Fig.5.5). For numerical values, we provide tooltips with ground-truth and predicted values across time points upon hovering over a symptom trajectory.

Temporal and Variable Symptom Importance

To look inside the model's black box, the variable importance components (Fig. 5.5.B) explain the Bi-LSTM's behavior. Using importance scores from the IMV-LSTM, it highlights which features (i.e., symptoms and treatments) contribute most to the prediction of a given symptom (T2). We can see the variable importances as weighted associations between the symptoms and treatment type and the M12 prediction for a given symptom. We use a matrix representation to show the variable importance of each symptom in predicting a given symptom. Each row lists the symptoms based on the dendrogram order from the

symptom trajectory component, and it visualizes the mean variable importance of a cohort for each symptom. The same order is followed in the columns. Brown corresponds to the current symptom, a gray color scheme is for the rest of the symptoms, while a purple color scheme is for the global importance of the treatment types. Lighter colors and lower opacity correspond to lower variable importance, while darker colors and higher opacity encode higher values. For each symptom, we highlight reliable results, characterized by a high AUC (>0.75), using a dark margin around the corresponding symptom's row. When hovering over a symptom label, the corresponding variable importance is highlighted in the row (Fig. 5.5.D) to illustrate how the symptom affects the predictions of other symptoms. During hovering, the column corresponding to the given symptom is highlighted to show the influence of the other symptoms in the prediction of the given symptom. The design is inspired by previous work for LSTM hidden states visualization [175]; however, we visualize both variable importance and temporal importance for each symptom. Moreover, we took inspiration from related work on cohort summaries using matrix encodings [113,200,203,206], but we highlight which items (i.e., symptoms) show reliable associations with other items through the rows' dark margins.

The design of the temporal importance component mirrors that of the variable importance component (Fig. 5.5.C). This consistency makes the temporal importance analysis more intuitive and easier to follow (T2). It shows the IMV-LSTM-generated importance of the symptoms' time points (i.e., B-M6) in predicting the M12 rating for a given symptom. In other words, we can see the temporal importance as a weighted association of the symptoms' time points to the prediction of the M12 rating. Following the matrix-based design, the row visualizes the mean temporal importance of a cohort for a symptom. Each row is split into smaller cells by time points on the vertical axis. The same interactions with the variable importance components apply here as well. Upon hovering over a symptom, the row highlights the influence of the given symptom on the rest. The highlighted corresponding column shows the impact of the other symptoms on the hovered one.

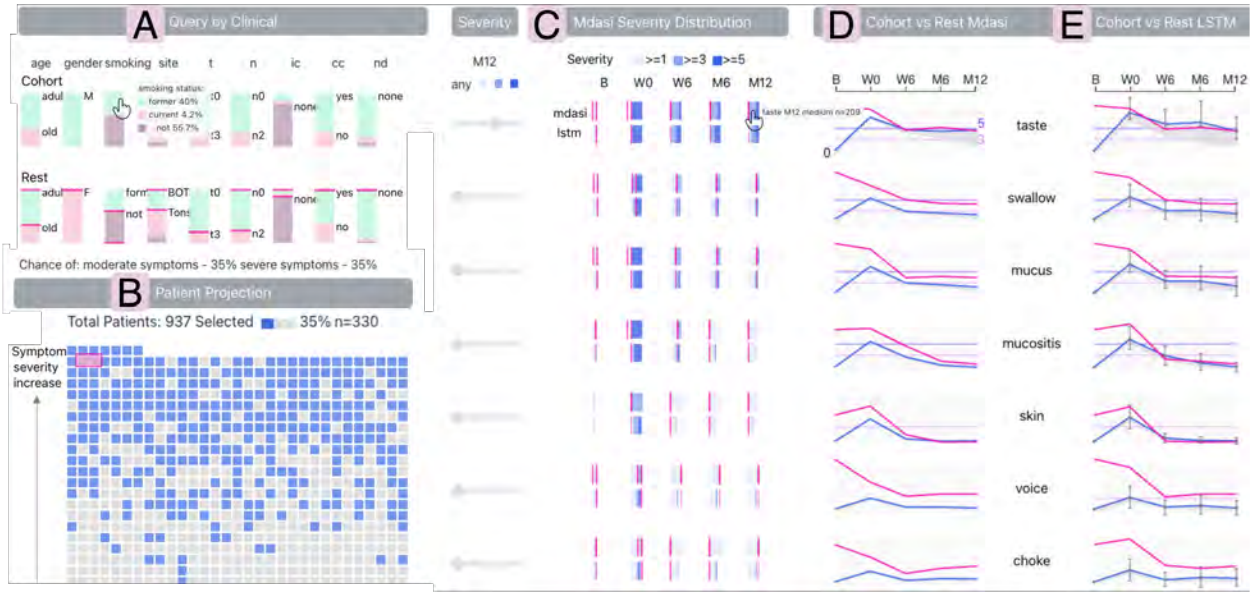

Figure 5.6: Model performance analysis for a custom cohort (clinician and modeler activity). A) Clinical component showing the queried patients: males with medium taste problems (right C) filters) that represent 35% of the patient population. B) Patient projection based on symptom burden, showing the queried patients on the upper, more severe half, which is represented by high symptom burden. Two female outliers (gray cells) are highlighted in the projection (magenta highlight in all components). C) Symptom severity distribution showing higher and similar prevalence for taste, swallow, mucus, and mucositis. Filters for symptom severity are displayed on the left of the component. D) Symptom trajectory for ground-truth and E) for predictions, showing differences in ground truth vs. predicted trajectories at W6-M12 for taste and significant differences from the brushed patients (magenta) in all symptoms at B-W0.

Symptom Severity Distribution

In a similar fashion to the performance metrics component (Fig. 5.2.C), this component (Fig. 5.6.C) uses a grid-based representation to display the temporal severity distributions. The top rows represent the ground-truth severity distributions and the bottom rows the predicted symptom distributions (T1, T4). With a similar design to the cohort attribute distribution component (Fig.5.4.A), each cell is represented by a stacked bar chart showing the distribution of symptoms (rating > 0), with light-to-dark blue colors representing mild ($r \geq 1$), medium ($r \geq 3$), and severe ($r \geq 5$) ratings. The horizontal bars support the side-by-side comparison between the ground-truth and predicted values. Tooltips provide numerical values upon hovering over the distribution rows. For custom cohort model analysis (A2), each symptom is accompanied by a severity slider, which filters the patients with the corresponding severity for the last time point, M12 (T5). The component highlights patterns in symptom presence and shows which symptoms are severe and need more attention during

analytical workflows (T6). Brushed patients from the patient projection are highlighted with magenta borders in this view. We chose this grid layout over other chart types because it provides a compact and efficient display of temporal, multivariate data. Different designs were considered, but would have occupied more screen space, such as box/violin plots and pie charts. This grid layout facilitates item (symptom) comparison, as well as comparison of temporal and data provenance (ground truth vs. predictions).

5.4.4 Workflows

The stratified model evaluation activity (A1), supports modeler tasks using predefined workflows represented by several panels. These panels are used to validate the Bi-LSTM and IMV-LSTM results on three patient clusters. An example is presented in Fig. 5.2, where the Bi-LSTM results are validated (T1) and compared (T3) between the patient clusters with mild and severe symptom burdens. After the selection of the two cohorts from the clinical component (Fig. 5.2.A), the cohorts are highlighted in the projection (Fig. 5.2.B). Their corresponding prediction metrics are represented in the corresponding component (Fig. 5.2.C) (T1). The cohort predictions are compared to the ground-truth on symptom trajectories (Fig. 5.2.D, E) (T3). Alternatively, another workflow examines the Bi-LSTM memory exposition to understand the model's behavior (T2) and its prediction decisions by analyzing the IMV-LSTM temporal and variable importances (Fig. 5.5.D, E) for a given cohort (Fig. 5.5.A).

A third panel supports the targeted model evaluation (A2), on user-defined cohorts (Fig. 5.6). This configuration enables both data modelers and clinicians to analyze modeling results for a specific cohort of interest. This workflow provides fewer model behavior details (e.g., no memory exposition) to lower the cognitive load during clinician analyses. The user can better understand how sensitive predictions are on diverse patient attribute inputs (T4) once they select a cohort from the cohort attribute distribution component (Fig. 5.6.A). The selection is highlighted in the patient projection (Fig. 5.6.B), alongside their symptom severity distribution on the right (Fig. 5.6.C) (T5), which connects symptom presence with prediction results. The selected cohort's ground-truth (Fig. 5.6.D) and predicted (Fig. 5.6.E)

trajectories, compared to the rest of the patient population (T5), are represented on the right side of the panel. Further selections on the patient matrix projection will highlight the corresponding patients in magenta. This panel and activity provide a common ground for hypothesis generation in clinician-modeler collaborations.

5.5 Evaluation

We evaluated L-VISP qualitatively through demonstrations and case studies with two data modelers, the visual computing team, and a senior research oncology expert, who had ML experience. The data modelers participated in the design of the visual analytics system and in the model building, and the oncologist provided occasional input and feedback. All evaluators are coauthors. Although the system serves modelers in cancer research, validating the results with a clinician was essential to ensure their clinical relevance.

The evaluation was based on pair analytics [11], where the main visual analytics designer was the navigator of the visual analytics tool (L-VISP), and the collaborators were the drivers of the tool. Although pair analytics requires two participants per session, we organized group evaluation sessions due to the collaborators' limited availability. Additionally, we observed that group sessions helped to generate more hypotheses and feedback. However, these sessions usually had two main drivers, namely a data modeler and the clinician. The evaluation was conducted online, through screen sharing, starting with demonstrations of the tool and then walking through case studies. The drivers (evaluators) were encouraged to think aloud and make hypotheses while the navigator was driving the interface, and a navigator helper (from the visual computing team) was taking notes. The data modelers described two case studies, presented below, on a cohort of 937 HNC patients treated at the MD Anderson Cancer Center, and the results were validated by the clinician.

5.5.1 Blended Models Insights and Evaluation

In the first case study (A1), the modelers were interested in evaluating the symptom modeling results on the pre-computed patient clusters (Fig. 5.2) and getting insights into how the

Bi-LSTM makes predictions (A1). The patient projection revealed that the clusters were separated by temporal symptom severity into the severe (top), the medium (center), and the mild cluster (bottom) (Fig. 5.2.B). Selecting the severe patient cluster, the modelers observed that it consistently showed higher severity in predictions as opposed to ground-truth data across all time points and all symptoms (Fig. 5.4) (T1). The modelers observed that most overpredictions occurred in W0, which was typically the highest-rated time point, and noted this biased result for future refinements. The symptom clusters highlighted by the dendrogram were similar between the ground-truth and predicted symptoms (T1, T3) (Fig. 5.4), showing that the model captured the same temporal patterns as the ground-truth (T1). The Bi-LSTM predictions revealed three consistent clusters across cohorts, with 'taste' as the first cluster (Fig. ref fig: fig51), 'Fig. 5.4' representing the second, and 'mucositis, swallow, and mucus' as the third (T3).

There were more high AUC and F1 score values for the mild patient cluster, suggesting that the model captures lower symptom ratings more effectively (Fig. 5.2.C) (T3). The highest RMSE was observed for M12 in the taste prediction, while the lowest was for skin problems. This was verified in the timelines (Fig. 5.2.D, E) where skin consistently had the lowest mean in M12, suggesting that a lot of people might not report skin (T1), while taste had the highest mean, which was expected as taste has shown to be one of the most prevalent symptoms in our previous symptom research work [61]. The modelers and the oncologist agreed that taste showed more severe and persistent patterns *"We see that taste is its own thing"*; *"I am not surprised taste is so common"*. The dendrograms highlighted similar symptom clusters for both the mild and severe patient groups (T3). This finding, despite the groups' differing severities, suggested that symptoms have consistent trajectory groupings regardless of severity. When checking the Bi-LSTM performance across clusters, (Fig. 5.2), the model consistently showed overprediction for the severe cluster, and underprediction for the mild cluster, across all symptoms (T1,3) (Fig. 5.2.D, E). The modelers agreed that the model tends to be more sensitive to severity extremes in the patient cohort.

The hidden states of the Bi-LSTM black box (Fig. 5.5) showed that only a couple of symptoms were highlighted as reliable predictions (T2). Swallow unsurprisingly showed associations with predictions for symptoms connected to the salivary domain (Fig. 5.5.D). However, its association with teeth was an unexpected finding, which the oncology expert suggested needed to be further investigated *"it's hard to tell the root cause of tooth pain, it can be from choking or pain, or a reflection of mucositis problems."* The temporal importance (Fig. 5.5.C) showed that most of the symptoms tend to be associated with M12 predictions at the end of the patient observation period, in W6 and M6 (T2), but did not show any common symptom patterns with the variable importance, which was surprising to the modelers.

5.5.2 Model Output Analysis for Targeted Cohorts

In the second case study (A2), the modelers, together with the clinician, explored the cohort to evaluate how the model predicted symptoms in subsets of patients (Fig. 5.6) (A2). They retrieved the male patients (Fig. 5.6.A)) with medium taste severity in M12 (Fig. 5.6.C), which were grouped at the top of the patient projection (Fig. 5.6.B), where patients with higher overall symptom severity were displayed. When selecting two outlier females against this cohort (Fig. 5.6.B) (T4), the clinician observed some high the Bi-LSTM predictions at M12 for swallow, voice, and choke, suggesting that the brushed patients show a higher risk for these symptoms (Fig. 5.6.E) (T5-6). The modelers recorded the patients' IDs of these outliers for further investigation.

Next, the evaluators checked how the Bi-LSTM model performed on the selected cohort (Fig. 5.6.C) (T5). The system revealed that the selected cohort consistently showed higher mean ratings for both the ground-truth and predicted values compared to the rest of the patient population (T4, T6), as well as higher trajectory ratings (Fig. 5.6.D, E). An interesting pattern was observed in the Bi-LSTM trajectories across symptoms that showed increases in M12, as opposed to the rest of the population, such as taste and mucus (T6) (Fig. 5.2.D, E). The Bi-LSTM outputs showed a second peak in M6 for these symptoms, suggesting that the model, by learning from both temporal directions, detected the increases

in M12 in the ground-truth data (T5). This was not obvious in the preliminary analyses of the model results.

The clinician expressed that looking at symptom statistics for the desired cohort is what he's primarily interested in *"Summaries of chances of having anything (symptoms) over 5 (rating)"*. He also added that this activity would benefit his clinician colleagues in analyzing cohorts of interest.

5.5.3 Expert Feedback

The evaluators' feedback was extracted from meeting notes and direct written feedback.

The modelers' feedback showed that L-VISP is valuable for their research practices: *"There is so much output data generated [...](L-VISP) is instrumental in facilitating the exploration of those outputs, comparing the performance of different patient groups, and visualizing the temporal symptoms importance. In the targeted evaluation, we can use patient filters that allow for hypothesis testing. The IMV-LSTM generates summary figures [...] these vary greatly between different cohorts, and it would not be possible to identify these differences without this"* and *"I appreciate how intuitive the system can show and compare Bi-LSTM's performance among different symptoms and patient cohorts"*.

The evaluation showed that L-VISP is fit to be used for clinical research. One modeler expressed: *"It helps me a lot to analyze and understand the behavior of the Bi-LSTM. The compact, yet informative, representation of [...] allows us to see not only which variables contribute to the target symptom, but also how important one symptom contributes to all other symptoms"*. The clinician appreciated the system's ability to analyze cohorts of interest *"I am interested in seeing simplified probabilities of severity, such as toxicity at X months... (given a cohort) which this (L-VISP) supports"*.

Statements from our evaluators regarding L-VISP's actual use in practice and agreeing on hypotheses during the evaluation showed that they trust the system's results. The oncologist expressed that they were considering showing the L-VISP results to their patients and coworkers: *"I can show these to my colleagues and even my patients"*. Furthermore, the

modelers agreed during one of the case studies that *"The LSTM overpredicts for the severe patient cluster and underpredicts for the mild cluster."*.

5.6 Discussion

L-VISP highlights symptom patterns and groupings generated by LSTM modeling that extend previous research in head and neck cancer post-treatment [60,61,186,187]. Our evaluation showed that L-VISP can blend results from multiple models, enabling tasks that range from evaluating Bi-LSTM performance on patient clusters (Fig. 5.2) to visualizing hidden states from the IMV-LSTM (Fig. 5.5) for deeper model insights. L-VISP helped the modelers capture input-output relationships in the Bi-LSTM results, showing increasing trends in targeted patient cohorts with severe symptoms (Fig. 5.5). Our visual system compared performances between cohorts and revealed that the Bi-LSTM showed consistent predicted symptom clusters among cohorts (Fig. 5.2.D, E). L-VISP validated the BI-LSTMs predictions, revealing mild mispredictions for the patient clusters with severe and mild symptoms (Fig. 5.4). L-VISP was able to capture insights into the Bi-LSTM decision-making by revealing associations between symptoms during prediction (Fig. 5.5). The modelers expected to see more reliable patterns in the model's behavior, which was not the case, but were overall content with the post-treatment symptom predictions.

L-VISP was developed mainly for data modelers to create interpretable models in clinical practice. Through expert feedback and generated hypotheses, our evaluation demonstrated that modelers can effectively summarize cohort modeling results and collaborate with clinical experts to clinically interpret the models. While our case studies target head and neck cancer patients, we generalize our design to multivariate, temporal patient cohorts where the focus is to evaluate and compare different model outcomes against ground-truth data, for multiple cohorts. We generalize most of our design choices to other fields that need complex temporal prediction output interpretation in multidisciplinary collaborations with multiple workflows. Specifically, L-VISP can support other variants of the LSTM family for

temporal predictions. The ACD collaborative and iterative approach ensured that L-VISP met technical requirements for data analysis and aligned with our collaborators' workflows. The ACD design process revealed a key insight for clinician-data modeler collaborations, which was to visually separate the results presented to modelers and clinicians. However, the clinician will analyze the results together with the modeler. Specifically, we separated model debugging tasks from clinical model interpretation tasks. Below, we present a couple of lessons learned from this multidisciplinary collaboration.

Design limitations include the inability to legibly visualize data for all 28 symptoms, such as the symptom trajectories, the variable importance, and temporal importance components (i.e., would require vertical scrolling). L-VISP does not support more than two-cohort comparisons, which in turn supports legible LSTM outcome visualization. The clinical component can support a limited number of attributes and sub-cohorts/clusters (i.e., would require horizontal scrolling) in the cohort attribute distribution component. On the other hand, the patient matrix can support 2D projections based on different combinations of attributes, and a larger cohort (i.e., thousands) at the cost of limiting individual patient selection/brushing. On large cohorts with tens of thousands of patients, the system can visualize the summarized LSTM results and the three-cluster stratification. However, individual patients depicted in the projection matrix would not be legible.

Future work will address updating the current design to scale for all 28 symptom categories and for larger patient cohorts. Cluster visualization would be an option instead of visualizing individual patients and symptoms. Another natural direction would be to update the Bi-LSTM model to account for the issues found during our evaluation (e.g., the overprediction for the severe patient cluster and underprediction for the mild cluster), or to visualize other cohorts with the same data attributes. This could help the data modelers use the model on future patient cohorts.

5.6.1 Research Questions

Q1. *How can visualization support cohort analysis?* We once again employed the ACD method and interviewed data modelers in head and neck cancer research for task-gathering. We did this because this project aimed to explain black-box, LSTM-based modeling for symptom prediction in patient cohorts. We also wanted to enable clinicians to visualize the results for desired sub-cohorts. Still, we realized that some tasks are modeler-specific and are not of interest to the clinician (e.g., analyze the model's hidden states).

Q2. *How to visually represent cohorts and their characteristics, and what interactions to support?* We proposed a multiple coordinated view design, separated into multiple front-end panels. Each panel was designed for specific activities; some are dedicated to modelers alone, while others are dedicated to a clinician-modeler collaborative analysis. All panels present consistent layouts, but are dedicated to different cohort analytical workflows (i.e., analyze target cohort prediction stats, validate cluster predictions, compare model performance between clusters). This design principle enabled more guided workflows and lowered the cognitive load associated to switching between user activities. Notable cohort characteristics that this system visualizes are weighted associations between symptoms, which are characteristics that the black-box LSTM uses to make predictions.

Q3. *What system implementations work for post-treatment decision-making?* Unlike THALIS and Roses, where we tackled either the clinician's needs first or considered both clinicians' and data modelers' activities, in this project, we focused on design decisions dedicated to the data modeler. This is because L-VISP aimed to support the evaluation and explanation of black-box models, which are more complex than the modeling approaches from the previous chapters. However, we support clinician-modeler collaborative analyses and the interpretation of a simplified version of the model outcomes by clinicians.

Q4. *What makes a visual analytics system valuable to biomedical users?* We considered the differences in the mental models of our users (i.e., the clinician's interest in clear results on desired cohorts, and the data modelers' needs to understand why the models take certain

decisions and output certain predictions). As a result, our system design separates the front-ends for the data modeler, which are focused on debugging the decisions the models, from the one dedicated to the clinician, which presents the results without the underlying model mechanisms, and makes the results appropriate for clinical interpretation. A key feature of this project is its versatility to both separate workflows and support collaborative workflows, which enhances hypothesis-making.

Takeaways. A key takeaway from this project was that when working with more complex cohort modeling, such as black-box models, it is more effective to visually separate user workflows and activities based on the audience category. During the software prototyping phase, we experimented with different front-end layouts and decided to provide predefined layouts that supported workflows of interest for each type of user. This is a more guided approach than what we had in Roses. Another notable finding was to reduce visual information density when necessary, or in other words, to choose quality over quantity. L-VISP was originally designed for a large display and incorporated modeling results for all 28 symptoms. However, this resulted in high information density. As a result, we redirected our efforts into interpreting modeling results for the symptoms of main interest, namely the 9 HNC symptoms, after which the modelers stated that they could see the information more clearly and could make hypotheses faster. Last but not least, another finding was to reuse visual components for multiple activities and keep a consistent layout across workflows. Given the project requirements, we minimized variability in the visual component design, and reused components for different purposes in order to lower cognitive load for end-users with different modeling expertise (i.e., modelers vs. clinicians), and to keep consistency across activities. This resulted in a lower learning curve for data modelers and enabled the clinician to quickly interpret results.

5.7 Conclusion

L-VISP introduced a visual analytics solution that supports model activity by enhancing the interpretation of symptom risk modeling for head and neck cancer patient cohorts. Our domain characterization for black-box cohort risk modeling revealed the different user workflows in interdisciplinary clinician - data modeler collaborations. Our proposed system uses predefined layouts while blending multiple cohort modeling methods to support different symptom analytical workflows on user-defined or machine-derived cohorts. We use custom visual encodings to explain model behavior and to evaluate model performance across different cohorts, and introduce an encoding that visualizes weighted associations between multivariate, temporal items.

In the following chapter, I will discuss how OpenDBM, THALIS, Roses, and L-VISP together answer the research questions from Chapter 1, discuss the feedback from our collaborators, and present the lessons learned from all projects.

Chapter 6

Discussion and Conclusion

This dissertation addresses several challenges in visual analytics for cohort modeling for research in post-treatment care. First, it documents the domain characterization for post-treatment adverse outcome risk modeling for two applications, in oncology and neurology. This is a critical step in understanding user tasks and design requirements. Second, it provides solutions for data visualization challenges, considering scalable, custom encodings that support multi-stage, temporal, multivariate, incomplete, and associated attributes in patient cohort data. Through iterative prototyping, several new visualizations are proposed, which summarize and detect patterns in these cohorts and support post-treatment risk analysis. Third, this thesis explores the application of unsupervised risk modeling methods for patient cohorts and proposes a rule mining and clustering approach for risk prediction. Through considerate design principles, the resulting visual analytics systems present visualization solutions for model evaluation and understanding. These visualizations support human-machine cohort analysis by focusing on the interpretability and actionability of modeling results in a clinical context. Finally, since post-treatment decision-making is a multidisciplinary domain, this work tackles collaborative workflows that visually bring together different data facets for both clinician and data modeler activities.

Next, I will present how I answer the research questions of this dissertation through the contributions of the four systems.

6.1 Research Questions

Q1. How can visualization support cohort analysis? This was supported by the domain characterization for cohort analysis in the digital biomarker and head and neck cancer domains. This step was crucial to understand what type of analyses are essential for domain experts to improve post-treatment hypothesis-making (e.g., the consideration of a multi-stage patient monitoring protocol and the fact that during treatment, data can influence post-treatment longitudinal outcomes). A general concept that was essential in all projects was to visualize different data facets together, at varying levels of detail, to facilitate understanding of health distributions in cohorts.

Q2. How to visually represent cohorts, their characteristics, and what interactions to support? Designing systems with coordinated multiple views gave us the flexibility to support different analytical workflows in cohort analyses and interactively combine different attributes together, which is crucial in post-treatment decision-making. This design choice helped with the adoption of novel encodings for non-conventional data. Although most visual encodings are conventional representations of temporal, multivariate data, we introduced new encodings when the data characteristics were unconventional. Such visualizations are the filament plot and the rose glyphs, which summarize multi-stage, multivariate time series and temporal networks and highlight patterns and associations between attributes.

Q3. What system implementations work for cohort stratification? Considering the importance of cohort modeling and XAI in post-treatment decision-making, it was critical to focus on system implementations that support human-machine workflows, more specifically, visual analytics for human interpretation of cohort modeling (machine-derived results). The ACD iterative and incremental design and development on each system, with regular feedback sessions, enhanced the interpretation, evaluation, and understanding of different methods for cohort modeling, and the analysis of said the results and methods in conjunction with relevant patient attributes.

Q4. What makes a visual analytics system valuable for biomedical users? The systems' evaluation with domain experts revealed that an essential consideration for post-treatment cohort research was to enable clinician-data modeler collaborations. This was supported by visual analytics that focused on interactive modeling activity for data modelers and appropriate visualizations for the clinical interpretation of the modeling results for clinicians. OpenDBM explored solutions that target large audiences (i.e., open source, which can include academics, clinicians, industry researchers, and so on); THALIS focused on visualizations useful for both clinicians and data modelers, but not on a configurable front-end; Roses, on the other hand, explored a highly configurable design for domain experts. In contrast, L-VISP focused on data modeler activities and separated the front-ends between clinicians and modelers for a more guided analysis.

In the following sections, I present the conclusions resulting from the systems' evaluations (i.e., resulting dimensions), then I move to the overall lessons learned from these projects, and then I touch on the generalizability, limitations, and future work of this dissertation.

6.2 Thematic Analysis

We performed a reflexive thematic analysis on the feedback from domain experts from all projects, and the results were coded into three dimensions. We investigated actionability, a concept often evaluated in medical visual analytics (VA) systems [33, 97], which helped us to understand whether domain experts think the systems are fit to be used in clinical practice and if they do the work they need to do (i.e., support the tasks and activities collected during the domain characterization process). A second dimension we looked at was perceived usefulness. Visual analytics systems usually measure usability, which focuses on how easy and efficient a system is to use. In our case, one of this dissertation's goals is to propose novel visual encodings for complex data, which can result in a slower adoption of new visualizations, and the users' need to get accommodated with "change". In exchange, we use perceived usefulness as an evaluated dimension, focusing on how much value the

domain experts believed that the systems provide and whether they think they would benefit from using the proposed systems in practice. Lastly, as trust is an ongoing challenge in the visualization community [79, 114], we also looked at this dimension. This helped us understand whether domain experts trust the systems enough to actually use and consult them during decision-making. Although recent decades have seen an expansion in medical visual analytics systems, especially XAI systems, medical experts still choose to trust their instinct over what a visual analytics system reports. With medicine being such a high-stakes domain, experts are right to show reluctance to new software; however, we were interested to see if they show confidence that the systems are dependable and trustworthy. Below, we present what facilitated the evaluation of these dimensions and provide some example feedback used to extract them.

Actionability. What helped: gradually presenting results and gradually expanding results from one project to the next (for example, the rule mining modeling, which was expanded from THALIS to Roses, then the clustering results from Roses influenced the modeling method from the future work project), and the occasional domain experts' interactions with the system during in-person campus visits. Clinical practitioners have noted that Roses and L-VISP were useful for cohort statistics and modeling, but not for in-clinic use by clinicians without modeling experience. This was not the case with THALIS, as it is a tool that supports more general cohort analyses, as a clinician noted that they'd show certain results to their patients. OpenDBM, on the other hand, was designed for open-source. Example feedback:

- *"I had someone looking away from the camera, this is actually picking up their data"* - noted by a domain expert about OpenDBM, while analyzing an individual's video data, noting the importance of knowing when patients don't look at the camera when recorded, and thus useful biomarker measurements are not extracted for analyses. They were able to spot when the patient was not facing the camera by combining the facial map (Fig. 6.1.A) with the facial measurements for eyes (Fig. 6.1.B).

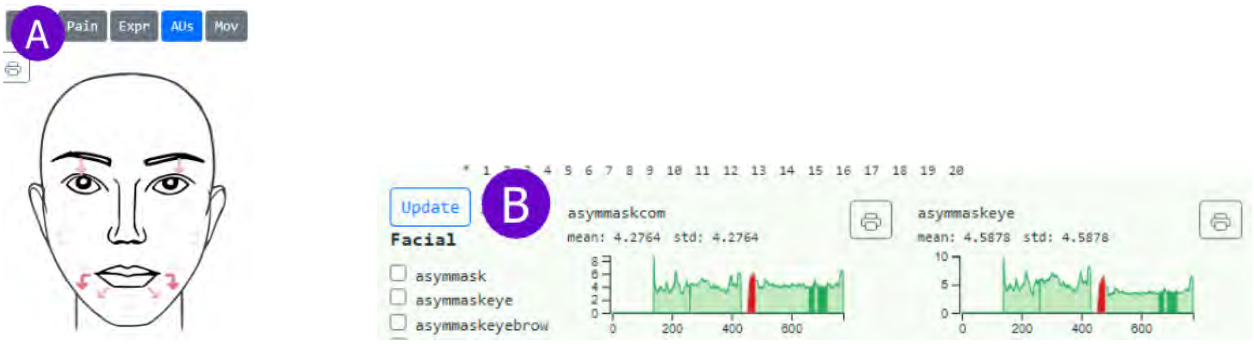

Figure 6.1: OpenDBM Combination of facial map (A) with facial measurements(B)

- *"I say – we're going to talk about dry mouth and swallowing, cause these two are really bad– and then there's all the other stuff. And then I see this [the ARM and heatmap and filaments], and here's this other stuff, that is usually at my periphery, but I don't focus on, although patients do mention it. If I were sitting with a patient and I'd look at this interface and ARMs—I get it, hey, there's actually a LOT of moving parts here [beyond dry mouth and swallowing], and they're related, and they have different time sources. It's sobering."* - noted by a clinical practitioner when combining patterns from the rule mining visualization (Fig. 6.2.A), together with the filament plots (Fig. 6.2.B); which showed understudied patterns in symptom trajectories and associations (clusters). In this case, they were surprised by how many strong associations drowsiness has with other symptoms (e.g., fatigue, daily activities).

- *"When I see a patient, this [taste-dry mouth] association in the late phase is the default picture I have in my mind. But here I see that also fatigue connects to drowsiness, and that these symptoms show up in the acute phase as well, and that I really need to discuss these issues with my patients."* - noted by the previous clinical practitioner during the evaluation of the symptom clusters in THALIS for the late stage (Fig. 6.2.A), as seen in the cluster with drowsiness at the center.

- *"This interface and the ARM provide great preliminary data for so many projects right off the bat!"* (Fig. 6.2.A)

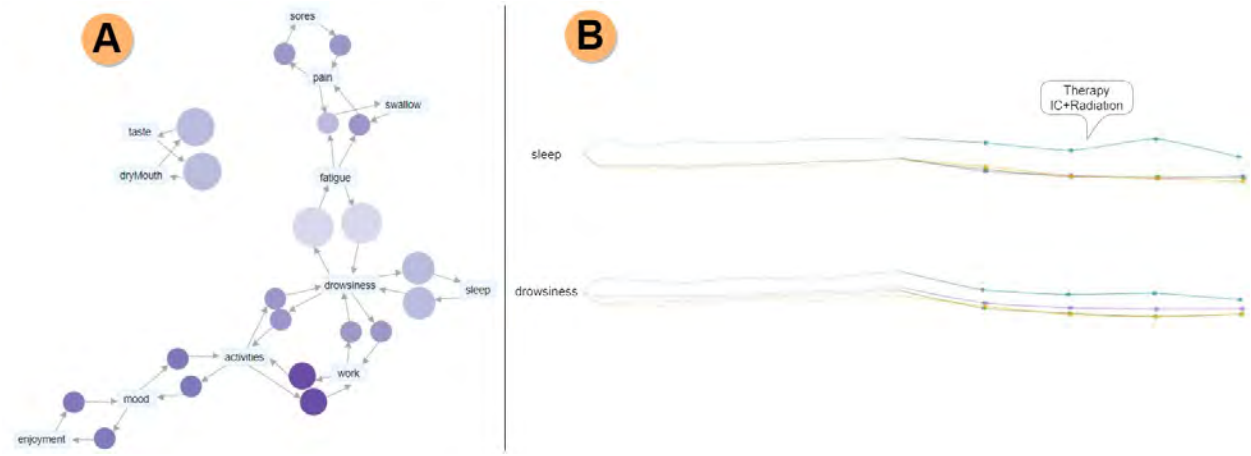

Figure 6.2: THALIS post-treatment symptom clusters (A) and trajectories (B)

- *"It helps me a lot to analyze and understand the behavior of the Bi-LSTM. The compact, yet informative, representation of [...] allows us to see not only which variables contribute to the target symptom, but also how important one symptom contributes to all other symptoms"* commented one modeler during the L-VISP evaluation while analyzing the variable importance view (Fig. 6.3)

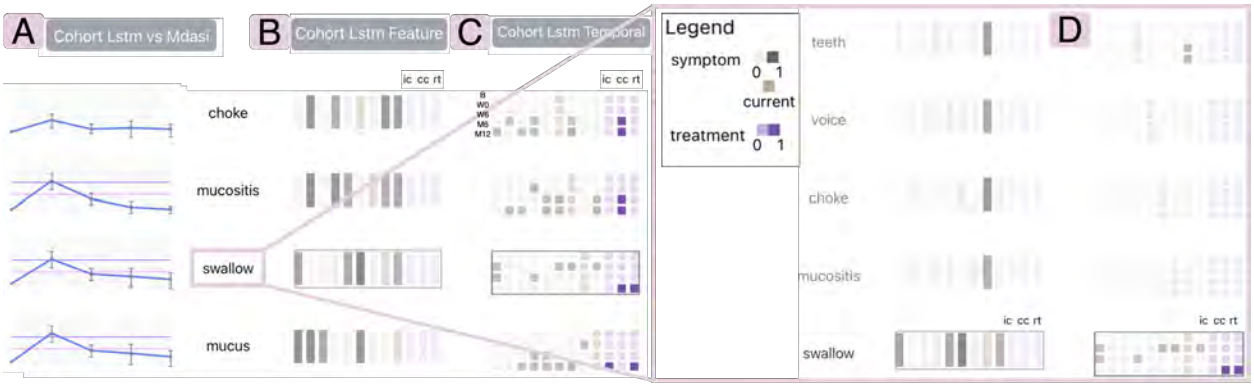

Figure 6.3: L-VISP symptom weighted associations to a selected symptom. A) symptom predictions against ground-truth. B) variable importance. C) temporal importance. D) Upon selecting swallowing, its corresponding column shows the influence of the other symptoms on the swallow prediction, and the highlighted items on its corresponding row show the influence of swallow to the other symptoms' predictions.

Perceived Usefulness. What helped: case studies, walkthroughs, and hypothesis generation during these cases. Example feedback:

- *"The utility for exploring the data here is very high, if you're talking about quantitative*

decision-making, this is very strong." - one clinical practitioner commented about Roses and its ability to adapt to different analytical tasks through its configurable front-end – namely the five visualization options for the four quadrants of the interface, which can be configured for the four treatment options – (Fig. 6.4). He also noted that the system is not suitable for clinicians without modeling experience.

- *"Very cool, so much better to use for the analysis we did last year, huge time saver"* - noted by a data modeler about the rose glyph projection in Roses (Fig. 6.4), which showed more comprehensive modeling results, which were time-based (treatment stage-based) associations and patterns between symptoms for each treatment (i.e., multiple sub-cohorts), as opposed to THALIS ("last year"), which showed limited (top 20 rules), within stage (non-temporal) results/rules for the entire cohort, without considering the treatment influence over the patient outcomes.

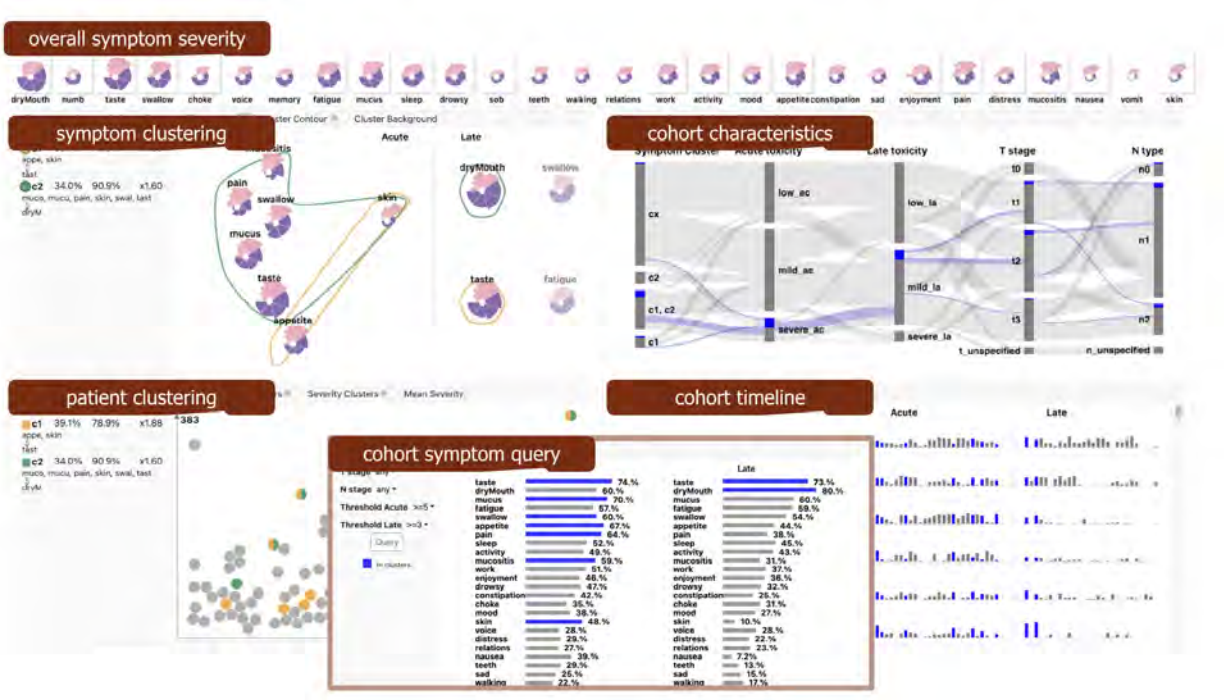

Figure 6.4: Roses five encoding options for configurable workflows

- *"I gotta be honest, [...] I get so much material for future research." "I like that when a patient is with [oncologist], they want percentages, e.g., 69% of patients have normal*

mucus after 1 week of treatment, and [THALIS] shows that" - declared by a clinician when looking at the statistics for mucus burden for the selected time point (1 week post-treatment)(Fig. 6.5).

- *"There is so much output data generated [...](L-VISP) is instrumental in facilitating the exploration of those outputs, comparing the performance of different patient groups, and visualizing the temporal symptoms importance. In the targeted evaluation, we can use patient filters that allow for hypothesis testing. The IMV-LSTM generates summary figures [...] these vary greatly between different cohorts, and it would not be possible to identify these differences without this"* mentioned a modeler about how valuable they find L-VISP for their research practices (Fig. 6.3)

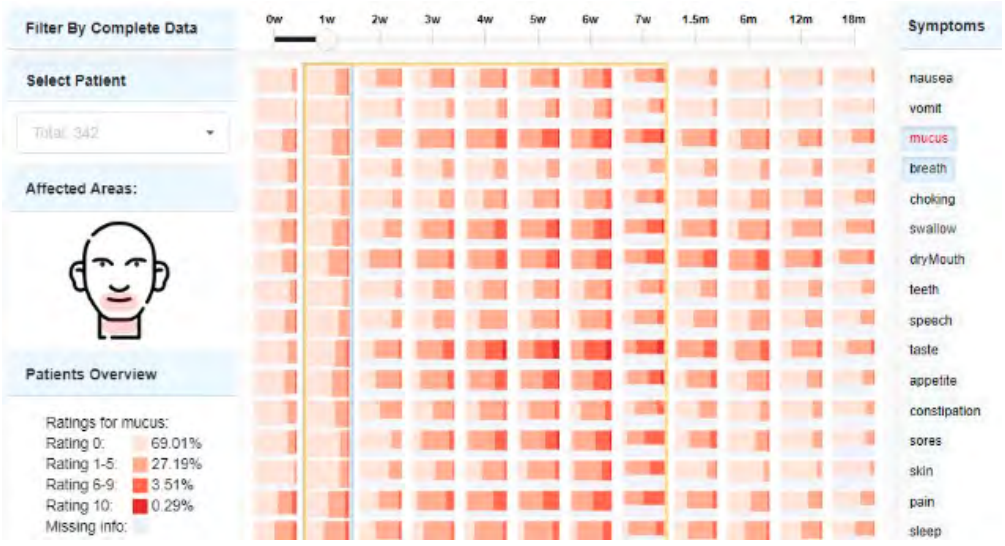

Figure 6.5: THALIS cohort longitudinal burden of a selected symptom (mucus) on the bottom left, and the severity and presence symptom distribution for the entire cohort.

Trust. What helped: showcasing results that confirm existing user knowledge (e.g., rose glyphs showed high mood-related symptoms at the beginning of treatment due to low patient morale, which was a known treatment consequence). One domain expert observed during an in-person demo session, using the rose glyphs, that the dataset was not mature due to missing data points (i.e., the cohort was under monitoring at that time and as time passed,

researchers had updated the dataset with more data points) and observed an example of wrong data summarization because specific time points in the roses showed higher ratings than they were supposed to. Example feedback:

- *"I expected that the more advanced staging you have, the more toxicity you get - it corrected my assumptions."* - one clinical practitioner noted when evaluating the Sankey-based encodings from Roses (Fig. 6.6), observing that certain cohort attributes, such as tumor size, are not necessarily associated with more severe symptoms.

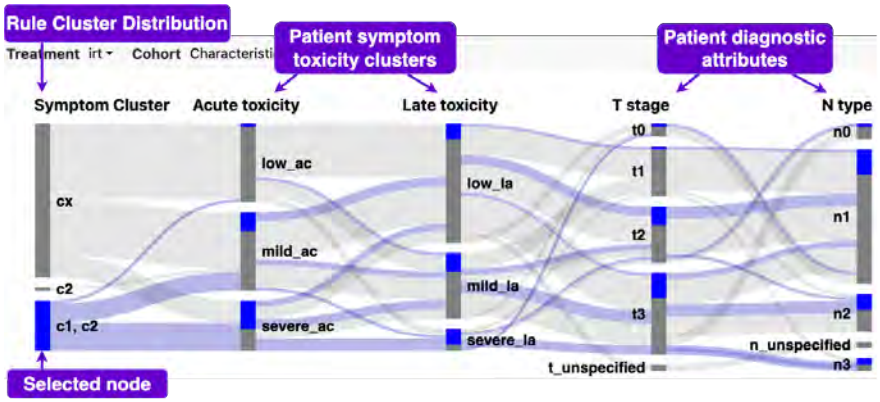

Figure 6.6: Roses Cohort attribute distribution

- *"When I see a patient, this [. . .] association in the late phase is the default picture I have in my mind. But here I see that also [. . .], and that these symptoms show up in the acute phase as well, and that I really need to discuss these issues with my patients." "I can share [this view] with my patients, to explain that pain and swallowing and fatigue are really tightly related – we don't know if it's causation, but they definitely show up together, so could you please, please, take your pain and anti-inflammatory meds, and could you please do the swallowing exercises we've talked about?"*- noted by a clinical practitioner when analyzing the rule node-link from THALIS, looking at the late symptom clusters (Fig. 6.2).

- *"I can show these to my colleagues and even my patients"* mentioned a clinician during the L-VISP evaluation, while the modelers trusted the results presented by the models:

"The LSTM overpredicts for the severe patient cluster and underpredicts for the mild cluster." (Fig. 6.7)

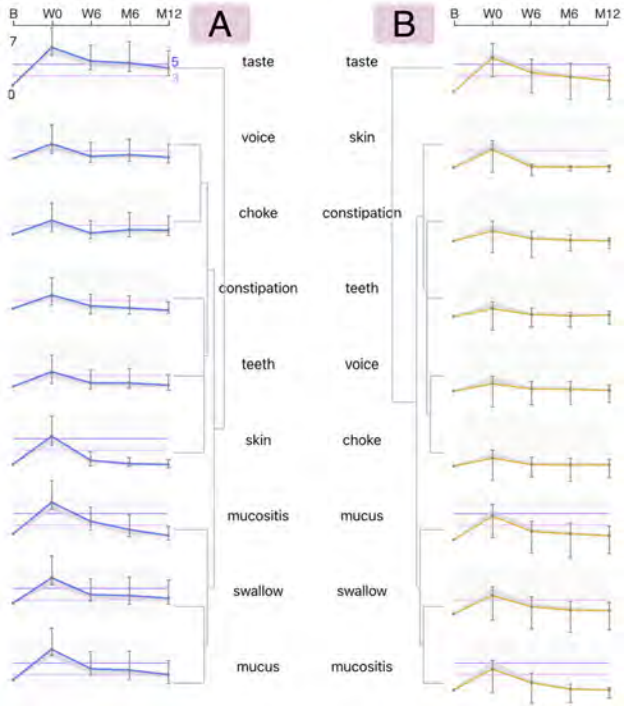

Figure 6.7: L-VISP predictions and errors for two patient clusters: severe symptom burden in A) and mild symptom burden in B)

6.3 Lessons Learned

Use the activity-centered paradigm in multidisciplinary collaborations. The ACD paradigm was used in all projects presented, helping to focus on user activities and tasks when designing visual analytics tools. It enabled a better understanding of the differences in mental models of the clinician and the modeler, as well as the differences in their analytical interests and tasks. In addition, it was a good strategy for multidisciplinary remote collaborations, providing steady progress and a structured design and development of the systems. In particular, unlike oncological applications, which were designed for a small, targeted audience of data modelers and clinicians, the neurology application project (OpenDBM) was designed for the open-source community. Fortunately, we were able to interview domain experts

representing various institutions and with different backgrounds, from clinicians to academics and industry data scientists. This resulted in many different interests and tasks for the project, but the ACD approach supported the prioritization of overlapping tasks, which resulted in a successful visualization tool. Another benefit of the ACD methodology was the incremental and iterative design process, which resulted in useful designs due to regular feedback sessions from the clients. In the end, using the same paradigm for all systems provided good results and positive feedback on the final systems.

Focus on domain sense and actionability. When designing visual analytics systems that focus on interpreting modeling results, usually referred to as XAI applications, it is important to focus on the application domain. More specifically, on how to make modeling results useful in a clinical setting. One way we supported actionability and domain sense was through custom visualizations. Some examples are: mapping patient-modeled multivariate spatial and non-spatial measurements on anatomical locations for a more approachable behavior interpretation for clinicians, in an attempt to humanize the data; providing additional context (i.e., adding clinician-targeted visualizations) for a more transparent data interpretation by clinicians; incorporating client knowledge into the systems for more trust in the modeled results (i.e., showing associations between specific symptoms from the literature to introduce new, more complex symptom associations). Another way to support actionability and domain sense was to focus on interactions and data analysis common to risk analysis. Some examples were patient against cohort analysis, which is available in all proposed systems, stratification of the patient cohort using user-selected attributes (present in OpenDBM, THALIS, and L-VISP), and connecting different data facets for a comprehensive outcome risk composition analysis.

Use visual scaffolding to introduce novel encodings. Although the proposed systems use custom visualizations and coordinated multiple views, which are critical to connect different data facets for comprehensive cohort analyses, each project introduces a novel encoding (i.e., the anatomical mask, the filament plot, the rose glyph, and the variable im-

portance matrix). These novel encodings, although they sometimes attracted reticence from some users, ended up being crucial components in describing complex concepts in analytical workflows. This was due to the presentation of client knowledge through the encodings, which helped with faster adoption of these designs, and due to the repeated use of these encodings when presenting project updates, which was a consequence of the ACD iterative process. The ACD incremental design process, which started with paper prototypes and then moved to digital prototypes, also helped with the introduction of these encodings to clients. In general, a balanced blend of standard encodings, such as scatterplots and barplots, and custom encodings helped with the adoption of novel visual representations.

6.4 Generalizability

Although the proposed visualization systems are designed for particular clients, tasks, and datasets, many concepts can be generalized to other visual analytics applications. First, these systems visualize cohorts that are heterogeneous, temporal, and multivariate, with most visual encodings suitable for other research domains (e.g., the matrix-based encoding in THALIS for multivariate, temporal distributions in cohorts with missing data points, Sankey-based encoding for multivariate cohort summarization, interactive scatterplots for item against cohort or cluster analysis, correlation matrix for multivariate attributes, etc.). Generalizable analytical tasks that these systems support are: association and pattern detection, interpretation of model output, and risk prediction, which can be found in other domains such as sports [65], (historical) documents [203], weather [173], financial [109], and visual analytics, etc. The novel encodings can also be applied in other domains. The anatomical representation from THALIS inspired the facial mapping encoding used in OpenDBM; in both cases, there was a need to provide context for multivariate attributes and map them to spatial locations. These types of representations can be reused in other medical applications. The filament plot summarizes temporal attributes and can be used in other applications where comparing and summarizing time series is a crucial task. An example is

an aircraft simulation publication [151] where we reused the filament plot to represent I/O parameters of aircraft simulation ensemble members. The rose glyph was essential in the representation of temporal networks with temporal nodes and was used to show associations and trajectory similarity in multivariate attributes. The rose glyph projection, representing temporal networks, is a contribution to summarizing and clustering associated, overlapping time series. By itself, the glyph can be used as a compact temporal representation, where comparing large sets of temporal items (n >20) is a key task. Another proposed encoding, the variable importance matrix, is used to expose LSTM-based model memories and can be used in other projects that use LSTMs. Moreover, this encoding supports the visualization of weighted associations between temporal and non-temporal items.

6.5 Limitations

The limitations of this work include items related to domain-specific applications. Although the design lessons learned from each project can be applied to other medical visualization systems, most of this work is aimed at clients with modeling experience, from data scientists to clinicians. Considering the limited number of users for most of the presented projects, the proposed work is evaluated using qualitative methods, through case studies, demonstrations, client observations, and direct feedback, and lacks quantitative evaluation. Moreover, since most of this work is focused on relatively new domains in clinical research, the designs presented focus on the actionability of the model over simplicity. In particular, there are limitations with respect to the availability of patient data that was used to develop and evaluate these projects. Considering the limited samples of patients with particular attributes for some of these projects (e.g., small samples of patients with specific treatment plans) and missing data points, the modeling results can present biases. Further data limitations include that one of the projects used synthetic data during the development and evaluation stages, while the other projects used manually extracted data from patient health records, which can be erroneous.

6.6 Future Work

Several extensions could be explored with some of this thesis's projects. One extension would be to improve the scalability of the L-VISP interface and support the analysis of more symptoms. This would be solved through using symptom clusters for symptoms with similar trajectories, instead of individual symptoms. Another future direction would be to support a guided exploration of the configurable interface from Roses. This could be enhanced through predefined workflows that could be selected from a menu, similar to L-VISP. Natural language queries could be used to faster generate desired workflows for clinicians or data modelers. The queries would configure the front-end with appropriate visual components (e.g., "Show me LSTM predictions"), on desired cohorts (e.g., "for female patients that had neck surgery"). This last extension would work well for L-VISP and Roses, which support multiple analytical workflows. It would help modelers to debug cohort modeling faster and clinicians to collaborate and find patients with similar characteristics to assess the risk for new patients.

There have been tens of visual analytics systems proposed throughout the years in the medical domain. In consequence, another direction would be to look into design standards for medical visual analytics applications in multidisciplinary clinician-modeler collaborations. What is a good design balance? Should we focus on one type of audience or both? A good step towards this direction would be to revise Guo's survey [79] that categorizes medical visual analytics into cohort, outcome, and prognosis analysis, and connect these to each type of audience and their corresponding tasks. Another relevant and newly proposed framework by Bernard [18] characterizes how humans, data, and models contribute and benefit from visual analytics processes. Combining the two studies and clearly characterizing each type of audience and how they play into medical visual analytics would help to create some standards for designing such systems.

Given the fast rise of Generative AI, a recent study from Monadjemi et al. [143] is proposing an updated theoretical framework for mixed-initiative (MI), human-machine analysis.

They pose very relevant questions for this dissertation, namely: how do artificial agents and humans contribute to MI visual analytics tasks? What are the characteristics of the MI tasks and the visual analytics environment? Although AI agents for medical visual analytics are out of scope in this dissertation, it is a concept that is rising fast and will shape the future of medical visual analytics. As for the MI framework, this dissertation has focused on less automated frameworks, where human agents (users) play a detrimental role in knowledge-centric tasks. However, the survey states that *mixed-initiative systems improve accuracy in comparison to either a human or an artificial agent performing the task alone.* In consequence, this survey is an excellent start for the new directions that can be explored in MI visual analytics for medical applications.

6.7 Conclusion

In conclusion, this thesis presented visual analytics methods that can facilitate heterogeneous cohort analysis for post-treatment care. This work documented the design, development, and evaluation of several visual analytics systems for cohort analysis. It did that by establishing the domain requirements for risk modeling in two medical applications, in neurology and oncology, then designing visualization systems that blend risk modeling with custom visualizations for patient cohorts, and by evaluating the usefulness of these systems for clinical research. The proposed visual analytics systems tackled data modeling and visualization challenges, supporting human-machine analysis in clinician-data modeler collaborations. This dissertation introduced novel encodings for multivariate, temporal, multi-stage patient cohorts, with unconventional characteristics such as multi-stage, associated spatial and non-spatial measurements, and with weighted associations, and adapts and proposes alternative, rule mining-based, outcome risk modeling approaches. The evaluated dimensions and the lessons learned presented important considerations for future research in visual analytics designed for multidisciplinary collaborations. More specifically, this research provided a deeper understanding of key considerations when doing research with domain experts from an ap-

plication domain (doctors) and data modelers, and on how data visualization designers can present the results generated by data modelers. Data visualization helped patient care research by finding and documenting more consistent patterns in patient cohorts. Many of the research lessons and contributions could be expanded to other application domains, not just medicine. Returning to this dissertation research statement, I conclude that this dissertation presents contributions that show that visual analytics can improve post-treatment research.

6.8 Appendix: Copyright Permissions

arXiv.org - Non-exclusive license to distribute

The URI http://arxiv.org/licenses/nonexclusive-distrib/1.0/ is used to record the fact that the submitter granted the following license to arXiv.org on submission of an article:

- I grant arXiv.org a perpetual, non-exclusive license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

Revision history

2004-01-16 - License above introduced as part of arXiv submission process
2007-06-21 - This HTML page created

Contact

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy
Carla Floricel , Nafiul Nipu , Mikayla Biggs , Andrew Wentzel , Guadalupe Canahuate , Lisanne Van Dijk , Abdallah Mohamed , C.David Fuller , G.Elisabeta Marai
Transactions on Visualization and Computer Graphics

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Carla Floricel — **Signature**

29-09-2021 — **Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

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

Carla Floricel , Andrew Wentzel , Abdallah Mohamed , C.David Fuller , Guadalupe Canahuate , G.Elisabeta Marai

Transactions on Visualization and Computer Graphics

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.

2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.

4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.

5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Carla Floricel — Signature

16-11-2023 — Date (dd-mm-yyyy)

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own

Cited Literature

[1] Digital biomarkers: Global markets (2022). https://www.bccresearch.com/market-research/biotechnology/digital-biomarkers-market.html.

[2] S. S. Abdullah, N. Rostamzadeh, K. Sedig, A. X. Garg, and E. McArthur. Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. *Informatics*, 7(2):17, 2020.

[3] E. Abel, E. Silander, J. Nyman, T. Björk-Eriksson, and E. Hammerlid. Long-Term Aspects of Quality of Life in Head and Neck Cancer Patients Treated with Intensity Modulated Radiation Therapy: A 5-Year Longitudinal Follow-up and Comparison with a Normal Population Cohort. *Advances in Radiation Oncology*, 5(1):101–110, 2020.

[4] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[5] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, p. 487–499. Morgan Kaufmann Publishers Inc., 1994.

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

[7] A. Aktas, D. Walsh, and L. Rybicki. Symptom clusters: Myth or Reality? *Palliative Medicine*, 24(4):373–385, 2010.

[8] P. Angelelli, S. Oeltze, J. Haász, C. Turkay, et al. Interactive Visual Analysis of Heterogeneous Cohort-Study Data. *IEEE Computer Graphics and Applications*, 34(5):70–82, 2014.

[9] M. Y. Ansari, A. Ahmad, S. S. Khan, G. Bhushan, et al. Spatiotemporal Clustering: A Review. *Artificial Intelligence Review*, 53:1–43, 2019.

[10] D. Antweiler and G. Fuchs. Visualizing Rule-based Classifiers for Clinical Risk Prognosis. *IEEE Transactions on Visualization and Computer Graphics*, pp. 55–59, 2022.

[11] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. *Hawaii International Conference on System Sciences*, 44:1–10, 2011.

[12] L. M. Babrak, J. Menetski, M. Rebhan, G. Nisato, M. Zinggeler, N. Brasier, K. Baerenfaller, T. Brenzikofer, L. Baltzer, C. Vogler, et al. Traditional and Digital Biomarkers: Two Worlds Apart? *Digital Biomarkers*, 3(2):92–102, 2019.

[13] L. Bartram and M. Yao. Animating Causal Overlays. In *Computer Graphics Forum*, vol. 27, p. 751–758. Wiley Online Library, 2008.

[14] T. Baumgartl, M. Petzold, M. Wunderlich, M. Hohn, et al. In Search of Patient Zero: Visual Analytics of Pathogen Transmission Pathways in Hospitals. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):711–721, 2021.

[15] M. S. Benda and K. S. Scherf. The Complex Emotion Expression Database: A Validated Stimulus Set of Trained Actors. *PloS one*, 15(2):e0228248, 2020.

[16] J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*, pp. 1–4. Springer, 2009.

[17] J. Benoit, H. Onyeaka, M. Keshavan, and J. Torous. Systematic Review of Digital Phenotyping and Machine Learning in Psychosis Spectrum Illnesses. *Harvard Review of Psychiatry*, 28(5):296–304, 2020.

[18] J. Bernard. The human-data-model interaction canvas for visual analytics. *arXiv preprint arXiv:2505.07534*, 2025.

[19] J. Bernard. VIVA: Virtual Healthcare Interactions Using Visual Analytics, With Controllability Through Configuration. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–18, 2025.

[20] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer. A Visual-interactive System for Prostate Cancer Stratifications. *Procceedings of IEEE VIS Workshop Visualizing Electronic Health Record Data*, 10, 2014.

[21] M. Biggs, C. Floricel, L. Van Dijk, A. S. Mohamed, C. David Fuller, G. E. Marai, X. Zhang, and G. Canahuate. Identifying Symptom Clusters from Patient Reported Outcomes Through Association Rule Mining. *International Conference on Artificial Intelligence in Medicine*, pp. 491–496, 2021.

[22] A. Bonifati, F. Del Buono, F. Guerra, and D. Tiano. Time2Feat: Learning Interpretable Representations for Multivariate Time Series Clustering. *Proceedings of the VLDB Endowment*, 16(2):193–201, 2022.

[23] M. Bostock, V. Ogievetsky, and J. Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[24] L. Brasseur. Florence Nightingale's Visual Rhetoric in the Rose Diagrams. *Technical Communication Quarterly*, 14(2):161–182, 2005.

[25] V. Braun and V. Clarke. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1):37–47, 2021.

[26] I. Brook. Late Side Effects of Radiation Treatment for Head and Neck Cancer. *Radiation Oncology Journal*, 38(2):84, 2020.

[27] D. Bruzzese and C. Davino. Visual Mining of Association Rules. In *Visual Data Mining*, p. 103–122. Springer, 2008.

[28] H. S. G. Caballero, A. Corvò, P. M. Dixit, and M. A. Westenberg. Visual Analytics for Evaluating Clinical Pathways. *IEEE Workshop on Visual Analytics in Healthcare*, pp. 39–46, 2017.

[29] Q. Cai, K. Zheng, H. Jagadish, B. C. Ooi, and J. Yip. CohortNet: Empowering Cohort Discovery for Interpretable Healthcare Analytics. *Proceedings of the VLDB Endowment*, 17(10):2487–2500, 2024.

[30] A. Cao, X. Xie, M. Zhou, H. Zhang, M. Xu, and Y. Wu. Action-Evaluator: A Visualization Approach for Player Action Evaluation in Soccer. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[31] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[32] B. C. Cappers and J. J. van Wijk. Exploring Multivariate Event Sequences Using Rules, Aggregations, and Selections. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):532–541, 2017.

[33] R. Cava, C. M. D. S. Freitas, and M. Winckler. ClusterVis: Visualizing Nodes Attributes in Multivariate Graphs. *Proceedings of the Symposium on Applied Computing*, p. 174–179, 2017.

[34] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided Visual Clustering Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, 2019.

[35] A. S. Chandrabhatla, I. J. Pomeraniec, and A. Ksendzovsky. Co-Evolution of Machine Learning and Digital Technologies to Improve Monitoring of Parkinson's Disease Motor Symptoms. *NPJ Digital Medicine*, 5(1):1–18, 2022.

[36] L. Chen et al. FSLens: A Visual Analytics Approach to Evaluating and Optimizing the Spatial Layout of Fire Stations. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[37] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang. Py-feat: Python Facial Expression Analysis Toolbox. *Affective Science*, 4(4):781–796, 2023.

[38] K. M. Christopherson, A. Ghosh, A. S. R. Mohamed, M. Kamal, et al. Chronic Radiation-Associated Dysphagia in Oropharyngeal Cancer Survivors: Towards Age-Adjusted Dose Constraints for Deglutitive Muscles. *Clinical and Translational Radiation Oncology*, 18:16–22, 2019.

[39] K. K. Chui, J. B. Wenger, S. A. Cohen, and E. N. Naumova. Visual Analytics for Epidemiologists: Understanding the Interactions Between Age, Time, and Disease with Multi-Panel Graphs. *PLoS one*, 6(2):1–8, 2011.

[40] C. S. Cleeland, T. R. Mendoza, X. S. Wang, C. Chou, M. T. Harle, M. Morrissey, and M. C. Engstrom. Assessing Symptom Distress in Cancer Patients: The M.D. Anderson Symptom Inventory. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 89:1634–46, 11 2000.

[41] S. Coghlan and S. D'Alfonso. Digital Phenotyping: an Epistemic and Methodological Analysis. *Philosophy and Technology*, 34(4):1905–1928, 2021.

[42] J. D. Cox and K. K. Ang. *Radiation oncology E-book: rationale, technique, results*. Elsevier Health Sciences, 2009.

[43] P. Crits-Christoph, M. B. C. Gibbons, S. Ring-Kurtz, R. Gallop, S. Stirman, J. Present, C. Temes, and L. Goldstein. Changes in Positive Quality of Life Over the Course of Psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 45(4):419, 2008.

[44] V. De Angel, S. Lewis, K. White, C. Oetzmann, D. Leightley, E. Oprea, G. Lavelle, F. Matcham, A. Pace, D. C. Mohr, et al. Digital Health Tools for the Passive Monitoring of Depression: A Systematic Review of Methods. *NPJ Digital Medicine*, 5(1):1–14, 2022.

[45] D. Defays. An Efficient Algorithm for a Complete Link Method. *The Computer Journal*, 20(4):364–366, 1977.

[46] J. Deogun and L. Jiang. Prediction Mining–an Approach to Mining Association Rules for Prediction. *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 98–108, 2005.

[47] S. Di Bartolomeo, Y. Zhang, F. Sheng, and C. Dunne. Sequence Braiding: Visual Overviews of Temporal Event Sequences and Attributes. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1353–1363, 2020.

[48] D. Dingen, M. van't Veer, P. Houthuizen, E. H. Mestrom, E. H. Korsten, A. R. Bouwman, and J. Van Wijk. RegressionExplorer: Interactive Exploration of Logistic Regression Models With Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[49] N. Donald. The Design of Everyday Things: Revised and Expanded Edition, 2013.

[50] S. T. Dong, D. S. Costa, P. N. Butow, M. R. Lovell, M. Agar, G. Velikova, P. Teckle, A. Tong, N. C. Tebbutt, S. J. Clarke, et al. Symptom Clusters in Advanced Cancer Patients: An Empirical Comparison of Statistical Methods and the Impact on Quality of Life. *Journal of Pain and Symptom Management*, 51(1):88–98, 2016.

[51] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-Efficacy. *Transactions on Computer-Humanan Interaction*, 17(4), 24 pages, 2011.

[52] F. Du, C. Plaisant, N. Spring, and B. Shneiderman. EventAction: Visual Analytics for Temporal Event Sequence Recommendation. *IEEE Symposium on Visual Analytics Science and Technology*, p. 61–70, 2016.

[53] A. K. Dubey, U. Gupta, and S. Jain. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *International Journal of Computer Assisted Radiology and Surgery*, 11(11):2033–2047, 2016.

[54] P. Ekman and W. V. Friesen. Facial Action Coding System. *Environmental Psychology & Nonverbal Behavior*, 1978.

[55] M. Elshehaly, K. Sohal, et al. Creative Visualisation Opportunities Workshops: A Case Study in Population Health. *Evaluation and Beyond-Methodological Approaches for Vis.*, pp. 11–19, 2022.

[56] S. A. Eraj, M. K. Jomaa, C. D. Rock, A. S. Mohamed, B. D. Smith, J. B. Smith, T. Browne, L. C. Cooksey, B. Williams, et al. Long-Term Patient Reported Outcomes Following Radiation Therapy for Oropharyngeal Cancer. *Rad. Onco.*, 12(1):150, 2017.

[57] G. Fan, L. Filipczak, and E. Chow. Symptom Clusters in Cancer Patients: A Review of the Literature. *Current Oncology*, 14(5):173–179, 2007.

[58] K. A. Ferreira, M. Kimura, M. J. Teixeira, T. R. Mendoza, J. C. M. da Nóbrega, S. R. Graziani, and T. Y. Takagaki. Impact of Cancer-Related Symptom Synergisms on Health-Related Quality of Life and Performance Status. *Journal of Pain and Symptom Management*, 35(6):604–616, 2008.

[59] 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.

[60] C. Floricel, N. Nipu, M. Biggs, A. Wentzel, G. Canahuate, L. Van Dijk, A. Mohamed, C. D. Fuller, and G. E. Marai. THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):151–161, 2021.

[61] C. Floricel, A. Wentzel, A. Mohamed, C. D. Fuller, G. Canahuate, and G. E. Marai. Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[62] A. G. Forbes, A. Burks, K. Lee, X. Li, P. Boutillier, J. Krivine, and W. Fontana. Dynamic Influence Networks for Rule-Based Models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):184–194, 2017.

[63] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo. CMRules: Mining Sequential Rules Common to Several Sequences. *Knowledge-Based Systems*, 25(1):63–76, 2012.

[64] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, V. S. Tseng, et al. Spmf: a java open-source pattern mining library. *Journal of Machine Learning Research*, 15(1):3389–3393, 2014.

[65] Y. Fu and J. Stasko. Hoopinsight: Analyzing and Comparing Basketball Shooting Performance Through Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):858–868, 2023.

[66] K. Furmanová, N. Grossmann, L. P. Muren, O. Casares-Magaz, V. Moiseenko, J. P. Einck, M. E. Gröller, and R. G. Raidou. VAPOR: Visual Analytics for the Exploration of Pelvic Organ Variability in Radiotherapy. *Computers and Graphics*, 91:25–38, 2020.

[67] K. Furmanová, L. P. Muren, O. Casares-Magaz, V. Moiseenko, J. P. Einck, S. Pilskog, and R. G. Raidou. PREVIS: Predictive Visual Analytics of Anatomical Variability for Radiotherapy Decision Support. *Computers and Graphics*, 97:126–138, 2021.

[68] I. Galatzer-Levy, A. Abbas, A. Ries, S. Homan, L. Sels, V. Koesmahargyo, V. Yadav, M. Colla, H. Scheerer, S. Vetter, et al. Validation of Visual and Auditory Digital Markers of Suicidality in Acutely Suicidal Psychiatric in Patients: Proof-of-Concept Study. *Journal of Medical Internet Research*, 23(6):e25199, 2021.

[69] I. R. Galatzer-Levy, A. Abbas, V. Koesmahargyo, V. Yadav, M. M. Perez-Rodriguez, P. Rosenfield, O. Patil, M. F. Dockendorf, M. Moyer, L. A. Shipley, et al. Facial and Vocal Markers of Schizophrenia Measured Using Remote Smartphone Assessments. *medRxiv*, 2020.

[70] M. Geiger and O. Wilhelm. Computerized Facial Emotion Expression Recognition. *Digital Phenotyping and Mobile Sensing*, pp. 43–56, 2023.

[71] N. R. Giuliani, J. C. Flournoy, E. J. Ivie, A. Von Hippel, and J. H. Pfeifer. Presentation and Validation of the DuckEES Child and Adolescent Dynamic Facial Expressions Stimulus Set. *International Journal of Methods in Psychiatric Research*, 26(1):e1553, 2017.

[72] J. C. Goldsack, A. Coravos, J. P. Bakker, B. Bent, A. V. Dowling, C. Fitzer-Attas, A. Godfrey, J. G. Godino, N. Gujar, E. Izmailova, et al. Verification, Analytical Validation, and Clinical Validation (V3): The Foundation of Determining Fit-for-Purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digital Medicine*, 3(1):55, 2020.

[73] D. Gotz and H. Stavropoulos. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1783–1792, 2014.

[74] G. B. Gunn, T. R. Mendoza, C. D. Fuller, I. Gning, S. J. Frank, B. M. Beadle, E. Y. Hanna, C. Lu, C. S. Cleeland, and D. I. Rosenthal. High Symptom Burden Prior to Radiation Therapy for Head and Neck Cancer: A Patient-Reported Outcomes Study. *Head & neck*, 35(10):1490–1498, 2013.

[75] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao. Visualizing Uncertainty and Alternatives in Event Sequence Predictions. *CHI Conference on Human Factors in Computer Systems*, p. 1–12, 2019.

[76] S. Guo, Z. Jin, D. Gotz, F. Du, H. Zha, and N. Cao. Visual progression analysis of event sequence data. *IEEE Transaction on Visualization and Computer Graphics*, 25(1):417–426, 2018.

[77] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. Eventthread: Visual Summarization and Stage Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):56–65, 2017.

[78] T. Guo, T. Lin, and N. Antulov-Fantulin. Exploring Interpretable LSTM Neural Networks Over Multi-Variable Data. In *International Conference on Machine Learning*, pp. 2494–2504. PMLR, 2019.

[79] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, and N. Cao. Survey on Visual Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5091–5112, 2021.

[80] S. Gupta and R. Mamtora. A Survey on Association Rule Mining in Market Basket Analysis. *International Journal of Information Technology and Computer Science*, 4(4):409–414, 2014.

[81] C. K. Gwede, B. J. Small, P. N. Munster, M. A. Andrykowski, 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.

[82] R. H. Birk and G. Samuel. Can digital data diagnose mental health problems? a sociological exploration of 'digital phenotyping'. *Sociology of Health & Illness*, 42(8):1873–1887, 2020.

[83] M. Hahsler and S. Chelluboina. Visualizing association rules: Introduction to the r-extension package arulesviz. *R proj. mod.*, pp. 223–238, 2011.

[84] P. Hanula, K. Piekutowski, J. Aguilera, and G. E. Marai. Darksky halos: Use-based exploration of dark matter formation data in a hybrid immersive virtual environment. *Frontiers in Robotics and AI*, 6:11, 2019.

[85] J. Hao, Q. Shi, Y. Ye, and W. Zeng. TimeTuner: Diagnosing Time Representations for Time-Series Forecasting with Counterfactual Explanations. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1183, 2024.

[86] T. A. Harbig, S. Nusrat, T. Mazor, Q. Wang, A. Thomson, H. Bitter, E. Cerami, and N. Gehlenborg. OncoThreads: Visualization of Large-Scale Longitudinal Cancer Molecular Data. *Bioinformatics*, 37:i59–i66, 2021.

[87] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, Inc., USA, 1999.

[88] C.-W. Huang, R. Lu, U. Iqbal, S.-H. Lin, P. A. Nguyen, H.-C. Yang, C.-F. Wang, J. Li, K.-L. Ma, Y.-C. Li, et al. A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records. *BMC Medical Informatics and Decision Making*, 15(1):92, 2015.

[89] Q. R. Huang, Z. Qin, S. Zhang, and C. M. Chow. Clinical Patterns of Obstructive Sleep Apnea and its Comorbid Conditions: A Data Mining Approach. *Journal of Clinical Sleep Medicine*, 04(06):543–550, 2008.

[90] K. Huat Ong, K. leong Ong, W. keong Ng, and E. peng Lim. Crystalclear: Active Visualization of Association Rules. In *International Workshop on Active Mining*, 2002.

[91] J. Illi, C. Miaskowski, B. Cooper, J. D. Levine, L. Dunn, C. West, M. Dodd, A. Dhruva, S. M. Paul, C. Baggott, 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.

[92] A. Inselberg. Multidimensional Detective. *Proceedings of VIZ: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, p. 100–107, 1997.

[93] P. Jaccard. The Distribution of the Flora in the Alpine Zone. 1. *New Phytologist*, 11(2):37–50, 1912.

[94] K. Jensen, K. Lambertsen, and C. Grau. Late Swallowing Dysfunction and Dysphagia After Radiotherapy for Pharynx Cancer: Frequency, iItensity and Correlation with Dose and Volume Parameters. *Radiotherapy and Oncology*, 85(1):74–82, 2007.

[95] W. Jentner and D. A. Keim. Visualization and Visual Analytic Techniques for Patterns. *High-Utility Pattern Mining: Theory, Algorithms and Applications*, p. 303–337, 2019.

[96] Z. Jiang, M. Luskus, S. Seyedi, E. L. Griner, A. B. Rad, G. D. Clifford, M. Boazak, and R. O. Cotes. Utilizing Computer Vision for Facial Behavior Analysis in Schizophrenia Studies: A Systematic Review. *PloS one*, 17(4):e0266828, 2022.

[97] Z. Jin, S. Cui, S. Guo, D. Gotz, J. Sun, and N. Cao. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare*, 1(1), 2020.

[98] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., USA, 2002.

[99] Josephine, N. and others. SG-RAD: A Visual Analytics System in Subgroup and Risk Factors Analysis and Discovery. In *Pacific Visualization Conference*, pp. 331–336. IEEE, 2024.

[100] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[101] M. Kamal, M. P. Barrow, J. S. Lewin, A. Estrella, G. B. Gunn, Q. Shi, T. M. Hofstede, D. I. Rosenthal, C. D. Fuller, K. A. Hutcheson, et al. Modeling Symptom Drivers of Oral Intake in Long-Term Head and Neck Cancer Survivors. *Supportive Care in Cancer*, 27(4):1405–1415, 2019.

[102] M. Karpefors and J. Weatherall. The Tendril Plot—A Novel Visual Summary of the Incidence, Significance and Temporal Aspects of Adverse Events in Clinical Trials. *Journal of the American Medical Informatics Association*, 25(8):1069–1073, 2018.

[103] M. Kaur and S. Kang. Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, 85:78–85, 2016.

[104] H.-J. Kim, I. Abraham, and P. S. Malone. Analytical Methods and Issues for Symptom Cluster Research in Oncology. *Current Opinion in Supportive and Palliative Care*, 7(1):45–53, 2013.

[105] J. Kim, S. Lee, H. Jeon, K.-J. Lee, H.-J. Bae, B. Kim, and J. Seo. Phenoflow: A Human-LLM Driven Visual Analytics System for Exploring Large and Complex Stroke Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[106] L. Kim and S. Myoung. Comorbidity Study of Attention-deficit Hyperactivity Disorder (ADHD) in Children: Applying Association Rule Mining (ARM) to Korean National Health Insurance Data. *Iranian Journal of Public Health*, 47(4):481–488, 2018.

[107] J. Kirkova, A. Aktas, D. Walsh, and M. P. Davis. Cancer Symptom Clusters: Clinical and Research Methodology. *Palliative Medicine*, 14(10):1149–1166, 2011.

[108] P. Klemm, S. Oeltze-Jafra, K. Lawonn, K. Hegenscheid, H. Völzke, and B. Preim. Interactive Visual Analysis of Image-Centric Cohort Study Data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1673–1682, 2014.

[109] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert. A Survey on Visual Analysis Approaches for Financial Data. *Computer Graphics Forum*, 35(3):599–617, 2016.

[110] R. Kost, B. Littenberg, and E. S. Chen. Exploring Generalized Association Rule Mining for Disease Co-Occurrences. *American Medical Informatics*, p. 1284–1293, 2012.

[111] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual Inspection of Black-Box Machine Learning Models. *CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697, 2016.

[112] A. Kumar, T. Jaquenoud, J. H. Becker, D. Cho, M. R. Mindt, A. Federman, and G. Pandey. Can You Hear Me Now? Clinical Applications of Audio Recordings. *medRxiv*, 2022.

[113] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng. DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways. *IEEE Transaction on Visualization and Computer Graphics*, 27(9):3685–3700, 2020.

[114] B. C. Kwon, M. Choi, J. T. Kim, et al. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019.

[115] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the Art of Visual Analytics for Explainable Deep Learning. In *Computer Graphics Forum*, vol. 42, pp. 319–355. Wiley Online Library, 2023.

[116] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable Decision Sets: A Joint Framework for Description and Prediction. *SIGKDD International Conference on Knowledge Discovery and Data Minining*, p. 1675–1684, 10 pages, 2016.

[117] J. A. Langendijk, P. Doornaert, I. M. Verdonck-de Leeuw, C. R. Leemans, N. K. Aaronson, and B. J. Slotman. Impact of Late Treatment-Related Toxicity on Quality of Life Among Patients With Head and Neck Cancer Treated With Radiotherapy. *Journal of Clinical Oncology*, 26(22):3770–3776, 2008.

[118] 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 Applied Statatistics*, 9(3):1350 – 1371, 2015.

[119] H. Li, Y. Wang, and H. Qu. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. *CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024.

[120] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale. Timespan: Using Visualization to Explore Temporal Multi-Dimensional Data of Stroke Patients. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):409–418, 2015.

[121] H.-Y. Lu, Y. Li, and K.-L. Ma. A Visual Analytics Design for Connecting Healthcare Team Communication to Patient Outcomes. *Proceedings of the International Conference on Medical Health Informatics*, 2024.

[122] T. Luciani, A. Burks, C. Sugiyama, J. Komperda, and G. E. Marai. Details-First, Show Context, Overview Last: Supporting Exploration of Viscous Fingers in Large-Scale Ensemble Simulations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):1225–1235, 2018.

[123] T. Luciani, A. Wentzel, B. Elgohari, H. Elhalawani, A. Mohamed, G. Canahuate, D. M. Vock, C. D. Fuller, and G. E. Marai. A Spatial Neighborhood Methodology for Computing and Analyzing Lymph Node Carcinoma Similarity in Precision Medicine. *Joural of Biomedical Informatics*, 112:100067, 2020.

[124] T. B. Luciani, B. Cherinka, D. Oliphant, S. Myers, W. M. Wood-Vasey, A. Labrinidis, and G. E. Marai. Large-Scale Overlays and Trends: Visually Mining, Panning and Zooming the Observable Universe. *IEEE Transactions on Visualization and Computer Graphics*, 20(7):1048–1061, 2014.

[125] C. Ma, T. Luciani, A. Terebus, J. Liang, and G. E. Marai. PRODIGEN: Visualizing the Probability Landscape of Stochastic Gene Regulatory Networks in State and Time Space. *BMC Bioinformatics*, 18(2):1–14, 2017.

[126] T. Ma and A. Zhang. Integrate Multi-Omic Data Using Afinity Network Fusion (ANF) for Cancer Patient Clustering. *International Bioinformatics and Biomedicine*, p. 398–403, 2017.

[127] N. S. Madiraju. *Deep Temporal Clustering: Fully Unsupervised Learning of Time-Domain Features*. PhD thesis, Arizona State University, 2018.

[128] S. Malik, F. Du, M. Monroe, E. Onukwugha, et al. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics, 2015.

[129] G. E. Marai. Activity-Centered Domain Characterization for Problem-Driven Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):913–922, 2018.

[130] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio, G. Canahuate, D. M. Vock, A. S. Mohamed, and C. D. Fuller. Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1732–1745, 2019.

[131] G. E. Marai and T. Möller. The Fabric of Visualization. In *Foundations of Data Visualization*, pp. 5–14. Springer, 2020.

[132] A. Maries, N. Mays, M. Hunt, K. F. Wong, W. Layton, R. Boudreau, C. Rosano, and G. E. Marai. GRACE: A Visual Comparison Framework for Integrated Spatial and Non-Spatial Geriatric Data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2916–2925, 2013.

[133] D. Martens, B. Baesens, and T. Van Gestel. Decompositional Rule Extraction From Support Vector Machines by Active Learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):178–191, 2008.

[134] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard. Acume: A New Visualization Tool for Understanding Facial Expression and Gesture Data. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 591–596. IEEE, 2011.

[135] 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.

[136] M. Meuschke, U. Niemann, et al. GUCCI-Guided Cardiac Cohort Investigation of Blood Flow Data. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[137] C. Miaskowski, A. Barsevick, A. Berger, et al. Advancing Symptom Science Through Symptom Cluster Research: Expert Panel Proceedings and Recommendations. *Journal of the National Cancer Institute*, 109, 2017.

[138] C. Miaskowski, B. A. Cooper, S. M. Paul, et al. Subgroups of Patients With Cancer With Different Symptom Experiences and Quality-of-Life Outcomes: A Cluster Analysis. *Oncology Nursing Forum*, 33(5):E79–89, 2006.

[139] D. Michaelis, T. Gramss, and H. W. Strube. Glottal-to-Noise Excitation Ratio–A New Measure for Describing Pathological Voices. *ACTA Acustica United with Acustica*, 83(4):700–706, 1997.

[140] M. L. Miller, I. M. Raugh, G. P. Strauss, and P. D. Harvey. Remote Digital Phenotyping in Serious Mental Illness: Focus on Negative Symptoms, Mood Symptoms, and Self-Awareness. *Biomarkers in Neuropsychiatry*, p. 100047, 2022.

[141] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352, 2018.

[142] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren. ProtoSteer: Steering Deep Sequence Model With Prototypes. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):238–248, 2019.

[143] S. Monadjemi, Y. Guo, K. Xu, A. Endert, and A. Crisan. A Scoping Review of Mixed Initiative Visual Analytics in the Automation Renaissance. *arXiv preprint arXiv:2509.19152*, 2025.

[144] M. Monroe, R. Lan, et al. Temporal Event Sequence Simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013.

[145] S. Moritz, M. Rufer, S. Fricke, A. Karow, M. Morfeld, L. Jelinek, and D. Jacobsen. Quality of Life in Obsessive-Compulsive Disorder Before and After Treatment. *Comprehensive Psychiatry*, 46(6):453–459, 2005.

[146] J. Müller, V. Zebralla, S. Wiegand, and S. Oeltze-Jafra. Interactive Visual Analysis of Patient-Reported Outcomes for Improved Cancer Aftercare. *EuroVA at the EuroVis Conference*, pp. 78–82, 2019.

[147] Multidisciplinary Larynx Cancer Working Group. Conditional Survival Analysis of Patients With Locally Advanced Laryngeal Cancer: Construction of a Dynamic Risk Model and Clinical Nomogram. *Scientific Reports*, 7(1):43928, 2017.

[148] D. Nguyen, W. Luo, D. Phung, and S. Venkatesh. LTARM: A Novel Temporal Association Rule Mining Method to Understand Toxicities in a Routine Cancer Treatment. *Knowledge-Based Systems*, 161:313–328, 2018.

[149] N. P. Nguyen, H. J. Smith, and S. Sallah. Evaluation and Management of Swallowing Dysfunction Following Chemoradiation for Head and Neck Cancer. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 15(2):130–133, 2007.

[150] E. Niklas and T. Philippas. Growing Squares: Animated Visualization of Causal Relations. *Symposium on Software Visualization*, 10:774833–774836, 2003.

[151] N. Nipu, C. Floricel, N. Naghashzadeh, R. Paoli, and G. E. Marai. Visual Analysis and Detection of Contrails in Aircraft Engine Simulations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):798–808, 2022.

[152] D. Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic books, 2013.

[153] C. M. Nutting, J. P. Morden, K. J. Harrington, T. G. Urbano, S. A. Bhide, C. Clark, E. A. Miles, A. B. Miah, K. Newbold, M. Tanay, et al. Parotid-Sparing Intensity Modulated Versus Conventional Radiotherapy in Head and Neck Cancer (PARSPORT): A Phase 3 multicentre Randomised Controlled Trial. *The Lancet Oncology*, 12(2):127–136, 2011.

[154] B. O'Sullivan, S. H. Huang, J. Su, A. S. Garden, E. M. Sturgis, K. Dahlstrom, N. Lee, N. Riaz, X. Pei, S. A. Koyfman, 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 Oncology*, 17(4):440–451, 2016.

[155] H. O'Reilly, D. Pigat, S. Fridenson, S. Berggren, S. Tal, O. Golan, S. Bölte, S. Baron-Cohen, and D. Lundqvist. The EU-Emotion Stimulus Set: A Validation Study. *Behavior Research Methods*, 48(2):567–576, 2016.

[156] G. Peake and J. Wang. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2060–2069, 2018.

[157] A. Piau, K. Wild, N. Mattek, and J. Kaye. Current State of Digital Biomarker Technologies for Real-Life, Home-Based Monitoring of Cognitive Function for Mild Cognitive Impairment to Mild Alzheimer Disease and Implications for Clinical Care: Systematic Review. *Journal of Medical Internet Research*, 21(8):e12785, 2019.

[158] C. Plaisant, B. Milash, A. Rose, S. Widoff, et al. LifeLines: Visualizing Personal Histories. *SIGCHI Conference on Human Factors in Computing Systems*, p. 221–227, 1996.

[159] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal. Alzheimer's Disease and Automatic Speech Analysis: A Review. *Expert Systems With Applications*, 150:113213, 2020.

[160] R. G. Raidou, O. Casares-Magaz, A. Amirkhanov, V. Moiseenko, L. P. Muren, J. P. Einck, A. Vilanova, and M. E. Gröller. Bladder Runner: Visual Analytics for the Exploration of RT-Induced Bladder Toxicity in a Cohort Study. *Computer Graphics Forum*, 37(3):205–216, 2018.

[161] R. G. Raidou, U. A. van der Heide, et al. Visual Analytics for the Exploration of Tumor Tissue Characterization. *Computer Graphics Forum*, 34:11–20, 2015.

[162] J. Rogers, N. Spina, A. Neese, et al. Composer—visual cohort analysis of patient outcomes. *App. Clin. Info.*, 10(2):278, 2019.

[163] E. L. Rosenberg and P. Ekman. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.

[164] D. I. Rosenthal et al. The MD Anderson Symptom Inventory–Head and Neck Module, a Patient-Reported Outcome Instrument, Accurately Predicts the Severity of Radiation-Induced Mucositis. *International Journal of Radiation Oncology, Biology, Physics*, 72(5):1355–1361, 2008.

[165] D. I. Rosenthal, T. R. Mendoza, M. S. Chambers, J. A. Asper, I. Gning, M. S. Kies, R. S. Weber, J. S. Lewin, A. S. Garden, K. K. Ang, et al. Measuring Head and Neck Cancer Symptom Burden: The Development and Validation of the M. D. Anderson Symptom Inventory, Head and Neck Module. *Head and Neck*, 29(10):923–931, 2007.

[166] D. I. Rosenthal, T. R. Mendoza, C. D. Fuller, K. A. Hutcheson, X. S. Wang, E. Y. Hanna, C. Lu, A. S. Garden, W. H. Morrison, C. S. Cleeland, et al. Patterns of Symptom Burden During Radiotherapy or Concurrent Chemoradiotherapy for Head and Neck Cancer: A Prospective Analysis Using the University of Texas MD Anderson Cancer Center Symptom Inventory-Head and Neck Module. *Cancer*, 120(13):1975–1984, 2014.

[167] L. Sbitan, N. Alzraikat, H. Tanous, A. M. Saad, and M. Odeh. From One Size Fits All to a Tailored Approach: Integrating Precision Medicine Into Medical Education. *BMC Medical Education*, 25(1):90, 2025.

[168] M. Schuster and K. Paliwal. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[169] A. Ševčík and M. Rusko. A Systematic Review of Alzheimer's Disease Detection Based on Speech and Natural Language Processing. In *2022 32nd International Conference Radioelektronika*, pp. 01–05. IEEE, 2022.

[170] H. M. Skerman, P. M. Yates, and D. Battistutta. Multivariate Methods to Identify Cancer-Related Symptom Clusters. *Research in Nursing and Health*, 32(3):345–360, 2009.

[171] A. M. Smith, W. Xu, Y. Sun, J. R. Faeder, and G. E. Marai. RuleBender: Integrated Modeling, Simulation and Visualization for Rule-Based Intracellular Biochemistry. *BMC Bioinformatics*, 13(8):1–16, 2012.

[172] M. Sondag, C. Turkay, K. Xu, L. Matthews, S. Mohr, and D. Archambault. Visual Analytics of Contact Tracing Policy Simulations During an Emergency Response. In *Computer Graphics Forum*, vol. 41, pp. 29–41. Wiley Online Library, 2022.

[173] C. A. Steed, J. R. Goodall, J. Chae, and A. Trofimov. CrossVis: A Visual Analytics System for Exploring Heterogeneous Multivariate Data Eith Applications to Materials and Climate Sciences. *Graphics and Visual Computing*, 3:200013, 2020.

[174] N. Steinhauer, M. Hörbrugger, A. D. Braun, T. Tüting, S. Oeltze-Jafra, et al. Comprehensive Visualization of Longitudinal Patient Data for the Dermatological Oncological Tumor Board. *EuroVis*, 2020.

[175] H. Strobelt, S. Gehrmann, et al. Lstmvis: A tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):667–676, 2017.

[176] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina. Discovering Symptom Patterns of COVID-19 Patients Using Association Rule Mining. *Computers in Biology and Medicine*, 131:104249, 2021.

[177] A. Tifentale and L. Manovich. Selfiecity: Exploring Photography and Self-Fashioning in Social Media. In *Postdigital Aesthetics*, pp. 109–122. Springer, 2015.

[178] S. Ueng, C. Luo, T. Tsai, and H. Chang. Voice Quality Assessment and Visualization. *International Conference on Complex, Intelligent, and Software Intensive Systems*, pp. 618–623, 2012.

[179] L. Van den Bosch, A. van der Schaaf, H. P. van der Laan, F. J. Hoebers, O. B. Wijers, J. G. van den Hoek, K. G. Moons, J. B. Reitsma, R. J. Steenbakkers, E. Schuit, et al. Comprehensive Toxicity Risk Profiling in Radiation Therapy for Head and Neck Cancer: A New Concept for Individually Optimised Treatment. *Radiotherapy and Oncology*, 157:147–154, 2021.

[180] S. Vasudevan, A. Saha, M. E. Tarver, and B. Patel. Digital Biomarkers: Convergence of Digital Health Technologies and Biomarkers. *NPJ Digital Medicine*, 5(1):1–3, 2022.

[181] R. Voleti, J. M. Liss, and V. Berisha. A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):282–298, 2019.

[182] Q. Wang, T. Mazor, T. A. Harbig, E. Cerami, and N. Gehlenborg. Threadstates: State-Based Visual Analysis of Disease Progression. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):238–247, 2021.

[183] T. Wang et al. VIEWER: an Extensible Visual Analytics Framework for Enhancing Mental Healthcare. *Journal of the American Medical Informatics Association*, p. ocaf010, 2025.

[184] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. In *SIGCHI Conference on Human Factors in Computing Systems*, p. 457–466. ACM, 2008.

[185] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1049–1056, 2009.

[186] Y. Wang, G. M. Canahuate, L. V. Van Dijk, A. S. R. Mohamed, C. D. Fuller, X. Zhang, and G.-E. Marai. Predicting Late Symptoms of Head and Neck Cancer Treatment Using LSTM and Patient Reported Outcomes. *International Database Engineering & Applications Symposium*, pp. 273–279, 2021.

[187] Y. Wang, L. Van Dijk, A. S. Mohamed, M. Naser, C. D. Fuller, X. Zhang, G. E. Marai, and G. Canahuate. Improving Prediction of Late Symptoms using LSTM and Patient-reported Outcomes for Head and Neck Cancer Patients. In *IEEE International Conference on Healthcare Informatics*, pp. 292–300. IEEE, 2023.

[188] C. Ware, E. Neufeld, and L. Bartram. Visualizing Causal Relations. *IEEE Information Visualization*, 99:39–42, 1999.

[189] A. Wentzel, S. Attia, X. Zhang, G. Canahuate, C. D. Fuller, and G. E. Marai. DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[190] A. Wentzel, G. Canahuate, L. V. Van Dijk, A. S. Mohamed, C. D. Fuller, and G. E. Marai. Explainable Spatial Clustering: Leveraging Spatial Data in Radiation Oncology. *IEEE Visualization Conference*, p. 281–285, 2020.

[191] A. Wentzel, C. Floricel, G. Canahuate, M. A. Naser, A. S. Mohamed, C. D. Fuller, L. van Dijk, and G. E. Marai. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Computer Graphics Forum*, 2023.

[192] A. Wentzel, P. Hanula, T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. Vock, C. D. Fuller, and G. E. Marai. Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):949–959, 2020.

[193] A. Wentzel, P. Hanula, L. V. Van Dijk, B. Elgohari, A. S. Mohamed, C. E. Cardenas, C. D. Fuller, D. M. Vock, G. Canahuate, and G. E. Marai. Precision Toxicity Correlates of Tumor Spatial Proximity to Organs at Risk in Cancer Patients Receiving Intensity-Modulated Radiotherapy. *Radiotherapy and Oncology*, 148:245–251, 2020.

[194] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[195] K. Wongsuphasawat and D. Gotz. Outflow: Visualizing Patient Flow by Symptoms and Outcome. *IEEE Workshop on Visual Analytics in Healthcare*, p. 25–28, 2011.

[196] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, et al. LifeFlow: Visualizing an Overview of Event Sequences. *SIGCHI Conference on Human Factors in Computing Systems*, p. 1747–1756, 2011.

[197] Y. Wu et al. Liveretro: Visual Analytics for Strategic Retrospect in Livestream E-commerce. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[198] Y. Yamashita, M. Onodera, K. Shimoda, and Y. Tobe. Visualizing Health With Emotion Polarity History Using Voice. *International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pp. 1210–1213, 2019.

[199] H. Yang, C. Rudin, and M. Seltzer. Scalable Bayesian Rule Lists. *International Conference on Machine Learning*, pp. 3921–3930, 2017.

[200] J. Yuan, G. Y.-Y. Chan, B. Barr, K. Overton, K. Rees, L. G. Nonato, E. Bertini, and C. T. Silva. SUBPLEX: A Visual Analytics Approach to Understand Local Model Explanations at the Subpopulation Level. *IEEE Computer Graphics and Applicaitons*, 42(6):24–36, 2022.

[201] J. Yuan, O. Nov, and E. Bertini. An Exploration and Validation of Visual Factors in Understanding Classification Rule Sets. In *IEEE Transactions on Visualization and Computer Graphics*, pp. 6–10. IEEE, 2021.

[202] V. Zebralla, J. Müller, T. Wald, A. Boehm, G. Wichmann, T. Berger, K. Birnbaum, K. Heuermann, S. Oeltze-Jafra, T. Neumuth, et al. Obtaining Patient-Reported Outcomes Electronically With "OncoFunction" in Head and Neck Cancer Patients During Aftercare. *Frontiers in Oncology*, 10:2502, 2020.

[203] W. Zhang, J. K. Wong, X. Wang, Y. Gong, R. Zhu, K. Liu, Z. Yan, S. Tan, H. Qu, S. Chen, et al. Cohortva: A visual analytic system for interactive exploration of cohorts based on historical data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):756–766, 2022.

[204] Z. Zhang, D. Gotz, and A. Perer. Iterative Cohort Analysis and Exploration. *Information Visualization*, 14(4):289–307, 2015.

[205] J. Zhao, Z. Dai, P. Xu, and L. Ren. ProtoViewer: Visual Interpretation and Diagnostics of Deep Neural Networks With Factorized Prototypes. In *IEEE Transactions on Visualization and Computer Graphics*, pp. 286–290. IEEE, 2020.

[206] J. Zhao et al. Matrixwave: Visual Comparison of Event Sequence Data. *Conference on Human Factors in Computer Systems*, pp. 259–268, 2015.

[207] X. Zhao, Y. Wu, et al. Iforest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):407–416, 2018.

Carla Floricel

Visual Analytics | Visual Data Science | HCI | UI
Ph.D. Candidate · University of Illinois Chicago
(574) 300-9378 | cflori3@uic.edu | carlafloricel.github.io | CarlaFloricel

Education

University of Illinois Chicago — *Chicago, IL*
Ph.D. in Computer Science — *2019 - present*

University of Illinois Chicago — *Chicago, IL*
M.S. in Computer Science — *2025*

Politehnica University of Bucharest — *Bucharest, Romania*
B.S. in Computer Science — *2019*

Work Experience

Electronic Visualization Laboratory, University of Illinois Chicago — *UIC, Chicago, IL*
Research Assistant — *2019 – present*

- Design and development of visual analytics tools, primarily for medical applications.
- Client interviewing, prototyping, design, development, evaluation.

Epsilon — *Chicago, IL*
PhD Data Visualization Intern — *2024*

- Developed a fully working prototype for the company's digital advertising platform (DiME).
- Prototyped, designed, and developed a visual component to better understand client behavior.

University of Illinois Cancer Center - Diversity in Cancer Research (DICR) — *UIC, Chicago, IL*
Research Supervisor — *2023*

- Co-mentored an under-represented high school student.

OpenDBM, AiCure — *AiCure, New York, NY*
Research Intern — *2022*

- Designed and developed an open source visualization tool for a feature extraction toolkit.
- Visual analysis of digital biomarkers extracted from audio and video sources.

Computer Science Department, University of Illinois Chicago — *UIC, Chicago, IL*
Teaching Assistant — *2019, 2020, 2022 Fall Semesters*

- Office hours, lab presentations and supervision, assignment grading.

Unicredit Business Integrated Solutions — *Unicredit Services, Bucharest, Romania*
Junior Full Stack Developer — *2018 – 2019*

- Development, debugging, and data cleaning for production deployment.

Publications and Posters

L-VISP: LSTM Visualization for Interpretable Symptom Prediction in Patient Cohorts — *2025*
C. Floricel, Y. Wang, A. Wentzel, C.D. Fuller, G.E. Marai, M.E. Papka, G. Canahuate — *under review at CGF*

PRO-based Stratification Improves Model Prediction for Toxicity and Survival of Head and Neck Cancer Patients *2024*

E. Anyimadu, Y. Wang, **C. Floricel**, S. Kamel, C.D. Fuller, X. Zhang, G.E. Marai, G. Canahuate *IEEE JBHI*

Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining *2023*

C. Floricel, A. Wentzel, A.S. Mohamed, C.D. Fuller, G. Canahuate, G.E. Marai *IEEE VIS*

DASS Good: Explainable Data Mining of Spatial Cohort Data *2023*

A. Wentzel, **C. Floricel**, G. Canahuate, M.A. Naser, A.S. Mohamed, C.D. Fuller, L.V. Dijk, G.E. Marai *EuroVis*

MouseScholar: Evaluating an Image+Text Search System for Biocuration *2023*

J. T. Trabucco, **C. Floricel**, C. Arighi, H. Shatkay, D. Raciti, M. Ringwald, G.E. Marai *IEEE BIBM*

Opening Access to Visual Exploration of Audiovisual Digital Biomarkers: an OpenDBM Analytics Tool *2022*

C. Floricel, J. Epifano, S. Caamano, S. Kark, R. Christie, A. Masino, A.D. Paredes *IEEE VIS Biomedical AI*

Visual Analysis and Detection of Contrails in Aircraft Engine Simulations *2022*

N. Nipu, **C. Floricel**, N. Naghashzadeh, R. Paoli, G.E. Marai *IEEE VIS*

THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy *2021*

C. Floricel, N. Nipu, A. Wentzel, G. Canahuate, L.V. Dijk, A.S. Mohamed, C.D. Fuller, G.E. Marai *IEEE VIS*

Identifying Symptom Clusters from Patient Reported Outcomes through Association Rule Mining *2021*

M. Biggs, **C. Floricel**, L.V. Dijk, A.S. Mohamed, C.D. Fuller, G.E. Marai, X. Zhang, G. Canahuate *AIME*

Parameter Analysis and Contrail Detection of Aircraft Engine Simulations (Poster) *2021*

N. Nipu, **C. Floricel**, N. Naghashzadeh, R. Paoli, G.E. Marai *IEE VIS LDAV*

Visualizing Symptom Development During Head and Neck Cancer Treatment (Poster) *2020*

C. Floricel, A. Wentzel, N. Nipu, G. Canahuate, L.V. Dijk, C.D. Fuller, G.E. Marai *IEEE VIS*

Curated Projects

Visual Analytics for Head and Neck Cancer Patient Cohorts *UIC, Chicago, IL*

Graduate Research Project

- Novel encodings to summarize patient cohorts and their temporal, multivariate characteristics.
- Patient data modeling using association and sequential rule mining, clustering.
- Support model actionability and interpretability using visual analytics.
- Enhance multidisciplinary clinician-modeler collaborations in medical decision-making.
- Qualitative and quantitative methods to evaluate actionability, perceived usefulness, and trust.

Visual Analytics of Global Fishing *UIC, Chicago, IL*

Visual Analytics Class Group Project

- Visual analytics of illegal, unreported, and unregulated (IUU) fishing around the world.
- Detection of IUU based on longitudinal vessel activity.

Visual Modeling of Aircraft Contrails *UIC, Chicago, IL*

Graduate Research Project

- Visualization of contrail formation from aircraft simulations.
- Visual analysis for I/O simulation parameters and ensemble members.

CAVE2 Airport Control Simulator *UIC, Chicago, IL*

Game Design Class Group Project

- Airport control simulator for a large-scale virtual-reality environment.
- Designed and developed for CAVE2 at EVL.

vINCI: Clinically-Validated Integrated Support for Assistive Care and Lifestyle Improvement, the Human Link *Politehnica University Bucharest, Romania*

Bachelor's Thesis

- Designed, tested and developed an interface to improve the quality of life of older adults using IoT.
- Second prize for best research project presentation at Politehnica University competition.

Talks and Presentations

2025	**Patient Cohort Visual Analytics for After Treatment Care**, Presented dissertation proposal	*Chicago, IL*
2024	**Epsilon Internship**, Virtually presented the results of the summer project	*Epsilon, Chicago*
2023	**Roses Have Thorns: Understanding the Downside of Cncological Care Delivery Through Visual Analytics and Sequential Rule Mining**, Paper presented during the IEEE VIS 2023 conference	*Melbourne, Australia*
2022	**Opening Access to Visual Exploration of Audiovisual Digital Biomarkers: an OpenDBM Analytics Tool**, Presented during the IEEE VIS 2022 Visualization in Biomedical AI Workshop	*Oklahoma City, OK*
2022	**OpenDBM Internship**, Virtually presented the resulting visual system prototype	*AiCure, NY*
2022	**THALIS:Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy**, Poster presented during the CRA-WP Grad Cohort for Women Workshop	*New Orleans, LA*
2021	**THALIS:Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy**, Paper presented during the IEEE VIS 2021 conference and at one of the satellite events	*Chicago, IL*
2021	**A personal experience: the doctoral program in the USA**, Presented graduate study opportunities in the USA for Romanian students	*Politehnica University Bucharest, Romania*
2021	**Visual Analysis of Patient Timelines**, Ph.D. Qualifier Examination	*UIC, Chicago, IL*

Volunteering

EuroVis/CGF 2022, CGF 2023, IEEE VIS/TVCG 2022 - 2024

Full Paper Reviewer *2022 - 2024*

IEEE VIS 2021 - 2024

Student Volunteer (2023, 2024 Captain) *2021 - 2024*

- Moderated sessions and student tasks, revised presentation videos, registered participants.

Electronic Visualization Laboratory

Student Volunteer *September 2019 - Present*

- Conducted student/faculty visits, conducted demos for students and faculty.

Awards

Wexler Award for good standing incoming UIC students *2019-2020*

UIC Cancer Center Trainee Travel Award *2024*