DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer

Andrew Wentzel, Serageldin Attia, Xinhua Zhang, Guadalupe Canahuate, Clifton David Fuller, and G.Elisabeta Marai

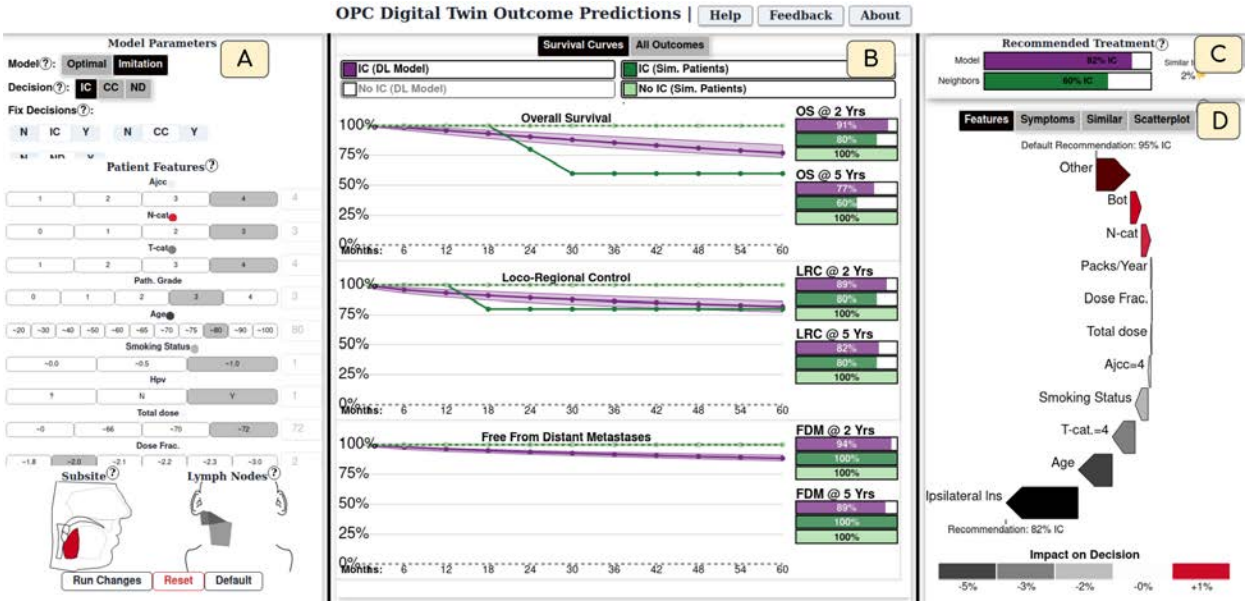

Fig. 1: Overview of DITTO. (A) Input panel to alter model parameters and input patient features. (B) Temporal outcome risk plots for the patient based on different models and treatment groups. (C) Treatment recommendation based on the twin model and similar patients. (D) Auxiliary data panel, currently showing a waterfall plot of how each feature cumulatively contributes to the model decision.

Abstract—Digital twin models are of high interest to Head and Neck Cancer (HNC) oncologists, who have to navigate a series of complex treatment decisions that weigh the efficacy of tumor control against toxicity and mortality risks. Evaluating individual risk profiles necessitates a deeper understanding of the interplay between different factors such as patient health, spatial tumor location and spread, and risk of subsequent toxicities that can not be adequately captured through simple heuristics. To support clinicians in better understanding tradeoffs when deciding on treatment courses, we developed DITTO, a digital-twin and visual computing system that allows clinicians to analyze detailed risk profiles for each patient, and decide on a treatment plan. DITTO relies on a sequential Deep Reinforcement Learning digital twin (DT) to deliver personalized risk of both long-term and short-term disease outcome and toxicity risk for HNC patients. Based on a participatory collaborative design alongside oncologists, we also implement several visual explainability methods to promote clinical trust and encourage healthy skepticism when using our system. We evaluate the efficacy of DITTO through quantitative evaluation of performance and case studies with qualitative feedback. Finally, we discuss design lessons for developing clinical visual XAI applications for clinical end users.

Index Terms—Medicine; Machine Learning; Application Domains; High Dimensional data; Spatial Data; Activity Centered Design

1 Introduction

Head and Neck Cancer (HNC) is a serious but treatable illness that affects up to 65,000 people each year in the United States alone. Care for HNC patients is a complex, multi-stage process that is dependent on the spatial location of the disease and its spread, and which includes potentially repeated cycles of surgery, chemotherapy, and radiation therapy. Determining the appropriate course of treatment for each patient is currently reliant on high level national guidelines and clinician cumulative experience. However, current guidelines do not adequately address the wide range of individual patient responses to treatments or the dynamic adjustments clinicians must make in response. For example, treating patients with chemotherapy before radiation treatment may reduce the overall tumor size and therefore reduce the risk of severe long-term side effects, but may also increase mortality risk. As a result, there is exceptional interest in digital twin (DT) models of the treatment process to help HNC oncologists better understand the potential risks and benefits of different treatment decisions at each state in the treatment process. Digital twins are data-driven simulations of patients and how they respond to treatment, which can be used to tailor treatments for individual patients based on how they are expected to respond to different interventions. DTs require complex simulations of a patient's health at multiple points in treatment, and thus rely on

- *Andrew Wentzel, Xinhua Zhang, and Liz Marai are with the University of Illinois Chicago E-mail: {awentze2 | gmarai}@uic.edu*
- *G. Canahuate is with the University of Iowa*
- *S. Attia and C.D. Fuller are with the MD Anderson Cancer Center*

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

models that are more complex than those typically used in clinical settings (e.g., logistic regression). Data visualization is an underutilized resource that can help clinicians interact more effectively with these digital twins.

Visualization for digital twins for subject-matter experts is an underexplored visualization challenge [44], with many additional challenges specific to HNC clinical decision-making. In terms of data, DTs consider multiple aspects of treatment, in addition to a combination of spatial and dynamic multivariate data to capture the patient state, which need to be visualized. In terms of outcomes, patient simulations yield dense, dynamic, and temporal outcome predictions, which need to be presented efficiently to users who may be interested in only a small subset of the resulting outcomes, depending on the context.

Furthermore, creating usable DT models also constitutes a visual explainable AI (XAI) challenge. While many XAI approaches have been developed for explaining models to model builders, less work has looked at the specific needs of model clients, who have unique requirements when considering both model performance and model explanations. For example, HNC clinical decisions may heavily depend on factors like spatial features and clinician experience, making simplification of results difficult. Issues with model explainability and actionability may be a factor in the low penetration of ML models in medicine (<2% [2]) beyond medical image analysis ML. Additionally, since existing models often contain biased or insufficiently diverse datasets to perfectly model the cohort, it is important to give recommendations that allow for model introspection and support appropriate trust in the recommendations while allowing physicians to identify cases when the model should be disregarded. Finally, complex model results need to be communicated to physicians while ensuring that the visualizations are sufficiently familiar so that they require minimal training.

In this work, we introduce a visual analysis interface for digital twins in oropharyngeal cancer treatment (DITTO). Our specific contributions are: 1) Requirements engineering of the factors that HNC oncologists consider when interacting with digital twin systems for treatment planning; 2) The design and implementation of a visual computing system with a dual digital-twin back-end, one twin (set of models) of the HNC patients, and one twin (set of models) of the HNC physician decisions; 3) The design of visual encodings for the visual computing front-end, with a focus on supporting clinicians and supporting both trust and skepticism in the models; and 4) A qualitative evaluation of the system with clinicians, resulting in visual digital twin design insights.

2 Related Work

2.1 Patient Risk Modeling

Research in head and neck (HNC) oncology focuses on evaluating ways of improving patient outcomes through changes in treatment. Current approaches have seen relatively high survival rates ($\sim$ 86%) in many HNC patients. As a result, current work often focuses on reducing side-effects (*toxicities* or *symptoms*) from treatment for patients with good survival probabilities. Earlier works have built interpretable models for predicting patient clinical outcomes for HNC patients such as survival and toxicity using clinical features [37], lymph node involvement [32, 68], tumor location [66, 67] and dose distributions [69], and radiomics [9]. This work is an extension of these approaches with a focus on temporally changing outcomes as well as intermediate treatment responses, which relies on more complex black-box models and post-hoc, instance based explanation methods for model interpretability.

Risk modeling for patients with censored time-to-event outcome data like survival [29] is generally modeled using approaches such as cox proportional hazard models [52], non-parametric Kaplan-Meier analysis, and fully parametric models such as linear regression and survival trees [60, 73]. This work adapts a deep-learning approach to survival modeling called deep survival machines (DSMs), which use a fully parametric mixture of distributions fitted to the training data [42]. Other approaches have adapted deep learning approaches to Cox proportional hazard models [71] and attention-based transformer models for predicting survival [30]. However, none of these models account for differences in patient response during treatment.

In terms of Reinforcement learning, VA for interpretable RL is usually focused on targeting model builders [58, 59]. For clinical models, several systems have proposed attention weights for interpreting temporal neural networks [10, 33]. In terms of visualization, RetainVis [26] focuses on exploring a recurrent neural network on temporal electronic health record data in patient cohorts. RMExplorer [27] uses subgroup statistics and feature attribution methods to explore model fairness in risk models.

More generally, DrugExplorer [61] proposed a general framework for XAI applied to drug discovery. In terms of presenting models to users, Suh et al [47] and Zitek et al [78] discuss strategies for communicating models to domain experts, but do not expand this to applications in decision support. Kaur et al. [24] showed that many users can "over trust" erroneous model explanations they don't understand properly. VISPUR [51] discusses methods of identifying spurious correlations in causal models, but do not focus on integrating domain expert knowledge.

2.2 Digital Twins

A digital twin is a digital model of a real-world system or process, that serves as the digital counterpart of it for practical purposes, such as simulation, integration, testing, monitoring, and maintenance. Although the term digital twin was introduced in 2010, visual steering of detailed computer simulations (i.e., fdigital twins) has been used before for flood simulation planning [63] and VR applications for manufacturing [77].

In healthcare, limited work has been done in exploring digital twins for patients using dashboards [23, 28] and 3D models to visualize blood flow [39]. Digital twin tools have also made for simulating physicians [49]. Other approaches have built digital twins for radiation dosage adaptation [53], glioblastoma treatment [13], and emergency department managementand [6], but do not integrate visualization or explainability. Marai et al [38] developed a web visualization tool for HNC patient risk based on similar patients that allows for what-if analysis. Our work uniquely integrates visualization for both a digital twin and digital physicians. Additional, to our knowledge, there has been no interactive visual computing approach for digital twins that can also factor temporal decision-making.

2.3 Decision Support Systems

Relevant to this work is clinical decision support (CDSS) systems. Jacobs et al. discuss a CDSS for clinical depression [21]. Other systems have focused on identifying ways of supporting physician workflows for heart implants [72], critical care patients [75] and diabetes care [8]. Other work has focused on model building for CDSS Bayes networks [40], and integrating feature explanations to help train physicians in diagnostics [45]. More generally, a recent study has suggested that users are more likely to use AI recommendations for harder tasks [18]. Despite this, few visual systems have focused on decision recommendation in the context of explainable ML recommendations.

Several systems have been developed specifically to visually communicate risk prediction to clinical users or patients, although none of them focus on deep learning-driven personalized patient outcomes. A majority of these systems focus on variants of Kaplan Meier plots to communicate patient survival based on general diagnostic features [11, 12, 54]. Oncofunction [74] focuses on helping patients plan post-treatment symptoms. PROACT [19] found patients were primarily interested in time left and survival risk at different time points using simple visualization methods. Vromans et al. [57] found that some information seeking was a coping mechanism for a percentage of the population, and personal quality-of-life measures were equally as important as patient survival. Floricel et al [14, 15] uses temporal glyphs and Sankey diagrams to show clusters of patient symptoms over time. Other common tools have used Kaplan-Meier curves [54] and nomograms [16] for HNC and prostate cancer. However, as far as we know, no online systems yet include digital twins with temporal state outcomes, or use model explainability methods with patient-specific predictions.

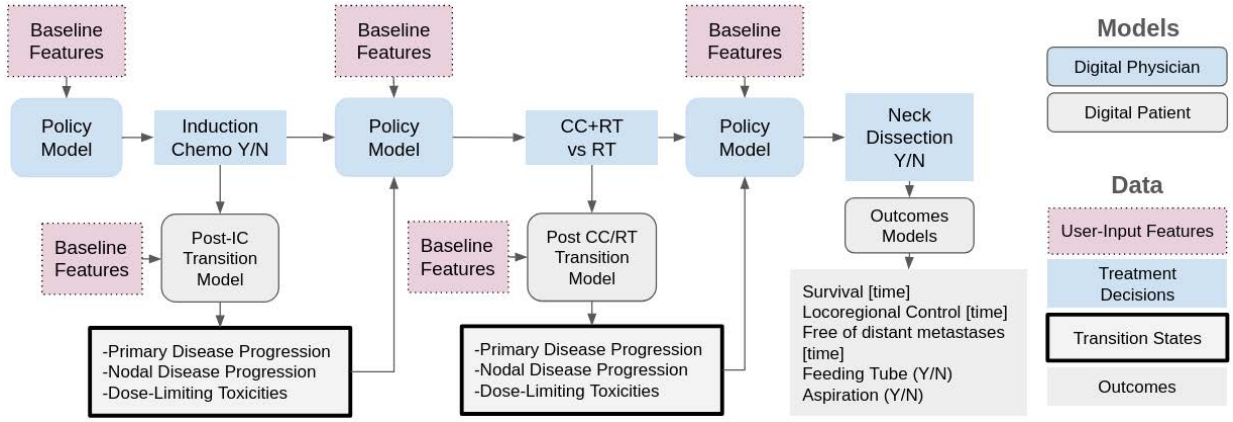

Fig. 2: Overview of the treatment sequence simulated by the digital twin models, along with the features to be considered. Intermediate results of induction and concurrent chemotherapy are used as inputs into the next decision. Final outcomes are a mixture of time-to-event curves and fixed binary outcomes. The DT model is trained to make optimal decisions with respect to the final outcomes.

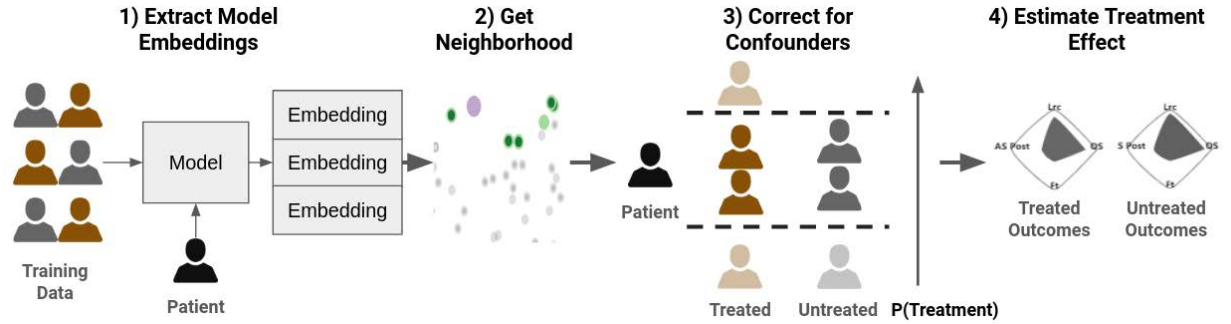

Fig. 3: Diagram neighbor-based models when predicting patient outcomes. Model embeddings from the policy model are used to extract the most similar patients. Neighbors are filtered by their estimated likelihood of receiving treatment from the imitation model for those closest to the new patient. the difference between untreated and treated filtered neighbors can then be used to estimate impact of treatment.

3 Methods

3.1 Requirement Analysis

This project was developed as part of a multi-year collaborative research project between HNC oncology radiotherapists from the MD Anderson Cancer Center, machine learning experts at the University of Illinois Chicago and University of Iowa, and visualization researchers at the University of Illinois Chicago. We followed the Activity-Centered Design (ACD) methodology, which has higher success rates in interdisciplinary settings than Human-Centered Design (63% vs. 25%) based on a survey of design studies [36]. Requirements for both the interface and models were initially gathered through interviews with three research oncologists, and gradually refined during weekly meetings through multiple rounds of parallel prototyping and feedback over the course of several months.

To gather formative feedback, a version of the interface designed based on our initial requirements was presented to a group of 11 physicians with clinical experience within the HNC oncology group at the MD Anderson Cancer Center. Several participants were familiar with the underlying dataset, but none had participated in the design of DITTO. During the session, participants were given an overview of the system components before being given a demo of the system and an online link where they were allowed to interact with the system on their own. This was followed by an open-ended discussion and a feedback interview.

3.2 Data Abstraction

Our dataset uses the patient cohort described in Tardini et al. [50]. The cohort consists of 526 anonymized patients with squamous cell oropharyngeal tumors treated using definitive radiation therapy at the MD Anderson Cancer Center between 2003 and 2013. All data were collected after approval from the MDACC IRB (PA16-0303 and RCR03-0800). All patients included also had either recorded deaths or a minimum followup time of 4 years. Patients diagnostic data, treatment sequence, and outcomes were collected using EHR records.

Standard treatments for patients include a mixture of *surgery*, *chemotherapy*, and *radiation therapy* (RT). Chemotherapy can either be given before RT (induction - IC) or with RT (concurrent - CC). While real treatment can involve multiple rounds of each therapy, our simplified treatment sequence models the treatment process as 3 decisions: chemotherapy before RT (IC), chemotherapy concurrent with RT (CC), and neck-dissection (ND), a common surgery. These decisions are critical decision points, aligned with the standard-of-care [1]. The entire treatment sequence model is shown in Fig. 2. Baseline features include age (cont.), if the patient is male or female/nonbinary (binary); race (binary x3); which regions of the neck have affected lymph nodes gregoire2014delineation (binary x14); smoking status (never, former, current) (ord.); total radiation dose to the tumor (cont.) and dose per-visit (cont.); tumor staging (ord. x3); and tumor subsite (categorical x6). Tumor staging features (T, N, and AJCC) are ordinal rankings of tumor severity used to decide on treatment regime based on tumor size and spread for the main tumor and nearby lymph nodes [3]. Race used a simplified grouping of demographics: White, African American/Black, Hispanic, and "Other", which is modeled as three one-hot variables in the data input. Minority inclusion reflects the demographics of the MD Anderson Cancer Center patient population which is approximately 85% Caucasian and 15% minority. Default gender is denoted as "male" in the model, to reflect the demographics of the patient population which is approximately 30% females and 70% males and did not have information on nonbinary individuals.

Each feature is associated with a feature-importance during model-

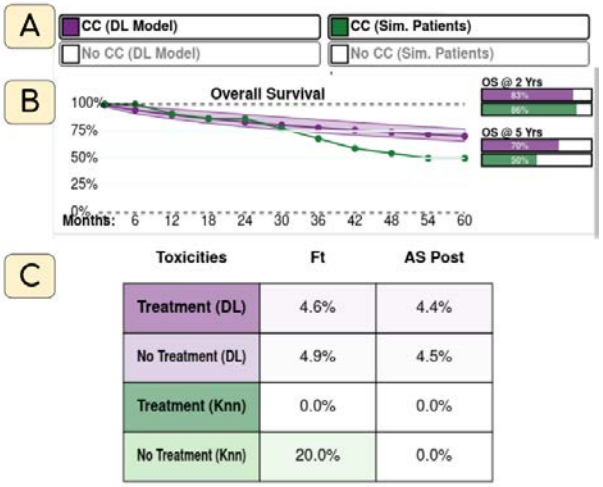

Fig. 4: Image of survival curves for a patient based on different models. (A) Legend with toggle-able models and outcomes, currently showing only treatment groups. (B) Survival plot for a patient, showing prediction with concurrent chemotherapy and 95% CI based on the DSM model (purple) and similar patients (green), along with fixed probabilities at 2 and 5 years. (C) Alternative outcomes view showing tables of predicted probabilities for additional toxicities (Ft - Feeding Tube, AS Post - Aspiration Post-Treatment).

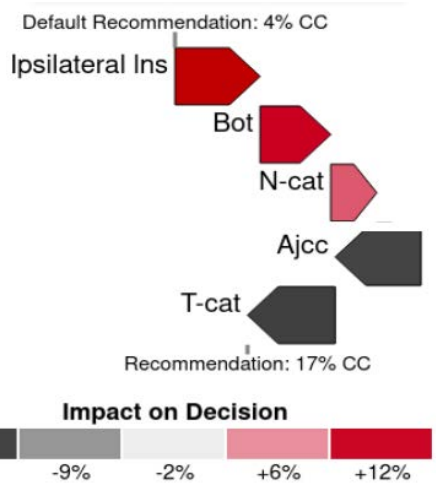

Fig. 5: Truncated feature contribution waterfall plot showing how each feature contributes to the final model recommendation, relative to the default (median) patient. Color double-encodes attributions

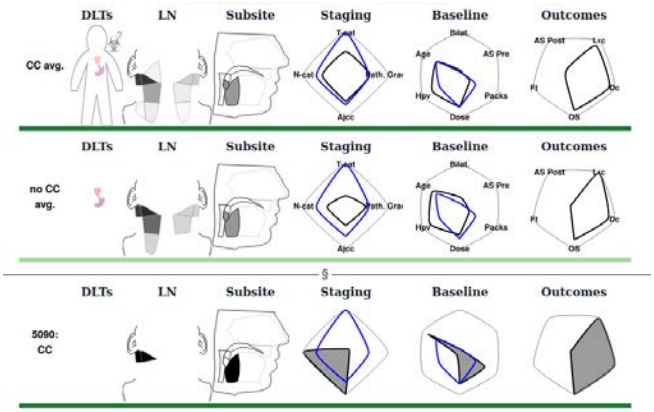

Fig. 6: Similar patients view showing the row for average treatment group. Each row shows toxicities, lymph node involvement, tumor subsite, staging, demographics, 4 year outcomes. Blue lines indicate the input features of the current patient in the staging and baseline Kiviat diagrams.

ing, based on how much it contributes to the twin model final decision, which is a value between 0 and 1.

After each decision, the patient response to treatment is modeled as a transition state in terms of the response (change in size) of the primary tumor, nodal tumors, and any dose-limiting toxicities (DLTs). Tumor response is categorized into 4 groups based on the amount of change in tumor size: progressive disease, stable disease, partial response, and complete response. DLT types considered in our model are: Hematological, Neurological, Dermatological, Gastrointestinal, or Other. All DLT categories with fewer than 3 instances in the dataset are grouped into the "other" class.

For temporal outcomes, we consider patient survival (OS), local-regional control (LRC), and distant control (FDM). For each of these, we collected whether the event occurred, as well as either the time of the event or last follow-up date. Additionally, we recorded whether the patient was hospitalized for a feeding tube (FT) or lung aspiration (AS) within 6 months after finishing treatment as binary toxicity outcomes. As an auxiliary outcome, we extracted symptom ratings from a separate dataset of 937 patients with self-reported outcomes after receiving radiation therapy [62], which is used in a secondary view to display possible symptom trajectories.

3.3 Digital Twins and Planning for Trust and Skepticism

One of our goals is to provide support for both trust and skepticism in the system recommendations. In prior work [65], we have discussed visualizing "counterfactuals", where the model recommendation and ground truth diverge, and adding cues to highlight when model predictions should be given more scrutiny. Because DITTO aims to provide treatment recommendations for a new patient, where the ground truth is not available, we implement instead "neighborhood-based models" that are shown alongside the treatment recommendation and predicted outcomes, to encourage both trust and skepticism in the digital twin recommendations. These neighbor-based models show outcomes from similar patients in the cohort and are described in Sec. 3.6.

Our core dual digital twin system is based on modeling of patient responses at each time point for a given patient, alongside modeling of the physician-recommended treatment. We specifically refer to the patient response models as the "Patient Simulator", and the predicted physician treatment decisions as the "Policy Model".

To further encourage trust and skepticism, and avoid reinforcing clinical bias, we planned to leverage and show recommendations from two deep learning models for the twin. In the context of modeling a physician we implement two approaches: imitation learning [76], which attempts to mimic what an expert would learn, and Deep Q Learning (DQN) [55], which attempts to find an optimal decision based on expected future losses. e also implement a preliminary imitation learning model for use by clinicians. We refer to the DQN strategy as the "Optimal Policy Model" and to the imitation learning strategy as the "Imitation Policy Model". For the purpose of the interface, viewers can select the specific strategy to be used, and examine results from that model strategy. These two supervised deep learning models (DQN and imitation) are "Digital Twins" of the physician decision process, in addition to the patient simulator.

In total, DITTO leverages and can show three recommendations: two based on the DTs of the physician decision process, and one based on the neighborhood models.

3.4 Task Analysis

Our system aims to help HNC oncology radiotherapists better understand the likely tradeoffs of adding other treatments to radiation therapy. Based on interviews, we found that clinicians generally consider treatment decisions at each stage individually, with a primary focus on identifying potential outcomes in terms of both immediate disease response, toxicity risk, and overall temporal outcomes up to 5 years.

Individual interests, degrees of information seeking, and visual literacy varied based on the individual practitioner and their backgrounds. As a result, our system was designed to have flexibility, with the most prominent views being presented by default, and more detailed views available on demand.

Additionally, we have found during our collaboration that some clinicians trust their past clinical experience over neural networks and cohort-based reasoning, while others tend to trust the model even when the system does not make sense. As a result, our main design focuses on simultaneously showing results from both the supervised deep learning models used in the digital twin and neighbor-based models that use similar patients in the cohort (Sec. 3.6), to cue the user to have appropriate trust and skepticism in the system.

Based on interviews and clinical feedback during the prototyping stage, we arrived at the following task abstraction:

T1. Identify the risk profile of a patient given a treatment selection.

1. Display the temporal risk of negative outcomes for the individual patient using the digital twin
2. Identify the cumulative patient risk in terms of the cohort of similar patients in the dataset
3. Identify the ideal treatment plan for the patient
4. Compare the patient to similar patients based on treatment and diagnostic data
5. Display expected patient symptom profiles after radiation therapy

T2. Identify relative benefit of treatment at the given time point.

1. Display the potential gain in therapeutic efficacy in terms of survival, disease control, and additional side effects
2. Compare expected cumulative tumor control and survival to the probability of additional toxicity due to treatment for the patient
3. Display the risk of dose-limiting toxicity due to chemotherapy or treatment complications

T3. Identify the trustworthiness of the model predictions and recommended treatment

1. Show the cumulative impact of each attribute on the recommended treatment in terms of percentage confidence
2. Flag when the patient is an outlier in the cohort
3. Display confidence intervals for the patient outcome predictions
4. Compare the prediction of the DT and neighbor-based models

Nonfunctional Requirements In addition to tasks, we determined a number of nonfunctional requirements. DITTO needed to build via visual scaffolding [35] on encodings in existing clinical tools, such as Kaplan-Meier plots, barcharts, and cumulative distribution histograms. Additionally, several clinicians desired to be able to show these results to patients, and thus designs needed to avoid causing patent anxiety (i.e., scale survival should show risk always compared to 0). Finally, DITTO needed to be responsive and available online to be used by clinicians at any time, with minimal (< 5 seconds) time to produce results for a new patient.

During the workshop, two participants requested information about data provenance and the model details, including limitations, available in the interface. Additionally, participants asked for the patient inputs to always be visible, and to only render additional views once an input has been manually submitted. Our original design also included both survival plots and barcharts of all outcomes and predicted transition states at the same time, in addition to median time to event for each temporal outcome. However, clinicians stated that most use-cases would focus on a smaller subset of results: the survival plots and survival at 2 and 5 years, with uncertainty values given, and that these designs should be centrally located, and additional results could be given on-demand as exact values.

3.5 Deep Reinforcement Learning Models

DITTO uses an extension of the dual digital twin system described in Tardini et al [50]. The full system is shown in Fig. 2. Patients are assumed to follow a series of 3 binary decisions: Induction chemotherapy (IC), Concurrent chemotherapy (CC), and Neck Dissection (ND).

Our digital twin is composed of multiple sub-models at each state in the treatment sequence, which are shown in more detail in Fig. 2. For the purpose of this section, we define terminology when referring to each of these sub-components. We call a model that predicts the patient's direct response to each treatment the "Transition Model", and the model that predicts long term temporal outcomes after definitive treatment is completed (i.e. survival and recurrence) the "Outcome Model". Following RL terminology, we refer to the model that simulates a physician as the "Policy Model". We have two versions of the policy model: The "Optimal Policy Model", and the "Imitation Policy Model", which attempts to predict the best treatment in terms of long term outcomes, and the treatment a physician would make, respectively. We only use one Policy model at a time, which is defined by the user. We use deep learning for all DT models due to their ability to deal with multimodal inputs with variable outputs and handle missing data [46]. The following section briefly discusses the details of each model.

To supplement the Digital Twin predictions, we show alternative predictions in the interface based on the most similar patients in the cohort at the given timepoint. We refer to this as the "Neighbor-based models" collectively, as we do not have to simulate responses at each step since all patients have ground truth decisions and patient responses available.

Bellow we briefly describe each model. Due to space constraints, full details, model parameters, and evaluation can be found in the supplemental material Appendix A.

Patient Simulator

To simulate the patient, we use a set of models to mimic intermediate response to treatment (transition models), and long-term response after treatment (outcome models).

Transition models predict patient response to treatment in terms of tumor shrinkage and severe toxicities from treatment. Specifically, we consider primary disease response (PD), and nodal disease response (ND), which are each 4 categorical ordinal variables, as well as 5 binary results for different types of dose-limiting toxicities (DLTs). For induction chemotherapy (IC), disease response is always assumed to be stable when no treatment is done.

For post-treatment outcomes, we predict a combination of temporal and static outcomes. We predict static outcomes using a deep neural network that predicts hospitalization due to two severe toxicities at up to 6 months after treatment: Aspiration (AS), and Feeding Tube insertion (FT). The temporal outcome model predicts cumulative patient risk over time for overall survival (OS), locoregional control (LRC), and distant metastases (FDM) for up to 5 years. Temporal risk models use a variant of deep survival machines (DSM) [42]. For all three outcomes, the DSM model returns a mixture of parametric log-normal distributions for the patient that can be used to provide a cumulative survival risk over time.

Because clinicians listed confidence intervals as important for reasoning about the model predictions (T3.3), all transition and outcome models are trained using dropout on the penultimate layer between 50% and 75%. During evaluation, we re-run each prediction with random dropout at least 20 times, and then save the 95% confidence intervals for each prediction.

Policy Modeling

The patient simulator models and ground truth responses are used as the environment to train a digital physician (policy model) . The policy model is a deep-learning based transformer encoder that predicts a binary treatment decision based on the baseline patient features, response to the previous treatment, previous decisions, and current timepoint.

Because we need to explain the policy model recommendations (T3), we use integrated gradients [48] to obtain feature importance for each decision relative to a baseline value. Integrated gradients was chosen as it satisfies the completeness axiom where attributions sum to the difference in the prediction between the baseline and actual recommendation, which was found to be easier to reason about with our clinicians. For our baseline, we assume the lowest possible rating

for most ordinal attributes such as tumor staging or disease response, and the most common value for categorical attributes such as gender, ethnicity, and tumor subsite, as well as age and dose to the main tumor, based on feedback from clinicians and what they found most intuitive.

3.6 Neighbor-based Models

To provide an alternative model prediction to improve user trust (Sec. 3.4). we provide methods for estimating different patient outcomes using similar patients in the cohort, based on the embeddings taken from the final layer in the policy model for the given time-point and output. Our approach uses a modified variant of average treatment effect, which is used in causal modeling for finding predicted effects from treatment while correcting for confounders.

For a new patient, we calculate a set of k patients whose embeddings are most similar at each time point in terms of embedding using euclidean distance. When predicting treatment policy (physician choices), we use a smaller subset of the $n, n < k$ most similar patients and report the percent of patients that received treatment. For other outcomes and patient response, we take from the k patients those with a predicted probability of receiving treatment that are within a certain value of the patient. We then calculate the relative prevalence of each outcome for the untreated and treated patients within this propensity-matched [4] group (Fig. 3). For our system, We calculate the value difference as a fixed percentage of the standard deviation of the logits of the propensity scores in the cohort, defined as:

$$cd = \alpha * \sqrt{\frac{1}{|X|}\sum_{x\in X}\left(\ln\left(\frac{p_x}{p_x-1}\right) - \frac{1}{|X|}\sum_{k\in X}\left(\ln\left(\frac{p_k}{p_k-1}\right)\right)\right)^2}$$

Where X is the cohort and p_n is the predicted probability of patient n receiving treatment. We use an α of .1 based on the suggested formula in [5], which is increase in increments of .1 until treated and untreated groups have at least 5 patients.

3.7 Implementation

Our back-end was implemented in Python using flask and pandas for data processing. Deep learning (digital twin) models use Pytorch, and deep survival machines use modified code taken from the auton-survival package [43]. Feature attributions were calculated using the Captum package [25]. Our system front-end uses react with d3.js. Our online interface requires approximately 3.6-4.5 seconds to return simulation results for a new patient with two cores on an AMD EPYC 7452 Processor and requires 4GB of ram with 4 worker processes on the server, based on test queries for 10 random patients in the cohort. Specific model parameters where chosen via model tuning are given in the supplemental material.

4 Design

4.1 Layout and Workflow

Our main system is divided into three main components: input, patient outcomes, and treatment recommendation + supplemental views. First, an input panel on the left is used to change model and patient details (Fig. 1-A). To minimize cognitive load, we focus on only showing one, user-selected treatment (IC, CC, or ND) at a time. Users can optionally decide on other treatment decision when calculating future patient outcomes, with the policy model handling the other treatment decisions when nothing is input by the user. Next, central views show patient survival outcomes (Fig. 1-B) as well as the recommended treatment for the patient (Fig. 1-C). Finally, additional views are shown via tabs to users who have an interest in more detailed information, such as model feature explanations (Fig. 1-D), similar patients, additional outcomes, and predicted symptoms ratings. These views are changed by toggling a set of buttons above the panel. Because many views are only of interest to certain users, we added functionality to resize width of each view via dragging the black vertical dividers, to allow users to expand auxiliary views as needed, while keeping the main goal of evaluating patient outcomes the main focus.

Whenever model predictions are shown in the interface, we present the deep-learning based Digital Twin predictions, and the neighbor based models. We use purple to encode Digital Twin predictions, and green to encode similar patient predictions.

4.2 User Input

The left panel allows inputting the relevant patient features and model parameters into the system. At the top, prompts are given for model input parameters: 1) whether the policy model should use the "optimal" or "imitation" strategy (Sec. 3.3); 2) what decision is being considered; and 3) if any of the other decisions in the system are assumed to be "fixed" (yes or no). By default, the decision is decided by the currently selected policy model's recommendation.

Below the model parameter input is a panel for the current patient (Fig. 1-A). By default, the average values for each feature are selected. We found that clinicians tend to think of continuous variables such as smoking pack-years and age in terms of discrete "bins" therefore, all features are shown using categorical stylized radio buttons to make selection easier, with free-text inputs on the side that allow users to use specific values when desired. These values are checked for validity based on the feature. When analyzing concurrent chemotherapy or neck-dissection, users can either specify the patient's primary and nodal tumor response to the previous round of chemotherapy, or allow the system to estimate this response automatically.

For the spatial inputs: affected lymph nodes and tumor subsites, we allow users to directly interact with diagrams of the respective areas. The diagram for the lymph nodes was previously developed alongside clinicians in our work with explainable lymph node clustering [64]. The diagram of tumor subsites were adapted from diagrams created by the MD Anderson Cancer Center. See the Appendix B for a labeled description of each spatial diagram.

In addition to feature input, we include color cues in the feature attribution plot for each of the features, described in detail in 4.4 (T3.1). These are shown as colored dots next to each feature for nonspatial inputs, and as a color fill in the spatial features.

Because we do not want to re-run the computationally expensive simulation every time a feature or parameter is changed, a new simulation is run using the updated features once the user selects the "run changes" button at the bottom. Additional buttons reset the feature inputs to the last time the simulation was run, and load the default patient features.

4.3 Survival Plots and Outcomes

When collecting feedback from HNC clinicians at the MD Anderson cancer center, several clinicians suggested that users with less information seeking behavior will primarily be interested in seeing tumor control and survival risk for treated and untreated groups over time. As a result, we centrally place an outcomes view panel (Fig. 4) that shows the model predictions for all relevant endpoints in our system. By default, we show temporal plots for survival, local-regional control, and distant metastasis for the treated and untreated groups using the Digital Twin outcome models (T1.1) and neighbor predictions (T1.2), up to 60 months post-treatment (Fig. 4-B). We also include 90% confidence intervals for Digital twin predictions as semi-transparent envelopes (T3.3). We chose to use temporal outcome plots as the main outcome plot, as oncologists often use variants of Kaplan Meier survival plots to assess patient risk. Additionally, the legend at the top can also be used to toggle off the visibility of certain models or treatment groups when the user only wants to see predictions for certain parameters (T3.4) (Fig. 4-A). Each output is color-coded, where hue encodes model group (Digital twin vs neighbor-based) and luminance encodes treatment group (darker for treated groups).

Because a subset of information-seeking clinicians were interested in more details regarding patient response, an alternative window (Fig. 4-C), shows static risk tables for all transition outcomes and temporal risk at 2 and 5 years for both Digital twin and neighbor-based predictions, for both the treated and untreated groups, for a total of 4 predictions each, via a toggle button (T2.1, T2.2, T2.3). This view relies on direct encoding of features. Additionally, each cell is color coded, with opacity encoding the risk percentage. These additional results were

originally encoded as a barchart shown alongside the survival plots, but were moved to a simpler, more explicit table shown on demand based on clinician feedback as well as recent findings suggesting that tables with explicit values are less prone to confirmation bias when reasoning about the data [70].

4.4 Treatment Recommendation

The right panel of DITTO is devoted to more detailed model results, based on the varying requirements cited by different clinicians. We show the recommended treatment based on both the policy model, and similar patients at the top, in terms of a percentage between 0 and 100% for the suggested treatment (Fig. 1-C) (T1.3). To provide a cue as to how reliable the model recommendation is, we calculate the Mahalanobis distance between the patient embedding taken from the model for each time point and the rest of the cohort (T3.2). We then calculate the relative percentile of the distance for this patient relative to the rest of the cohort (e.g., 0 to 100%), which is shown next to the recommendation. We show a symbol (thumbs-up vs thumbs-down) based on if the percentile is below or above 75%, respectively. This feature was based on a specific clinician request for a cue regarding whether the new patient recommendation can be trusted based on the cohort being used. Our original design included a full histogram. However, during the workshop, several clinicians misread the histogram, as some assumed being in the middle was better and others assumed the left was better. Additionally, clinicians did not find seeing the distribution of the full training cohort useful, and thus recommended using a text rating.

Below the model recommendation, a panel shows additional custom model details. By default, the view shows a waterfall chart variant(Fig. 5). This view shows the cumulative impact of each attribute on the final decision in terms of percentage confidence in the given treatment on the x-axis (T3.1). The baseline shows the decision impact for a "default" patient, which is either the lowest possible value for ordinal (e.g., tumor staging) or continuous values, or the most common value for categorical features. We then show the impact of each feature as an error moving the decision along the x-axis. Because the integrated-gradients feature attribution method satisfies "completeness", the final position at the bottom is equal to the position of the final decision relative to the first decision point. Each bar is drawn as an error that uses a diverging color scheme to double-encode impact size. All values below a certain threshold (1%) are aggregated into an "other" value as they have negligible interest to users. Features are shown in order of positive impact from the top to the bottom. This view was finalized as waterfall charts are an established method of showing feature attributions [20], with the arrows and color encoding added to improve intuitiveness of the system. Additionally, it was very well received by clinicians during prototyping, and described as "very intuitive" by a collaborator with no prior experience with feature attributions.

4.5 Similar Patients

Based on interviews and previous experience with clinicians, many HNC oncologists are interested in using previous patients to reason about likely outcomes and the trustworthiness of the prediction and improve domain sense. As a result, we include an optional view that shows details on the similar patients used in the Average Treatment Effect estimates (Fig. 6) (T1.2, T1.4). The view shows feature summaries of each patient, as well as the average values for the treated and untreated groups. Each patient is encoded as a single row of patients. We show the tumor subsite and lymph nodes as heatmaps using the diagrams described in Sec. 4.2, as well as a diagram for and dose-limiting toxicity from the current treatment (see supplemental material). Additionally, we show three Kiviat charts with distributions of the most relevant features: diagnostic tumor staging (T-stage, N-stage, Overall Stage, and pathological grade), important clinical features (HPV, smoking status, age, etc.), and patient outcomes at 4 years (survival, local-regional control, distant control, aspiration, and feeding tube). The features for the current patient for non-outcome features are overlaid on top of each patient in blue, to support comparison between the groups and the current patient. This design was based on prior work showing promising results for diagram based spatial encodings [64, 65], and radial charts to encode clinical features [34, 37] when displaying similar patients for clinicians, along with positive feedback from collaborators.

We use colored borders and labels to indicate which patients are in the treated and untreated groups. This view is included as it was found to be useful for clinicians that value inspecting individual patients, or identifying confounders that may impact the recommendation of the neighbor-based predictions. However, since many clinicians said this functionality was only a secondary concern, it is hidden by default.

4.6 Symptoms

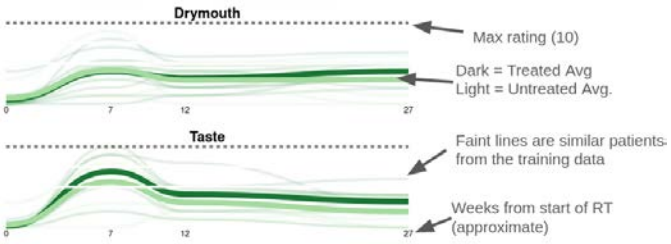

Fig. 7: Symptoms prediction for a patient. Dark green indicates average of patients that receive a selected treatment, light green is average of patients that don't receive treatment. Faint lines indicate trajectories of the cohort patients used to make the prediction.

Finally, because several oncologists expressed a desire to see the effect of treatment on long-term subclinical side effects, we include a KNN-based symptom progression model for the patient (Fig. 7) (T1.5). Due to data constraints, this view only includes a neighbor-based model, with no deep-learning based model as the cohorts were different and we were unable to get a sufficiently accurate model. This view shows self reported symptom progression for 10 different symptoms for a period of 6 months after the start of radiation treatment. Each similar patient is shown as a faint line, and group median values for treated and untreated groups are shown as bold lines. Symptoms are ordered by mean rating at the end of the time-period, as clinicians are most interested in long-term side effects that are more likely to be permanent.

5 Qualitative Evaluation

A quantitative evaluation of the models used in the system is included in the supplementary materials. To further evaluate DITTO, we performed two case studies with two users: one HNC clinician with 9 years of experience, 4 years of which were at the MD Anderson Cancer center, along with one Data Mining researcher, to find out how oncologists interact with the system. The case studies covered the evaluation of a single patient each and were performed via Zoom meetings with desktop sharing. To assess how different model recommendations might affect the users, we selected one patient that had both the neighbor-based and DT model agree with the true patient recommendation (non-counterfactual) and a case where the neighbor-based and DT disagreed with each other (counterfactual). The policy model strategy was set to "Imitation" based on clinician preference. Qualitative feedback was collected via a debriefing interview derived from the System Usability Scale [7] structure.

5.1 Typical Recommendation

Our first case study was taken from an example patient where both the neighbor-based and Imitation policy model agreed with the clinical ground truth. Starting with the patient input, the patient was notable for having a high T-stage (large primary tumor) and "Not Otherwise Specified" tumor location, suggesting that the patient had a large, irregularly positioned tumor, and being African American. The clinician then moved to the treatment recommendation (Fig. 8-A) to confirm that the model recommendation lined up with the similar patients in the cohort, where 100% of patients receive chemotherapy. Looking at the feature importances for the policy model recommendation (T3.1),

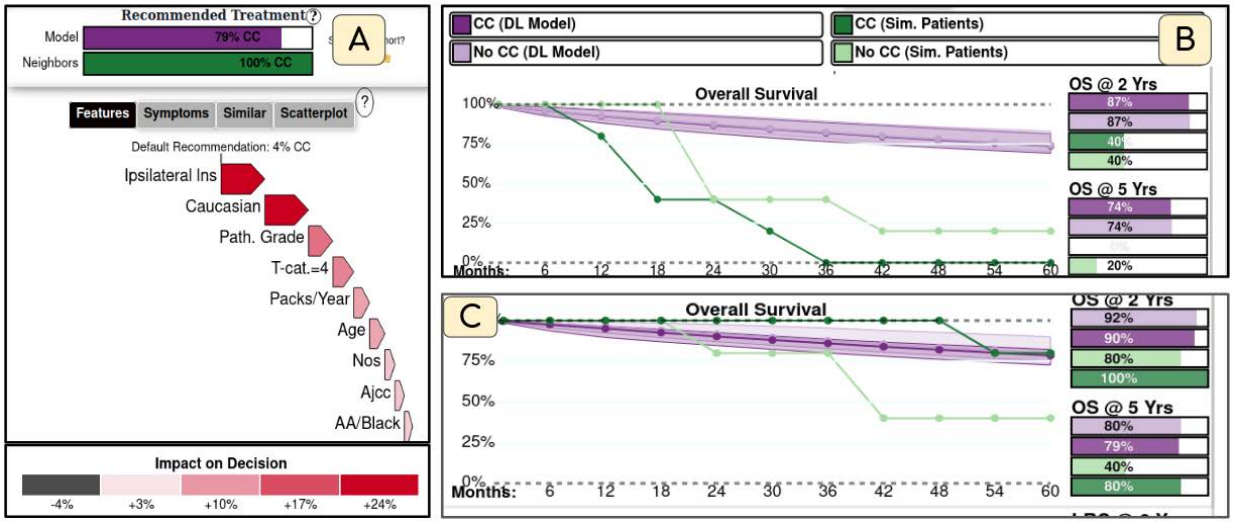

Fig. 8: First case study. (A) Feature importance and recommendation (truncated) showing that LN spread and Race are the main predictors of the patient receiving CC. (B) Patient survival curves. Green lines show very low survival for both treated (dark green) and untreated (light green) groups, but high survival from the DSM model (purple). (C) Survival curves for the patient when their race is changed to "white/caucasian". Similar patients have much higher survival rates.

they noted that the most prominent features are the LN spread, the patient's race, the pathological grade, and the T-staging. While this finding mostly lined up with clinical reasoning, the impact of race was surprisingly high (+33% chance of CC).

In the survival outcomes (Fig. 8-B), they noted that there was a large discrepancy between the predicted survival, and those reported by the cohort (T3.4): only 40% of similar patients survived 2 years (T1.2), and none of the treated group survived 5 years despite a predicted survival rating of 89% with high confidence (T1.1, T3.3). Interestingly, outcomes were better in the untreated group. Looking at the similar patients, we could see higher T-stage and pathological grade in the CC group, which may account for the difference, although it was unclear if race also impacted this (T1.4).

Looking back at the issue of race, the group tested this patient by changing only their race to "white/Caucasian". Indeed, this changes both the predicted treatment from the Deep Policy model (73% no cc), and the patient outcomes, with significantly higher rates of predicted survival in the similar patients (Fig. 8-C) (T3.2). This led to a discussion on the use of race in the model, where we discussed issues of bias and confirmed that, indeed, race has an impact on physician treatment and patient outcomes, which requires further study [41]. Interestingly, the clinician also tested the "optimal" policy model, which only showed a minimal impact of race (< 1%) on the recommended treatment, which was no CC. Notably, the optimal policy model listed the low pathological grade, tumor location, and AJCC stage as reasons to not give CC, while the LN spread is given as the primary reason to give CC. Based on the predicted outcomes, we noticed a much higher risk of side-effects (5.9% chance increase in feeding tube and 3.6% chance increase in Aspiration), with non-significantly higher predicted chance of tumor response or control, which led to the no-CC recommendation (T2.2, T2.3), as well as slightly higher incidence of severe symptoms in the symptom plot for the CC group (T1.5).

5.2 Counterfactual Recommendation

In this second case study, we examined a patient where the Deep Policy Model predicted no CC, while the most similar patients all received CC. In this case, the group noticed that patients had relatively low staging and low smoking, which the clinician confirmed lined up with the patient not needing CC in most cases (T1.3). They speculated the difference may be due to physician preference or other factors, such as features not accounted into the model but present in the lab notes (e.g., the patient having only one kidney). Additionally, they noted that in this case, the existing guidelines cite smoking and Lymph node levels as the main causal factors. We can see in the input LN diagram that the patient had LN levels II effects (Fig. 9-B, left), the most common levels, which have an impact on an increased chance of CC.

Noting in the outcome panel that there is a relatively high chance of survival for both groups given the low risk (T1.1, T1.2), and similar risk profile for treated and untreated groups (T2.2), the clinician moved to the similar patient panel. Notably, both treated and untreated groups had similar characteristics, but the untreated group actually had more nodal extension to level 3, higher staging, a higher average smoking rate, and lower survival and tumor control after 4 years (Fig. 9-A) (T1.4). They noted that this may confirm that the difference in the cohort treatment may be due to the physician or other factors.

Moving to the input panel, the clinician tested the impact of the two changes given by the physician: lymph node extension to level IV, and smoking > 20 pack-years, and confirmed that with these changes the model indeed changed to predict CC (52% chance), and that LN level IV was a major factor in the change in treatment (Fig. 9-B) (T3.1).

5.3 Qualitative Feedback

Feedback from HNC clinicians at the MD Anderson Cancer Center was very positive, stating that the system was "really attractive" and "amazing". When asked about their favorite features of the interface, multiple participants stated that they liked the views of similar patients, as well as symptom progression in the auxiliary panels. They also felt that many clinicians would be more interested in just the outcomes in the center. In response to this feedback, we turned the additional panels into a separate on-demand view. The participants also found the feature attributions interesting, saying "I also like the neighborhood panel and the multiple outcomes together". Some were particularly interested in the lymph node involvement levels for similar patients, as well as how this relates to feature importance in the policy model. They were also able to identify possible sources of data bias in the predictions by looking at the treated and untreated groups. When asked about the usefulness of the simplified three-decision model, the most senior clinician commented that "The 3 decision points are critical decision points, aligned with the standard of care. We could get more granular, but it's a great start."

6 Discussion

Our results show that DITTO is an effective tool for treatment planning for HNC clinicians using a novel Digital Twin system. Clinician feedback was very positive, with a variety of "favorite" components

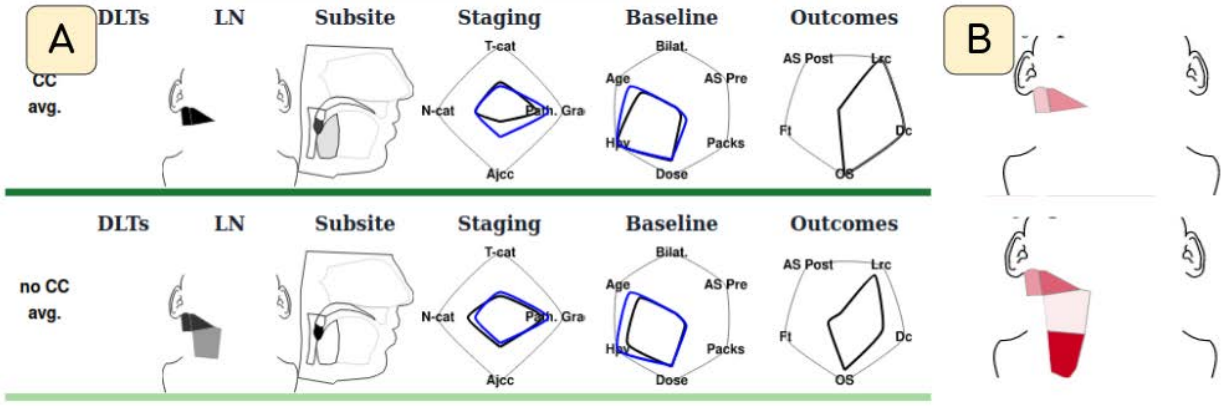

Fig. 9: Second case study. (A) Average of treated and untreated groups. Treated patients have lower LN spread and staging, and higher survival rates, which is counter-intuitive. (B) Feature attributions for ipsilateral LN levels II (top) vs levels II-IV (bottom), where dark red encodes higher likelihood of receiving CC. Changing the patient to have LN level IV involvement significantly increases confidence that the patient should receive CC.

and background, suggesting that DITTO can handle a variety patient treatment goals. Additionally, while we had initial concern that the use of two model outputs would prove confusing, our case studies show that investigating model discrepancies indeed leads to interesting discussion into how patients should be treated and how physician preference or uncounted variables may impact certain recommendations.

6.1 Design Lessons

A majority of explainable ML work has focused on visualizations meant for model builders or clinical researchers to use in research context, while most clinician-facing systems focus on relatively simple models [11, 31, 54]. In this regard, this system is a novel attempt to deliver the results of a complex Digital Twin system to clinical end users. In particular, this work focused on two challenges: delivering many potential results to clinicians in a way that allows them in a way that is relatively accessible, and to find a way to balance encouraging oncologists to use the system while not overly relying on potentially incorrect predictions. We list here the specific design insights we've developed during this participatory design process.

L1. Use visual scaffolding. Our users were clinicians who had experience with risk modeling visualization. We found our best results by scaffolding, such as relying on temporal plots and spatial anatomical diagrams. Previous attempts at novel encodings such as histograms or unique glyphs were less successful with wider audiences.

L2. Account for different information-seeking needs. We found in our interviews and literature reviews that the degree of information seeking behavior, as well as attitudes towards different models, varied greatly between clinicians. For example, we found that some users were completely uninterested in seeing similar patients or a scatterplot of the training cohort, while others listed the similar patients as their "favorite" part of the system and were able to identify potential confounder bias by looking at the similar patients. We also found that many users were primarily interested in seeing only the recommended treatment and time-to-survival, so this information could be communicated to patients, and felt additional features were distracting in the interface. As a result, we altered our design to afford these additional features in secondary tabs, and allowed for resizing of the different parts of the interface, while highlighting only the survival plots by default.

L3. Provide access to multiple models, and cues such as counterfactuals and confidence intervals to balance user expectations of the model. In using our system, we found that clinicians have a tendency to either fully trust or distrust a model in the absence of additional cues, and expressed a desire for "honesty" in terms of model confidence. To encourage users to "think slowly" [22, 24] about the model predictions, we relied on multiple cues: showing different model predictions side-by-side, using model confidence intervals when available, and placing the feature attribution plot prominently in the visualization. Still, there is necessarily a design tradeoff between interface simplicity, user acceptance of the model, and the number of additional cues. While many lay-users may prefer only being given a single prediction, we consider this a questionable design tradeoff with respect to XAI. As a result, we initially show users all model predictions, with the option to toggle off information.

6.2 Limitations and Future Work

In terms of limitations, our current models are limited by data availability and model performance. Our dataset requires modeling 19 different outcomes and transition state variables while relying on less than 600 patients from a single institution with limited demographic diversity. Furthermore, a more granular digital twin system could consider multiple rounds and dosages of chemotherapy and surgery. Our available dataset is also specific to oropharyngeal HNC patients. In terms of our interface, we focus on a limited group of HNC clinicians, a few of whom may have above-average visual literacy and information seeking behavior.

Our imitation learning model inherits existing treatment biases, and diversity shortcomings in the training data. While the initial goal of the system is to reveal these biases as shown in our case studies, there is the potential for users to interpret these explanations as justification for biased reasoning when over-trusting the system. Regardless, this bias should not be reflected in the risk prediction or optimal model, which should theoretically contradict the treatment recommendation in such a case.

In terms of generalizability, the general approach can be applied to any similar treatment sequence that can be simplified into discrete decision stages, and a majority of our visualization system is domain-agnostic, except for the tumor subsite and lymph node spread diagrams. Regarding visualizations, our algorithms for uncertainty, feature attribution, and similarity are specific to deep learning classification, but these values can be obtained more generally through bootstrapping, Shapley values, and appropriate distance metrics, respectively, and can be visualized in the same way.

7 Conclusion

In conclusion, we have implemented a visual clinical decision support system based on a temporal deep-reinforcement learning model that is capable of simulating patient treatment outcomes. To our knowledge, this is one of the few attempts at an explainable AI focused interface for clinical users, as well as one of the first attempts at a visual interface to explore a dual digital twin system in a healthcare setting. Through our participatory design, we highlight several findings with a focus on balancing information density, usability, and encouraging appropriate trust for a variety of of end users. In our future work, we hope to evaluate this interface on a large range of clinical end users in practice, as well as extend our work to even more detailed decision-making that can consider more patient quality-of-life measures.

Appendix A contains additional details about the model (Appendix A.1) and an evaluation of the deep learning models, along with details about the data cohort (Appendix A.2).
Appendix B contains additional figures of the images of related content for the system (Appendix B.1) and images of earlier prototypes (Appendix B.2).

A Appendix A: Model Details and Evaluation

A.1 Model Details

A.1.1 Patient Simulator Models

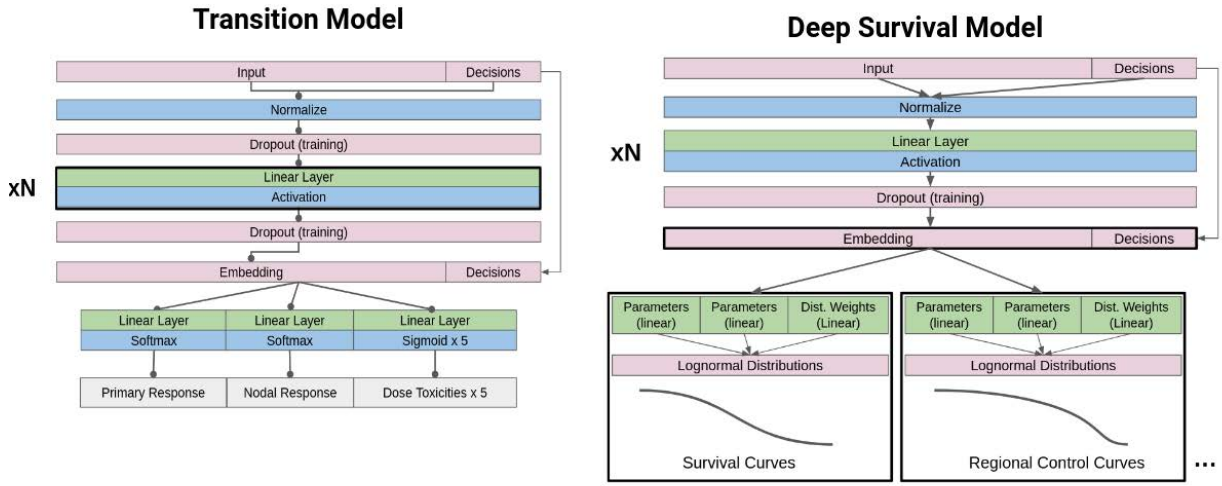

Fig. 10: Architecture for the transition and deep survival models (DSM). Patient state and previous state treatment decision use a standard DNN with input dropout to improve the models ability to deal with unknown data. The decision is concatenated to the penultimate layer in order to prevent the model from relying only on correlated features due to the use of dropout during training. DSM models predict a mixture of model parameters for each patient from a pre-trained set of user-defined number of mixtures.

To simulate the patient, we use a set of models to mimic intermediate response to treatment (transition models), and long-term response after treatment (outcome models). Fig. 10 Shows the architecture for the transition models and deep survival machines used to model temporal patient outcomes (DSMs). Each time-point uses a separate transition state model. For induction chemotherapy, we constrain the model to not allow for any tumor response when no chemotherapy is given, as this would indicate no treatment at this point.

Transition models predict patient response to treatment in terms of tumor shrinkage and severe toxicities from treatment. Specifically, we consider primary disease response (PD), and nodal disease response (ND), which are each 4 categorical ordinal variables, as well as 5 binary results for different types of dose-limiting toxicities (DLTS). For the case of Induction chemotherapy, disease response is always assumed to be stable when no treatment is done. Seperate models are trained for post-IC and post-CC transitions, as this resulted in better performance.

For the outcome model, two separate models are used. The first is a deep neural network that predicts toxicity risk using binary variables: Aspiration (AS), and Feeding Tube (FT) at 6 months after treatment.

The second outcome model predicts cumulative patient risk over time for overall survival (OS), locoregional control (LRC), and distant metastases (FDM) for up to 5 years. Temporal risk models use a variant of deep survival machines (DSM) [42]. For all three outcomes, the DSM model returns a mixture of parametric log-normal distributions for the patient that can be used to provide a cumulative survival risk over time.

Because clinicians listed confidence intervals as important for reasoning about the model predictions (T3.3), all transition and outcome models are trained using dropout on the penultimate layer between 50% and 75%. During evaluation, we re-run each prediction with random dropout at least 20 times, and then save the 95% confidence intervals for each prediction. [17].

All models implemented in pytorch and trained using the Adam optimizer. Models were trained using early stopping until the validation loss stopped increasing for at least 10 epochs. Transition models, static outcome models, and Deep Survival Machines for temporal outcomes used a dropout of 10% on the input layer and 50% on the penultimate layer during training. Transition models and static outcome models used 2 hidden layers with an output size of 500 each. The deep survival machines used a single hidden layer with a size of 100 and 6 different distributions for each outcome.

A.1.2 Policy Models

The patient simulator models and ground truth responses are used as the environment to train a digital physician (policy model) (Fig. 11). Because there is disagreement among users as to whether they prefer to see what a physician would do, or what the "best" choice should be, we jointly train two versions of the policy model: one that minimizes a combination of patient risks based on the patient simulator responses (optimal policy model), and one that predicts what a physician would do based on the cohort data (imitation policy model).

Each policy model (optimal and imitation) is trained using a dual loss function: prediction of the ground truth (or optimal) decision sequence, and triplet loss. Triplet loss is included as it was found to increase model performance in terms of AUC and accuracy for the imitation model. We use AUC as it is a measure of the relative ranking of patient risks, and is thus commonly used to identify rarer events. Specifically, the loss for a given patient p at each epoch for a given output (optimal or imitation) is given by:

$$L(p) = w_1 \cdot \sum_{i=0}^{2} BCE(\hat{y_{p,i}}, y_{p,i}) + w_2 \cdot max\big(d(a_p, b_p) - d(a_p, c_p) + 1, 0\big)$$

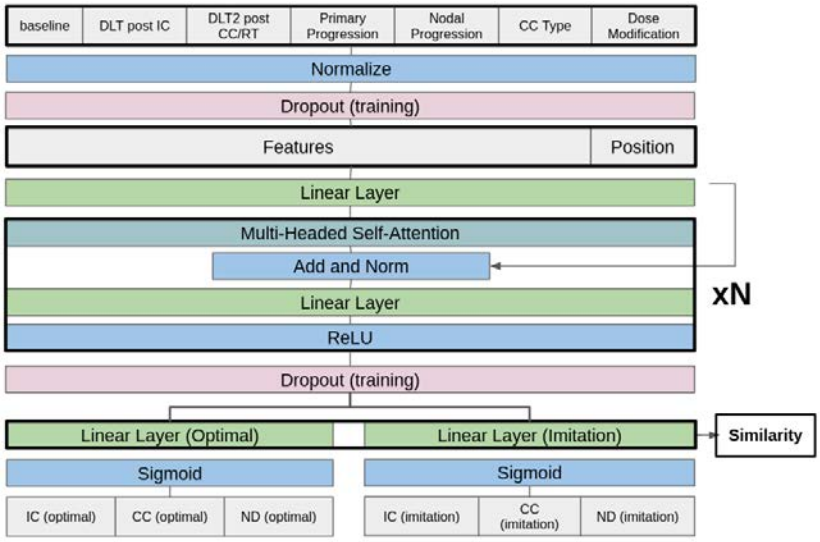

Fig. 11: Architecture for policy model used to simulate a physician decision. Both the optimal and imitation models use a shared embedding with a custom position token at each stage, followed by a separate layer for each output, with additional fully connected layers unique to each model before the output. Model activations for the penultimate layers are used when calculating similar patients. Policy models use a modified version of a transformer encoder that saves the cohort at each time point into memory at training time.

Where y_i and $\hat{y}_i$ are the ground truth and predicted decisions, respectively. $d(\cdot,\cdot)$ is the euclidean distance. a_p is the final hidden layer weight vector for the patient. b_p are the hidden weights for a randomly sampled patient with the same ground truth treatment sequence, and c_p is a randomly sampled patient with a different treatment sequence. w_1 and w_2 are user-decided weights. For our implementation we use $w_1 = 1$ and $w_2 = .2$ for both outputs.

Both optimal and imitation policy models use shared layers until the penultimate layers, which are unique to each output, and are re-trained each epoch (Fig. 11). This allows for joint learning of important features from each other. To encourage our model to explicitly consider other patients in the cohort, our policy model architecture uses a transformer encoder and uses a position token to encode the temporal state of the patient. During evaluation on a new datapoint, the cohort data for the current state is used as the Query input of the multi-headed attention as described in Vaswani et al. [56].

Imitation policy model decisions are trained using the unaltered ground truth states in the data to predict the decision made by the clinician. The optimal model decision is, in contrast, trained on using random data augmentation on the pre-treatment variables for the patient for each epoch. Specifically, each column (feature) has a 25% probability of being pseudo-randomly shuffled in the training sample, and the predicted patient response to treatment using the deep learning transition models.

When determining the treatment sequence for the optimal decision, we calculate the decisions that minimize a combination of all predicted outcomes, given by:

$$L = w_{tox} \sum_{z \in Z} w_z P(z = 1) + w_s \sum_{o \in O} \frac{w_o}{\tilde{T}(o)}$$

Where $z \in Z$ is the set of binary outcomes (e.g toxicity, 4 year survival, 4 year locoregional control), $o \in O$ is the set of temporal survival outcomes (survival, locoregional control, distant control), $\tilde{T}(o)$ is the median predicted time-to-event of outcome o, and $w \in W$ a set of user defined weights for each aspect of the loss function.

Because we need to explain the policy model recommendations (T3), we use integrated gradients [48] to obtain feature importance for each decision relative to a baseline value. Integrated gradients was chosen as it satisfies the completeness axiom where attributions sum to the difference in the prediction between the baseline and actual recommendation, which was found to be easier to reason about with our clinicians. For our baseline, we assume the lowest possible rating for most ordinal attributes such as tumor staging or disease response, and the most common value for categorical attributes such as gender, ethnicity, and tumor subsite, as well as age and dose to the main tumor, based on feedback from clinicians and what they found most intuitive.

All models implemented in pytorch and trained using the Adam optimizer. Models were trained using early stopping until the validation loss stopped increasing for at least 10 epochs. Our policy model used an input dropout of 10% and 25% dropout on the final layers, with a single transformer encoder of size 1000 for the joint embedding, and a linear layer of size 20 for both the optimal and imitation model outputs.

A.1.3 KNN-based Symptom Prediction

Our symptom prediction model uses a different cohort of patient and relies on a KNN predictor using the embeddings taken from a model trained to predict symptom trajectories. Specifically, we trained a fully connected deep learning model to predict symptom ratings for each symptom and each time point in the data. Time points considered were at 0, 7, 12, and 27 weeks after starting radiation therapy. Outputs were treated as independent values with a sigmoid loss function that was scaled to be between 0 and 10. Input features were gender, packs-years, HPV status, treatment dose and dose frction, race, tumor laterality, tumor subsite, T-category, N-category, and treatment decisions for induction chemotherapy (IC) and concurrent chemotherapy (CC).

Patient embeddings for the cohort were taken from the model activations in the batch-normalized penultimate layer in the deep learning model. When predicting a new patient, we take the new patient's model embeddings and extract the 10 most similar patients, based on euclidean distance, from the embeddings of both the treated and untreated patients, respectively. Patient symptom profiles are taken for these patient seperately

During deep learning model training we used an 80/20 train validation split on the data for parameter tuning, using the mean-squared-error loss (MSE). Missing symptom values were ignored in the loss function. All models were trained using the ADAM optimizer in pytorch using early stopping on the validation loss. Our final model used a single hidden layer of size 10 with the ReLu activation, followed by batch normalization, with no dropout.

A.2 Model Evaluation

536 patients were used to evaluate our system. The dataset was split into a training cohort of 389 patients and an evaluation cohort of 147 patients before beginning the development of the models. For evaluation purposes, the training sample was stratified in order to get a minimum of 3 patients with each endpoint, and treatment decision in the model. Because we could not achieve enough samples of patients with several dose limiting toxicities, all toxicities that were not present in both cohorts were aggregated into an "other" category for the purpose of modeling and evaluation. The features used for the entire cohort, excluding lymph node patterns, is shown in (Tab. 2), stratified by treatment sequence. An anova F-test was used to analyze correlations between each feature set and the treatment sequence, and p-values are included in the table.

Performance of the policy model with and without triplet loss is shown in (Appendix A.2). We see an increase in imitation model performance, with a slight decrease in "optimal" model performance for accuracy but increase in AUC. This is likely due to the heavy imbalance in the optimal outcomes: only 10.8% of cases recommended concurrent chemotherapy and 19% of cases recommended neck dissection, as rare events were predicted with higher prediction confidence. Given that a majority of users preferred to use the "imitation" model, the triplet model was used in practice.

In general, AUC tended to perform better than Accuracy in the optimal model, likely due to the heavy imbalance in the optimal outcomes: only 10.8% of cases recommended concurrent chemotherapy and 19% of cases recommended neck dissection. In general, model performance is comparable to similar outcome models from earlier studies, considering the added difficulty of optimizing for 23 different outcomes and 6 treatment decisions. Interestingly, our optimal model suggested induction chemotherapy followed by radiation alone a majority of the time, which contradicts the standard practice where concurrent chemotherapy is standard while induction is used for patients with very large tumor spread that needs to be reduced before applying radiation. However, the data is largely limited by confounders and lack of detailed information on how changes in patient's health affect treatment and outcomes. Additionally, we have been told that the specific grade of dose-limiting toxicity is an important factor in treatment and side effects, which our model does not consider.

Performance of transition models are shown in (Appendix A.2). Because the outcomes we want to predict are often rare events, we compared default training performance with basic cross-entropy loss with a balanced loss function. Non-balanced models generally performed better in terms of AUC with similar accuracy.

To evaluate time series models, we calculate F1 and ROC AUC scores at 12, 24, 36, and 48 months after treatment (Tab. 4). We exclude longer periods as we tend to have fewer followup data available after 48 months. OS, FDM and LRC models tend to have high F1 score but modest AUC scores, possibly due to the fact that failures are rare events in the data.

Decision	Optimal		Imitation	
	AUC	Accuracy	AUC	Accuracy
With Triplet Loss				
IC	0.84	0.58	0.79	0.88
CC	0.97	0.73	0.93	0.78
ND	0.95	0.79	0.90	0.81
No Triplet Loss				
IC	0.82	0.71	0.60	0.87
CC	0.96	0.91	0.74	0.78
ND	0.94	0.88	0.84	0.81

Table 1: Physician Simulator Policy Model Performace with and without use of triplet loss.

Treatment Sequence	CC	None	CC + ND	IC + CC	IC + CC + ND	IC	ND	IC + ND	P-Value
Count	223	57	51	100	36	45	11	13	1
HPV+	56.50%	80.70%	54.90%	50.00%	61.11%	42.22%	54.55%	61.54%	6.44E-03
HPV Unknown	6.28%	1.75%	7.84%	16.00%	11.11%	2.22%	18.18%	7.69%	
Age (Mean)	59.3	61.3	57.7	58.5	58.3	57.6	59.6	57.0	4.92E-01
Pack-years	17.6	10.5	18.9	17.6	21.8	15.4	16.7	4.8	1.83E-01
Male	88.34%	80.70%	92.16%	87.00%	91.67%	88.89%	81.82%	92.31%	6.76E-01
Smoker	19.28%	19.30%	35.29%	22.00%	22.22%	24.44%	18.18%	0.00%	2.38E-01
Former Smoker	42.15%	40.35%	29.41%	34.00%	33.33%	33.33%	54.55%	30.77%	
Bilateral	4.48%	3.51%	5.88%	4.00%	2.78%	2.22%	0.00%	0.00%	9.50E-01
T-category_1	18.83%	63.16%	5.88%	6.00%	13.89%	28.89%	54.55%	30.77%	1.30E-18
T-category_2	42.15%	33.33%	54.90%	33.00%	27.78%	48.89%	45.45%	61.54%	4.45E-02
T-category_3	24.66%	3.51%	21.57%	29.00%	27.78%	17.78%	0.00%	7.69%	3.25E-03
T-category_4	14.35%	0.00%	17.65%	32.00%	30.56%	4.44%	0.00%	0.00%	6.69E-08
N-category_1	52.91%	80.70%	52.94%	27.00%	16.67%	22.22%	63.64%	61.54%	4.96E-14

N-category_2	39.46%	12.28%	43.14%	65.00%	75.00%	73.33%	27.27%	38.46%	7.13E-14
N-category_3	1.79%	0.00%	0.00%	8.00%	8.33%	4.44%	0.00%	0.00%	1.91E-02
AJCC_2	15.25%	5.26%	15.69%	16.00%	22.22%	22.22%	9.09%	7.69%	1.66E-16
AJCC_3	9.42%	5.26%	13.73%	22.00%	25.00%	0.00%	18.18%	0.00%	2.10E-04
AJCC_4	36.77%	12.28%	39.22%	49.00%	38.89%	57.78%	18.18%	38.46%	5.90E-05
subsite_BOT	50.22%	35.09%	47.06%	56.00%	55.56%	57.78%	18.18%	46.15%	7.85E-02
subsite_GPS	0.90%	1.75%	1.96%	2.00%	8.33%	0.00%	0.00%	7.69%	7.50E-02
subsite_Soft palate	0.90%	1.75%	3.92%	1.00%	0.00%	0.00%	0.00%	0.00%	6.48E-01
subsite_Tonsil	41.26%	54.39%	41.18%	36.00%	33.33%	40.00%	81.82%	30.77%	4.78E-02
Pathological Grade_1	0.90%	0.00%	0.00%	3.00%	0.00%	0.00%	9.09%	0.00%	1.04E-01
Pathological Grade_2	28.25%	31.58%	27.45%	28.00%	33.33%	28.89%	45.45%	7.69%	6.63E-01
Pathological Grade_3	50.67%	54.39%	56.86%	48.00%	55.56%	46.67%	36.36%	61.54%	8.39E-01
Pathological Grade_4	0.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%	5.00E-02
White/Caucasion	93.27%	89.47%	96.08%	86.00%	86.11%	93.33%	90.91%	92.31%	3.57E-01
Aspiration Pre-Therapy	2.24%	0.00%	1.96%	7.00%	5.56%	2.22%	0.00%	0.00%	2.14E-01
Total Dose (gy)	68.99	66.86	69.47	69.36	69.33	67.42	68.05	67.23	1.05E-14
Dose Fractions	2.10	2.16	2.08	2.11	2.08	2.15	2.17	2.18	7.82E-05
Survival (Months)	76.26	71.24	80.52	74.10	74.47	87.80	98.57	97.21	1.76E-02
Locoregional control (Months)	74.46	67.63	67.28	71.36	63.85	86.54	72.10	94.07	2.50E-02
FDM (months)	74.69	71.00	77.95	72.38	69.56	84.34	94.86	97.21	5.02E-02
Overall Survival	75.34%	82.46%	70.59%	75.00%	61.11%	86.67%	63.64%	100.00%	4.36E-02
Locoregional Control	91.03%	92.98%	68.63%	85.00%	69.44%	91.11%	63.64%	84.62%	1.74E-05
FDM	89.69%	96.49%	86.27%	90.00%	80.56%	88.89%	72.73%	100.00%	1.25E-01
FT	17.49%	5.26%	21.57%	25.00%	38.89%	8.89%	18.18%	0.00%	4.80E-04
Aspiration Post-Therapy	17.49%	3.51%	25.49%	22.00%	41.67%	8.89%	18.18%	7.69%	1.87E-04
CR Primary	0.00%	0.00%	0.00%	34.00%	38.89%	66.67%	0.00%	46.15%	2.99E-50
PR Primary	0.00%	0.00%	0.00%	52.00%	55.56%	28.89%	0.00%	30.77%	2.02E-51
CR Nodal	0.00%	0.00%	0.00%	10.00%	2.78%	11.11%	0.00%	0.00%	1.79E-06
PR Nodal	0.00%	0.00%	0.00%	75.00%	88.89%	86.67%	0.00%	84.62%	1.21E-148
DLT after CC	0.00%	0.00%	0.00%	75.00%	50.00%	64.44%	0.00%	84.62%	0.00E+00
CR Primary 2	83.41%	91.23%	68.63%	90.00%	72.22%	91.11%	81.82%	92.31%	4.86E-03
PR Primary 2	16.14%	7.02%	25.49%	10.00%	19.44%	4.44%	18.18%	7.69%	3.56E-02
CR Nodal 2	52.02%	52.63%	17.65%	58.00%	19.44%	57.78%	9.09%	7.69%	1.38E-09
PR Nodal 2	43.95%	35.09%	80.39%	34.00%	77.78%	37.78%	90.91%	69.23%	4.46E-11
DLT After CC/RT	27.80%	0.00%	15.69%	29.00%	25.00%	0.00%	0.00%	0.00%	2.95E-07

State	Outcome	Metric	Value
After IC	Primary Response	accuracy	0.404
	Primary Response	auc_micro	0.801
	Primary Response	auc_weighted	0.674
	Nodal Response	accuracy	0.333
	Nodal Response	auc_micro	0.853
	Nodal Response	auc_weighted	0.533
	Dose Modification	accuracy	0.333
	DLT_Gastrointestinal	accuracy	0.804
	DLT_Other	accuracy	0.946
	DLT_Dermatological	accuracy	0.893
	DLT_Hematological	accuracy	0.786
	DLT_Neurological	accuracy	0.911
	DLT_Gastrointestinal	auc	0.497
	DLT_Other	auc	0.415
	DLT_Dermatological	auc	0.420
	DLT_Hematological	auc	0.511
	DLT_Neurological	auc	0.557
After RT + CC	Primary Response	accuracy	0.333
	Primary Response	auc_micro	0.887
	Primary Response	auc_weighted	0.568
	Nodal Response	accuracy	0.372
	Nodal Response	auc_micro	0.756
	Nodal Response	auc_weighted	0.545
	DLT_Gastrointestinal	accuracy	0.918
	DLT_Other	accuracy	0.980
	DLT_Dermatological	accuracy	0.966
	DLT_Hematological	accuracy	0.952
	DLT_Neurological	accuracy	0.966
	DLT_Gastrointestinal	auc	0.564
	DLT_Other	auc	0.727
	DLT_Dermatological	auc	0.625
	DLT_Hematological	auc	0.613
	DLT_Neurological	auc	0.552
After All Treatment	Feeding Tube	accuracy	0.803
	Feeding Tube	auc	0.683
	Feeding Tube	f1	0.216
	Aspiration Post-therapy	accuracy	0.803
	Aspiration Post-therapy	auc	0.775
	Aspiration Post-therapy	f1	0.065

Table 3: Model Performance for all transition states and toxicity

outcome	months	metric	value
OS	12	AUC	0.52
		F1	0.99
	24	AUC	0.63
		F1	0.96
	36	AUC	0.64
		F1	0.95
	48	AUC	0.60
		F1	0.94
Locoregional Control	12	AUC	0.64
		F1	0.97
	24	AUC	0.56
		F1	0.94
	36	AUC	0.57
		F1	0.93
	48	AUC	0.57
		F1	0.92
Distant Control	12	AUC	0.42
		F1	0.98
	24	AUC	0.62
		F1	0.95
	36	AUC	0.62
		F1	0.93
	48	AUC	0.57
		F1	0.92

Table 4: Model Performance for deep survival machines at 12, 24, 36, and 48 months.

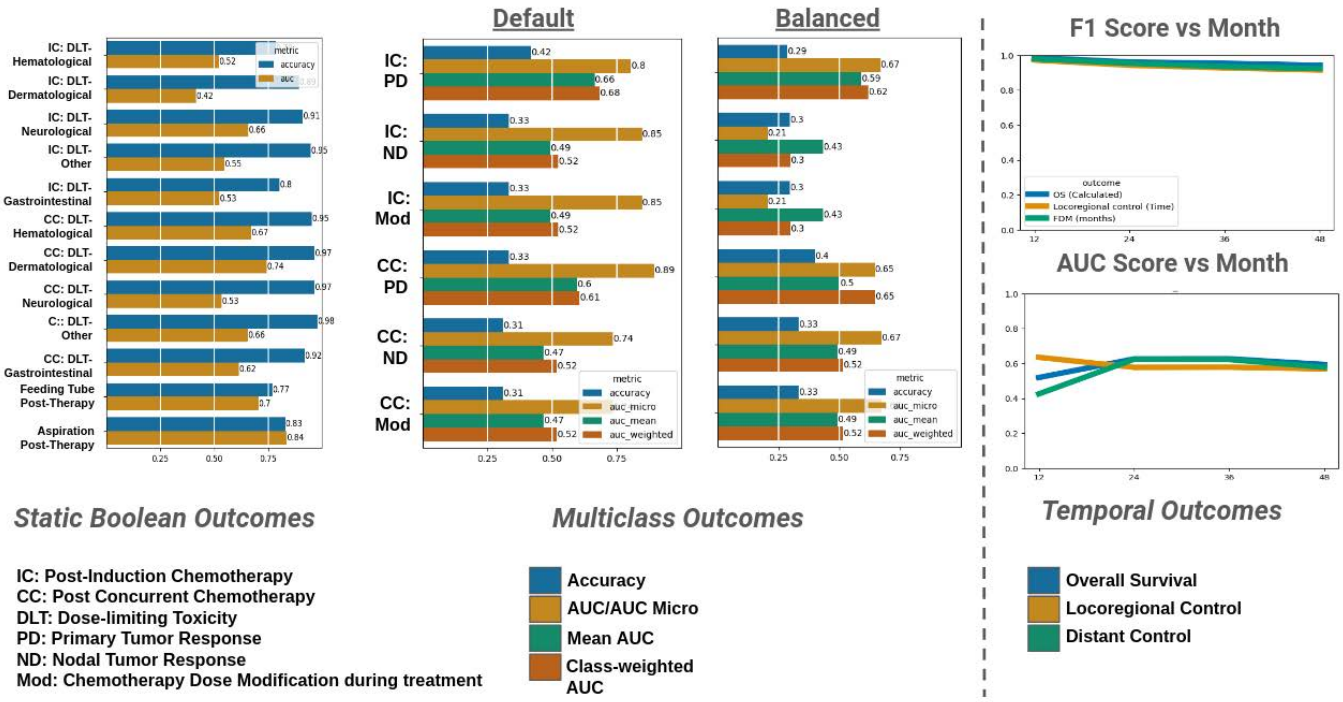

Fig. 12: All Transition State Outcomes. (Right) Accuracy and AUC score for boolean outcomes such as toxicities. Models perform well in terms of accuracy and late toxicity (FT and Aspiration), but have mixed AUC results for dose-limiting toxicities due to the heavy imbalance in the data and low number of positive samples to learn from. (Center) Model performance for multi-class transition states (disease response and dose modification) using accuracy and micro, macro, and weighted AUC score for both unweighted and balanced loss weights. Models perforce best in terms of macro AUC score. Balanced models generally performed worse. (Right) F1 score and AUC score for temporal outcomes at 12, 24, 36, and 48 months after treatment. F1 scores tend to be very high while AUC scores stay around .6, likely due to issue with imbalanced data and incomplete censoring.

B Appendix B: Prototypes and Additional Figures

Appendix B.1 Shows additional figures from the interface that were removed for space. Appendix B.2 Shows earlier designs of the interface. Fig. 16 Shows a diagram of the user workflow the system was design around.

B.1 Additional Figures

Fig. 13 Shows a scatterplot in the interface of the patient policy model embeddings, which was used during model development to explore the predictions in the cohort and find example patients to test the interface on. The scatterplot is not in the main manuscript as it does not contribute to the clinical user goals. Fig. 14 shows the spatial feature diagrams for Dose-limiting toxicities, lymph node levels, and tumor subsites used in the interface. Fig. 15 shows the full patient input panel. Fig. 16 Shows a diagram of the user workflow for the system. Fig. 17 Shows the psuedocode for the ATE patient sampling algorithm.

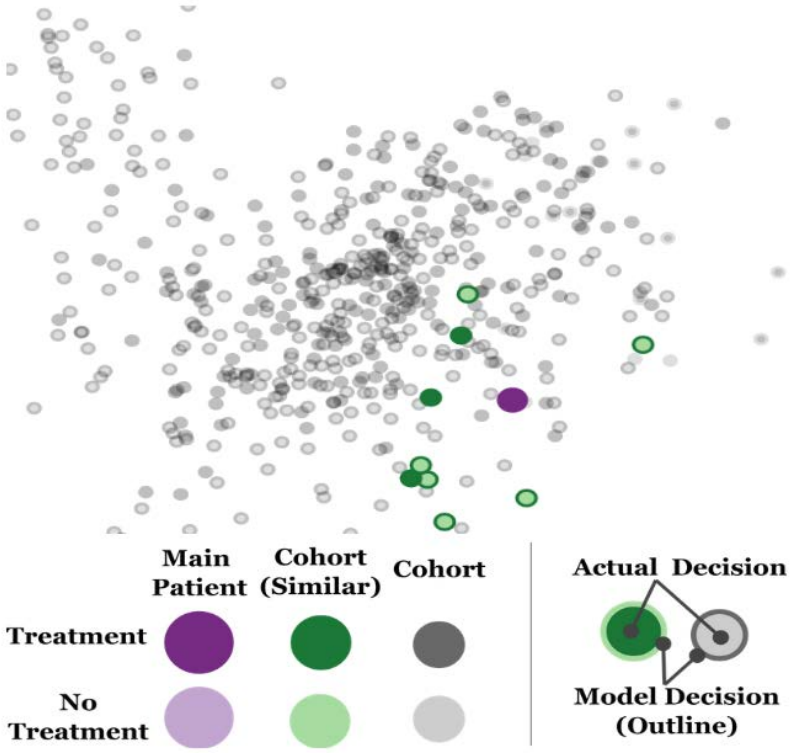

Fig. 13: Scatterplot used during model development using the 2 principal components of the policy model embeddings of the cohort and main patient. Outer and inner color encodes model predicted treatment and ground truth treatment, respectively. Hue is used to differentiate the current patient, the most similar patients, and the rest of the cohort.

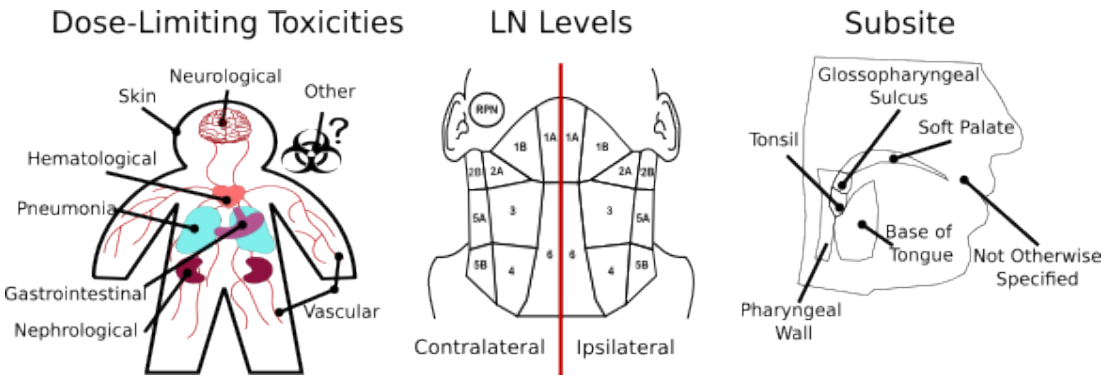

Fig. 14: Diagrams used for spatial features in the visualization. (Left) Dose-limiting toxicities. (Center) Lymph node regional involvement. (Right) Primary tumor subsite. All regions not included in the diagram are considered "Not Otherwise Specified".

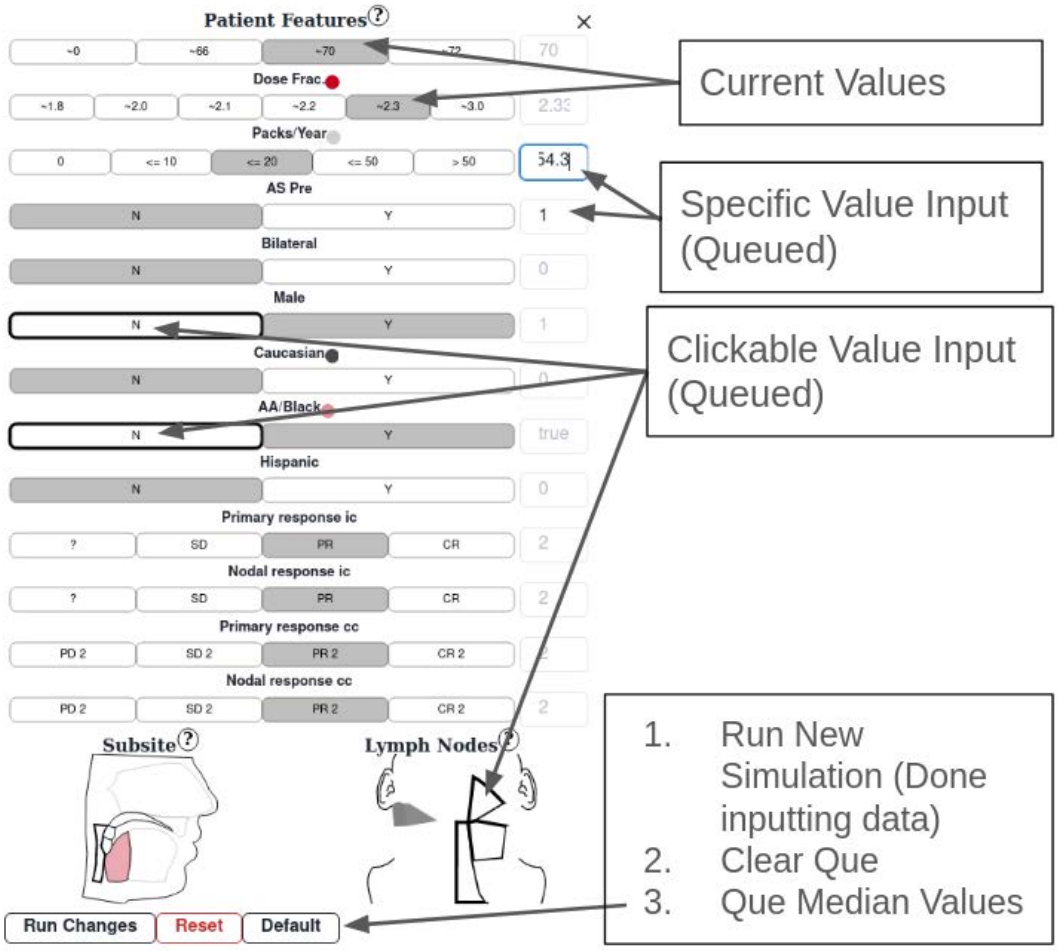

Fig. 15: Full view of the user input panel

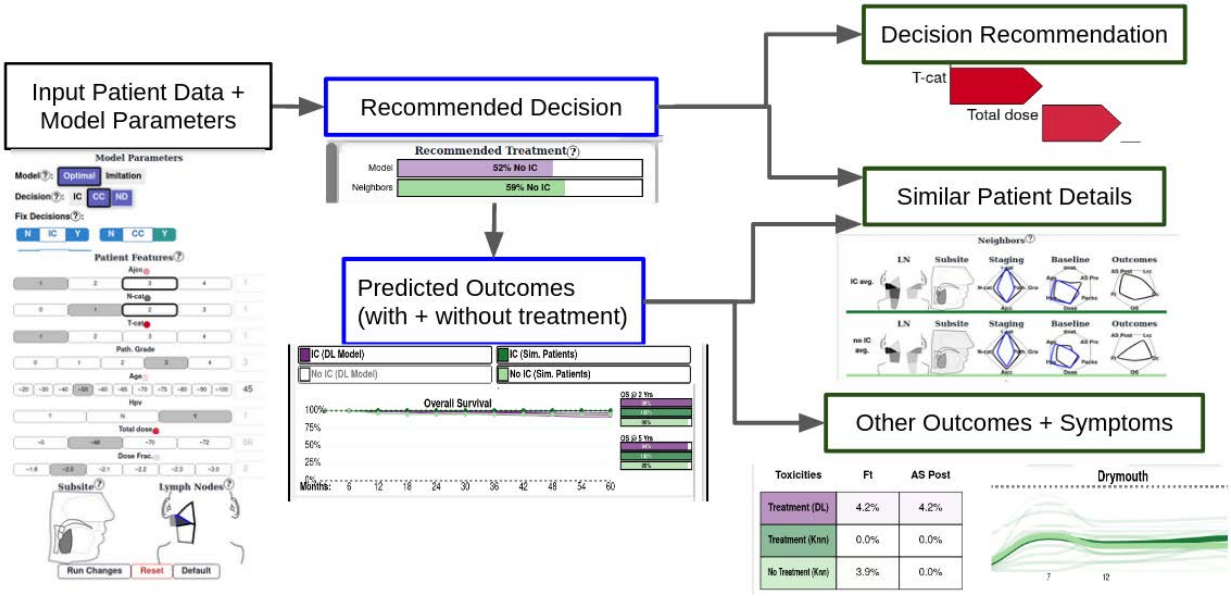

Fig. 16: Diagram of user workflow using the interface. (Left) Users start by inputting patient features and setting the model parameters, including the treatment to be considered. (Center) Users start by viewing the most prominent information: treatment recommendations and long-term patient outcome risk plots for survival and disease recurrence. (Right) Users who wish for more information can view additional views such as model explanations, similar patients, and additional patient risk prediction results.

```
###PseudoCode for treating the patient
def getGroup(isTreated,candidates, new_patient_propensity, minimum_group_size,caliper_distance_base,scale_increment):
   ##gets patients in the treated or untreated group that have a propensity similar to new_patient_propensity
   group = []
   caliper_distance_scale = scale_increment
   #gradually increase caliper size until we have minimum_group_size patients or we run out of candidates
   while len(group) < minimum_group_size and len(candidates) > 1:
      #update caliper distance
      caliper_distance = caliper_distance_base*caliper_distance_scale
      for patient in candidates:
         #check if the patient is in the treated group and has a close enough similarity score
         if patient.treatment == isTreated and absolute_value(new_patient_propensity - patient.propensity) < caliper_distance:
            #avoid repeating patients in subsequent loops
            delete patient from candidates
            #add patient to group if they are valid
            group.push(patient)
      #update caliper distance by scale_increment %
      caliper_distance_scale = caliper_distance_scale*scale_increment
   return group

def getNeighbors(cohort,new_patient_propensity, similiarity_filter_size, minimum_group_size, caliper_distance_scale, scale_increment):
   #filter out the top patients by similarity
   cohort = sort(cohort, key = lambda paient: patient.similarity_with_new_patient)
   cohort = cohort[0:similarity_filter_size]

   #calculate the standard deviation of the logit of the propensity scores of the cohort
   cohort_propensities = [patient.propensity for patient in cohort]
   caliper_distance_base = standard_deviation(logit(cohort_propensities))
   caliper_distance_scale = .1

   #calculate treated and untreated groups with scaling propensity independently
   treated_group = getGroup(True, copy(cohort), *args)
   untreated_group = getGroup(False, copy(cohort), *args)
   return treated_group, untreated_group

def localAverageTreatementEffect(outcome, *args):
   treated_group, untreated_group = getNeighbors(*args)
   treated_average = mean([patient[outcome] for patient in treated_group])
   untreated_average = mean([patient[outcome] for patient in untreated_group])
   return treated_average - untreated_average
```

Fig. 17: Pseudocode for the method of estimating average treatment effect for a patient

B.2 Prototypes

Fig. 18 and Fig. 19 show early versions of the interface. Fig. 20 Shows an early version of the outcomes view in more detail.

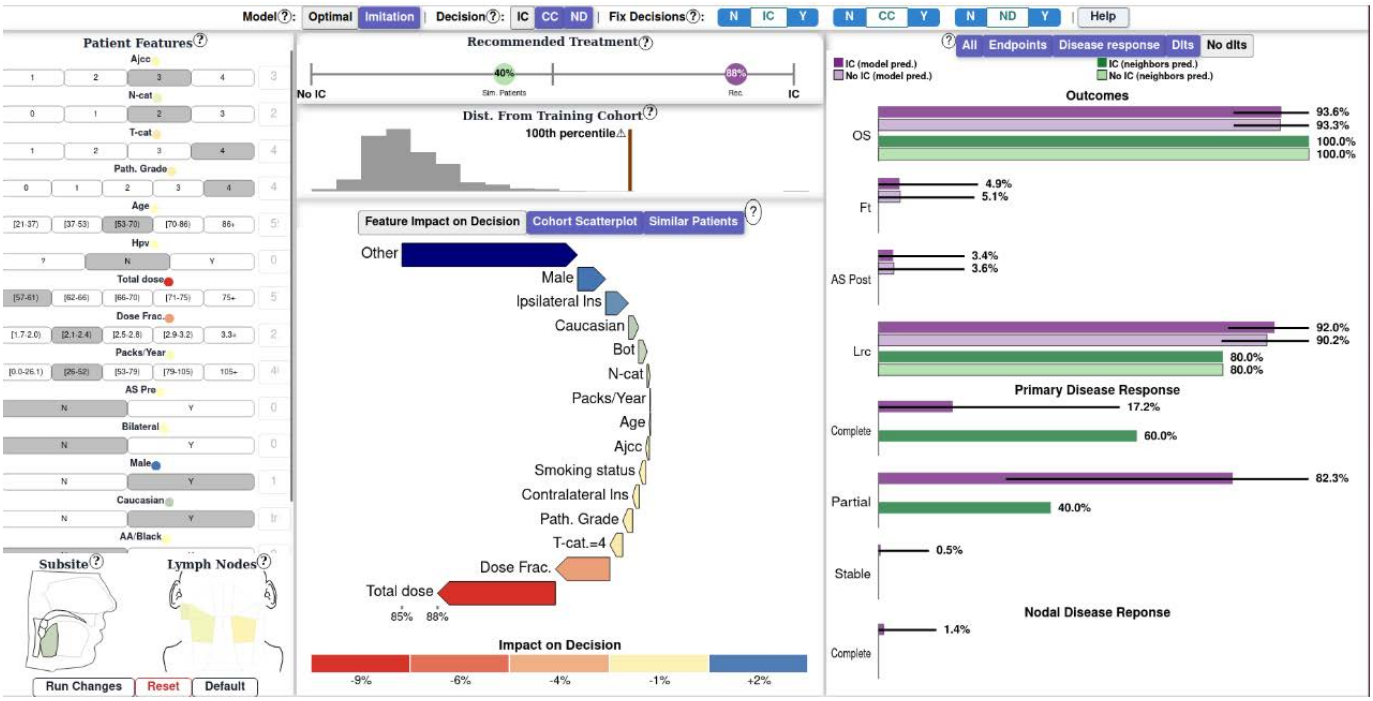

Fig. 18: Early version of the interface before integrating temporal outcomes. In this version we used a different encoding for treatment recommendation that users found intuitive as it showed the raw model output as percentage of confidence in the patient receiving treatment. This version also showed an additional histogram of the mahalanobis distances for the cohort. We also used a different colorscheme. Additionally, outcomes where shown only as barcharts with a toggle to change the set of outcomes being shown (transition states, dlts, or 4 year post-treatment outcomes). Model parameters where shown at the top instead of alongside the patient panel.

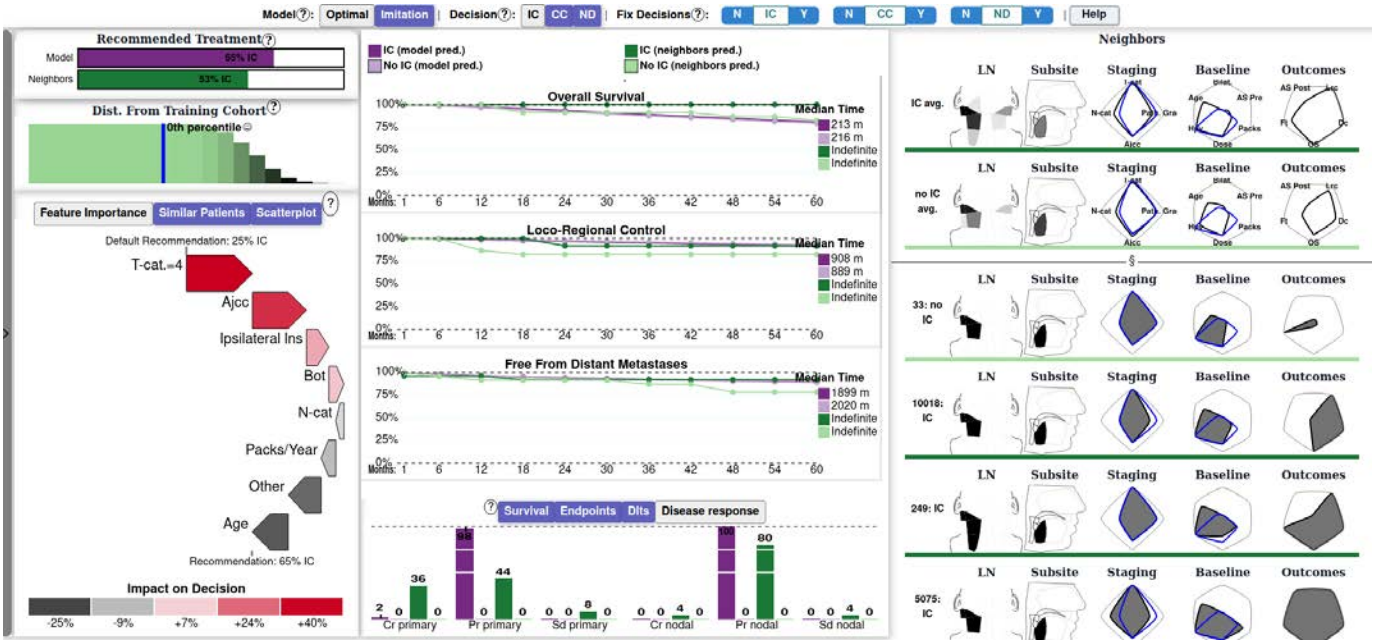

Fig. 19: Early version of the interface before the workshop. In this version the patient input panel was hidden in a "drawer" and could be pulled out via the grey section on the far left, once an initial patient was input. This version includes barcharts with alternative patient outcomes alongside temporal outcomes. Model parameters where shown at the top instead of alongside the patient panel.

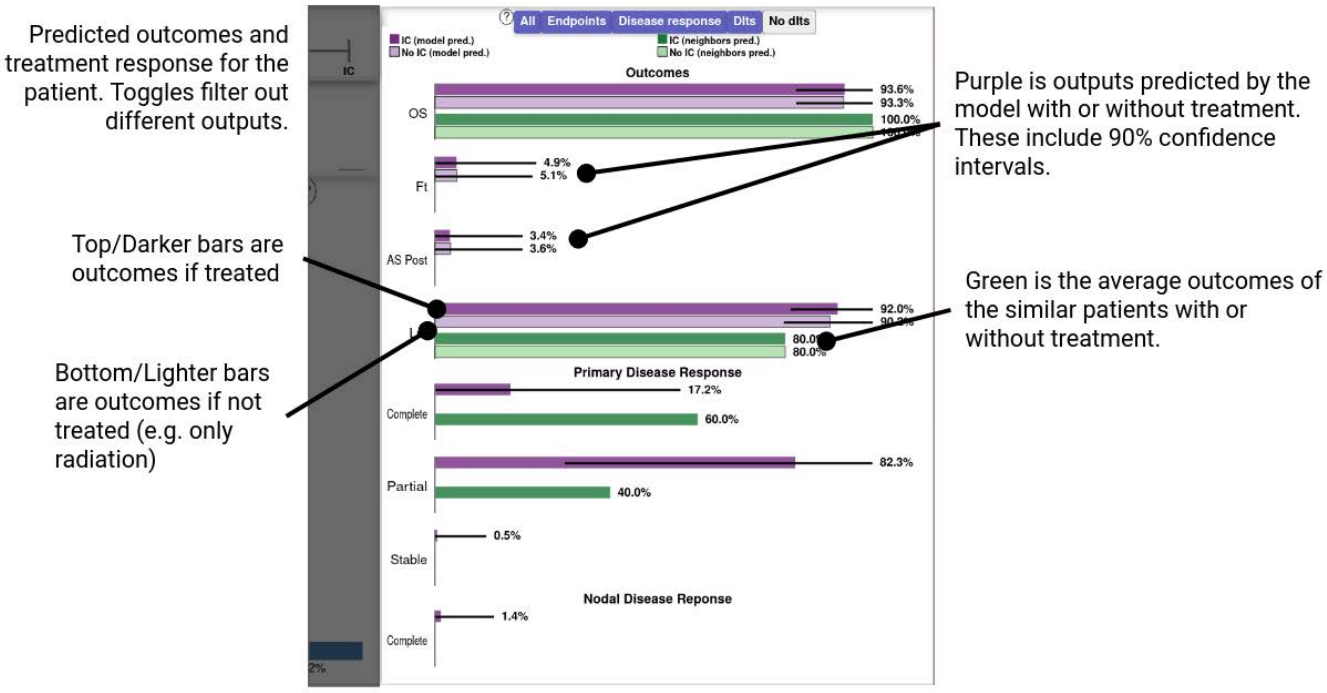

Fig. 20: Early version of the outcome view. Our original variant used only static outcomes (4 year survival etc) and focused on barcharts of multiple symptoms, based on the original DT model which used binary outcomes only. This was altered after clinicians states that they were used to dealing with temporal risk plots when reasoning about risk profiles, which also required the addition of the Deep survival machine outcome models.

References

[1] Treatment option overview for oropharyngeal cancer. https://www.cancer.gov/types/head-and-neck/hp/adult/oropharyngeal-treatment-pdq#_49. accessed July 22, 2024. 3

[2] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proc. ACM-BCB*, pp. 559–560, 2018. doi: 10.1109/ICHI.2018.00095 2

[3] M. Amin and al. *AJCC Cancer Staging Manual 8th edition*. Wiley Online Library, 12 2016. 3

[4] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. behav. res.*, 46(3):399–424, 2011. doi: 10.1080/00273171.2011.568786 6

[5] P. C. Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161, 2011. doi: 10.1002/pst.433 6

[6] G. Bouleux, H. B. E. Haouzi, V. Cheutet, G. Demesure, W. Derigent, T. Moyaux, and L. Trilling. Requirements for a Digital Twin for an Emergency Department. In *Proc. SOHOMA*, pp. 130–141. Springer, Cham, Switzerland, February 2023. doi: 10.1007/978-3-031-24291-5_11 2

[7] J. Brooke. *SUS – a quick and dirty usability scale*, pp. 189–194. 01 1996. 7

[8] E. R. Burgess, I. Jankovic, M. Austin, N. Cai, A. Kapuścińska, et al. Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. In *Proc. CHI*, CHI '23, article no. 15, 19 pages. ACM, 2023. doi: 10.1145/3544548.3581251 2

[9] G. Canahuate, A. Wentzel, A. S. R. Mohamed, L. V. van Dijk, D. M. Vock, B. Elgohari, H. Elhalawani, C. D. Fuller, and G. E. Marai. Spatially-aware clustering improves AJCC-8 risk stratification performance in oropharyngeal carcinomas. *Oral Oncol.*, 144:106460, September 2023. doi: 10.1016/j.oraloncology.2023.106460 2

[10] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016. doi: 10.48550/arXiv.1608.05745 2

[11] Y.-C. Chu, W.-T. Kuo, Y.-R. Cheng, C.-Y. Lee, et al. A survival metadata analysis responsive tool (smart) for web-based analysis of patient survival and risk. *Sci. reports*, 8(1):12880, 2018. doi: 10.1038/s41598-018-31290-z 2, 9

[12] A. Corvò, H. S. G. Caballero, and M. A. Westenberg. Survivis: Visual analytics for interactive survival analysis. In *Eurovis Workshop of Vis. Analytics*, pp. 73–77, 2019. doi: 10.2312/eurova.20191128 2

[13] A. Ebrahimi Zade, S. Shahabi Haghighi, and M. Soltani. Deep neural networks for neuro-oncology: Towards patient individualized design of chemo-radiation therapy for glioblastoma patients. *J. of Biomed. Informatics*, 127:104006, 2022. doi: 10.1016/j.jbi.2022.104006 2

[14] 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. *Trans. Vis. Comp. Graph.*, 28(01):151–161, January 2022. doi: 10.1109/TVCG.2021.3114810 2

[15] C. Floricel, A. Wentzel, A. Mohamed, 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. *Trans. Vis. Comp. Graph*, 2024. doi: 10.1109/TVCG.2023.3326939 2

[16] A. Gafita, J. Calais, T. R. Grogan, B. Hadaschik, H. Wang, et al. Nomograms to predict outcomes after 177lu-psma therapy in men with metastatic castration-resistant prostate cancer: an international, multicentre, retrospective study. *The Lancet Onc.*, 22(8):1115–1125, 2021. doi: 10.1016/S1470-2045(21)00274-6 2

[17] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *int. conf. on mach. learn.*, pp. 1050–1059. PMLR, 2016. doi: 10.48550/arXiv.1506.02142 10

[18] S. Ha, S. Monadjemi, and A. Ottley. Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics. *Computer Graphics Forum*, 43(3):e15108, June 2024. doi: 10.1111/cgf.15108 2

[19] A. Hakone, L. Harrison, A. Ottley, N. Winters, et al. Proact: Iterative design of a patient-centered visualization for effective prostate cancer health risk communication. *Trans. Vis. Comp. Graph*, 23(1):601–610, 2017. doi: 10.1109/TVCG.2016.2598588 2

[20] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proc. CHI*, 2019. doi: 10.1145/3290605.3300809 7

[21] M. Jacobs, J. He, M. F. Pradier, B. Lam, A. C. Ahn, et al. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proc. CHI*, pp. 1–14. ACM, May 2021. doi: 10.1145/3411764.3445385 2

[22] D. Kahneman. *Thinking, fast and slow*. macmillan, 2011. 9

[23] M. Karabacak, P. Jagtiani, A. Carrasquilla, I. M. Germano, and K. Margetis. Prognosis individualized: Survival predictions for who grade ii and iii gliomas with a machine learning-based web application. *NPJ Digital Medicine*, 6(1):200, 2023. doi: 10.1038/s41746-023-00948-y 2

[24] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proc. CHI*, 14 pages, p. 1–14, 2020. doi: 10.1145/3313831.3376219 2, 9

[25] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint*, 2020. doi: 10.48550/arXiv.2009.07896 6

[26] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, et al. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *Trans. Vis. Comp. Graph*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027 2

[27] B. C. Kwon, U. Kartoun, S. Khurshid, M. Yurochkin, and ohters. Rmexplorer: A visual analytics approach to explore the performance and the fairness of disease risk models on population subgroups. In *Vis. Conf.*, pp. 50–54, 2022. doi: 10.1109/VIS54862.2022.00019 2

[28] M. W. Lauer-Schmaltz, I. Kerim, J. P. Hansen, G. M. Gulyás, and H. B. Andersen. Human digital twin-based interactive dashboards for informal caregivers of stroke patients. In *Proc. PETRA*, PETRA '23, 7 pages, p. 215–221. ACM, 2023. doi: 10.1145/3594806.3594824 2

[29] K.-M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual rev. of pub. health*, 18(1):83–104, 1997. doi: 10.1146/annurev.publhealth.18.1.83 2

[30] X. Li, V. Krivtsov, and K. Arora. Attention-based deep survival model for time series data. *Reliability Engineering and System Safety*, 217:108033, 2022. doi: 10.1016/j.ress.2021.108033 2

[31] M. Louise Davies. A New Personalized Oral Cancer Survival Calculator to Estimate Risk of Death From Both Oral Cancer and Other. *JAMA Otolar.. Head. Neck Surg.*, 149(11):993–1000, November 2023. doi: 10.1001/jamaoto.2023.1975 9

[32] 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. *J. of biomed. informatics*, 112:100067, 2020. doi: 10.1016/j.yjbinx.2020.100067 2

[33] L. Ma, C. Zhang, Y. Wang, W. Ruan, et al. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proc. AAAI*, vol. 34, pp. 833–840, 2020. doi: 10.1609/aaai.v34i01.5428 2

[34] M. Mantovani, A. Wentzel, J. T. Trabucco, J. Michaelis, and G. E. Marai. Kiviat Defense: An Empirical Evaluation of Visual Encoding Effectiveness in Multivariate Data Similarity Detection. *J. Imag.Sci. Tech.*, 67:1–13, November 2023. doi: 10.2352/J.ImagingSci.Technol.2023.67.6.060406 7

[35] G. E. Marai. Visual scaffolding in integrated spatial and nonspatial analysis. In *EuroVis Works. Vis. Ana. (EuroVA)*. Eurographics Association, 2015. 5

[36] G. E. Marai. Activity-centered domain characterization for problem-driven scientific visualization. *Trans. Vis. Comp. Graph.*, 24(1):913–922, 2017. doi: 10.1109/TVCG.2017.2744459 3

[37] 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. *Trans. Vis. Comp. Graph.*, 25(4):1732–1745, 2018. doi: 10.1109/TVCG.2018.2817557 2, 7

[38] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio, et al. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE Trans. Vis. Comp. Graph. (TVCG)*, pp. 1732–1745, 4 2019. doi: 10.1109/TVCG.2018.2817557 2

[39] J. W. S. McCullough, R. A. Richardson, A. Patronis, R. Halver, R. Marshall, et al. Towards blood flow in the virtual human: efficient self-coupling of HemeLB. *Interface Focus*, 11(1):20190119, February 2021. doi: 10.1098/rsfs.2019.0119 2

[40] J. Müller-Sielaff, S. B. Beladi, S. W. Vrede, M. Meuschke, et al. Visual assistance in development and validation of bayesian networks for clinical decision support. *Trans. Vis. Comp. Graph*, 29(8):3602–3616, 2023. doi: 10.1109/TVCG.2022.3166071 2

[41] A. O. Naghavi, M. I. Echevarria, T. J. Strom, Y. A. Abuodeh, et al. Treat-

ment delays, race, and outcomes in head and neck cancer. *Cancer Epidem.*, 45:18–25, December 2016. doi: 10.1016/j.canep.2016.09.005 8

[42] C. Nagpal, X. Li, and A. Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *J. of Biomed. and Hea. Info.*, 25(8):3163–3175, 2021. doi: 10.1109/JBHI.2021.3052441 2, 5, 10

[43] C. Nagpal, W. Potosnak, and A. Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2204.07276 6

[44] E. Oral, R. Chawla, M. Wijkstra, N. Mahyar, and E. Dimara. From information to choice: A critical inquiry into visualization tools for decision making. *Trans. Vis. Comp. Graph*, 30(01):359–369, jan 2024. doi: 10.1109/TVCG.2023.3326593 2

[45] Y. Ouyang, Y. Wu, H. Wang, C. Zhang, F. Cheng, C. Jiang, L. Jin, Y. Cao, and Q. Li. Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning. *Trans. Vis. Comp. Graph*, 30(01):1238–1248, 2024. doi: 10.1109/TVCG.2023.3326929 2

[46] S. Phung, A. Kumar, and J. Kim. A deep learning technique for imputing missing healthcare data. In *Conf. IEEE EMBC*, pp. 6513–6516, 2019. doi: 10.1109/EMBC.2019.8856760 5

[47] A. Suh, G. Appleby, E. W. Anderson, L. Finelli, R. Chang, and D. Cashman. Are metrics enough? guidelines for communicating and visualizing predictive models to subject matter experts. *Trans. Vis. Comp. Graph.*, pp. 1–16, 2023. doi: 10.1109/TVCG.2023.3259341 2

[48] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Int. conf. on machine learning*, pp. 3319–3328. PMLR, 2017. doi: 10.48550/arXiv.1703.01365 5, 11

[49] A. K. Talukder, E. Selg, and R. E. Haas. Physicians' Brain Digital Twin: Holistic Clinical & Biomedical Knowledge Graphs for Patient Safety and Value-Based Care to Prevent the Post-pandemic Healthcare Ecosystem Crisis. In *Knowledge Graphs and Semantic Web*, pp. 32–46. Springer, Cham, Switzerland, November 2022. doi: 10.1007/978-3-031-21422-6_3 2

[50] E. Tardini, X. Zhang, G. Canahuate, A. Wentzel, A. S. Mohamed, L. Van Dijk, C. D. Fuller, and G. E. Marai. Optimal treatment selection in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: Deep q-learning with a patient-physician digital twin dyad. *J. med. Int. res.*, 24(4):e29455, 2022. doi: 10.2196/29455 3, 5

[51] X. Teng, Y. Ahn, and Y. Lin. Vispur: Visual aids for identifying and interpreting spurious associations in data-driven decisions. *Trans. Vis. Comp. Graph.*, 30(01):219–229, 2024. doi: 10.1109/TVCG.2023.3326587 2

[52] T. M. Therneau. Extending the cox model. In *Proce. first Seattle symp.in biostat.: surv. anal.*, pp. 51–84. Springer, 1997. doi: 10.1007/978-1-4684-6316-3_5 2

[53] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, and I. El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med. Phys.*, 2017. doi: 10.1002/mp.12625 2

[54] L. V. van Dijk, Sr. Abdallah Mohamed, S. Ahmed, N. Nipu, et al. Head and neck cancer predictive risk estimator to determine control and therapeutic outcomes of radiotherapy (hnc-predictor). *European J. of Cancer*, 178:150–161, 2023. doi: 10.1016/j.ejca.2022.10.011 2, 9

[55] H. van Hasselt, A. Guez, and D. Silver. Deep Reinforcement Learning with Double Q-Learning. *AAAI*, 30(1), March 2016. doi: 10.1609/aaai.v30i1.10295 4

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Adv. in neur. info. proc. sys.*, 30, 2017. doi: 10.48550/arXiv.1706.03762 11

[57] R. D. Vromans, S. Hommes, F. J. Clouth, D. N. Lo-Fo-Wong, et al. Need for numbers: assessing cancer survivors' needs for personalized and generic statistical information. *BMC Medi. Informatics and Dec. Making*, 22(1):1–14, 2022. doi: 10.1186/s12911-022-02005-2 2

[58] J. Wang, L. Gou, H.-W. Shen, and H. Yang. Dqnviz: A visual analytics approach to understand deep q-networks. *Trans. Vis. Comp. Graph*, 25(1):288–298, 2019. doi: 10.1109/TVCG.2018.2864504 2

[59] J. Wang, W. Zhang, H. Yang, C.-C. M. Yeh, and L. Wang. Visual analytics for rnn-based deep reinforcement learning. *Trans. Vis. Comp. Graph*, 28(12):4141–4155, 2022. doi: 10.1109/TVCG.2021.3076749 2

[60] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), article no. 110, 36 pages, feb 2019. doi: 10.1145/3214306 2

[61] Q. Wang, K. Huang, P. Chandak, M. Zitnik, and N. Gehlenborg. Extending the nested model for user-centric xai: A design study on gnn-based drug repurposing. *Trans. Vis. Comp. Graph.*, 29(1):1266–1276, 2022. doi: 10.1109/TVCG.2022.3209435 2

[62] Y. Wang, L. V. Dijk, A. S. R. Mohamed, M. Naser, et al. Improving prediction of late symptoms using lstm and patient-reported outcomes for head and neck cancer patients. In *Proc. ICHI*, pp. 292–300, 2023. doi: 10.1109/ICHI57859.2023.00047 4

[63] J. Waser, R. Fuchs, H. Ribičič, B. Schindler, G. Blöschl, and E. Gröller. World lines. *Trans. Vis. Comp. Graph.*, 16(6):1458–1467, 2010. doi: 10.1109/TVCG.2010.223 2

[64] 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. In *Vis. Conf. (short paper)*, pp. 281–285. IEEE, 2020. doi: 10.1109/VIS47514.2020.00063 6, 7

[65] A. Wentzel, C. Floricel, G. Canahuate, M. Naser, A. S. Mohamed, C. D. Fuller, L. van Dijk, and G. E. Marai. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Comp. Graph. For.*, 24(3), Jun 2023. doi: 10.1111/cgf.14830 4, 7

[66] 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. *Trans. Vis. Comp. Graph.*, 26(1):949–959, 2019. doi: 10.1109/TVCG.2019.2934546 2

[67] 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 Onc.*, 148:245–251, 2020. doi: 10.1016/j.radonc.2020.05.023 2

[68] A. Wentzel, T. Luciani, L. V. van Dijk, N. Taku, B. Elgohari, A. S. Mohamed, G. Canahuate, C. D. Fuller, D. M. Vock, and G. Elisabeta Marai. Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas. *Radiotherapy & Onco.*, 161:152–158, 2021. doi: 10.1016/j.radonc.2021.06.016 2

[69] A. Wentzel, A. S. R. Mohamed, M. A. Naser, L. V. van Dijk, K. Hutcheson, A. M. Moreno, C. D. Fuller, G. Canahuate, and G. E. Marai. Multi-organ spatial stratification of 3-d dose distributions improves risk prediction of long-term self-reported severe symptoms in oropharyngeal cancer patients receiving radiotherapy: development of a pre-treatment decision support tool. *Front. in onc.*, 13, 2023. doi: 10.3389/fonc.2023.1210087 2

[70] C. Xiong, E. Lee-Robbins, I. Zhang, A. Gaba, and S. Franconeri. Reasoning affordances with tables and bar charts. *Trans. Vis. Comp. Graph*, pp. 1–13, 2022. doi: 10.1109/TVCG.2022.3232959 7

[71] L. Xu and C. Guo. Coxnam: An interpretable deep survival analysis model. *Expert Systems with Applications*, 227:120218, 2023. doi: 10.1016/j.eswa.2023.120218 2

[72] Q. Yang, A. Steinfeld, and J. Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proc. CHI*, p. 1–11. ACM, 2019. doi: 10.1145/3290605.3300468 2

[73] L. Zdilar, D. M. Vock, G. E. Marai, C. D. Fuller, A. S. Mohamed, H. Elhalawani, B. A. Elgohari, C. Tiras, A. Miller, and G. Canahuate. Evaluating the effect of right-censored end point transformation for radiomic feature selection of data from patients with oropharyngeal cancer. *JCO clin. cancer informativd*, 2:1–19, 2018. doi: 10.1200/CCI.18.00052 2

[74] V. Zebralla, J. Müller, T. Wald, A. Boehm, et al. Obtaining patient-reported outcomes electronically with "oncofunction" in head and neck cancer patients during aftercare. *Front. in onc.*, 10:549915, 2020. doi: 10.3389/fonc.2020.549915 2

[75] M. Zhang, D. Ehrmann, M. Mazwi, D. Eytan, M. Ghassemi, and F. Chevalier. Get to the point! problem-based curated data views to augment care for critically ill patients. In *Proc. CHI*, CHI '22, article no. 278, 13 pages. ACM, 2022. doi: 10.1145/3491102.3501887 2

[76] B. Zheng, S. Verma, J. Zhou, I. W. Tsang, and F. Chen. Imitation Learning: Progress, Taxonomies and Challenges. *Trans. Neu. Net. Learn. Sys.*, pp. 1–16, October 2022. doi: 10.1109/TNNLS.2022.3213246 4

[77] Z. Zhu, C. Liu, and X. Xu. Visualisation of the Digital Twin data in manufacturing by using Augmented Reality. *Proc. CIRP*, 81:898–903, January 2019. doi: 10.1016/j.procir.2019.03.223 2

[78] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *Trans Vis. Comp. Graph*, 28(1):1161–1171, 11 pages, jan 2022. doi: 10.1109/TVCG.2021.3114864 2