Capstone Technical Report

Andrew Balch
xxv2zh@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

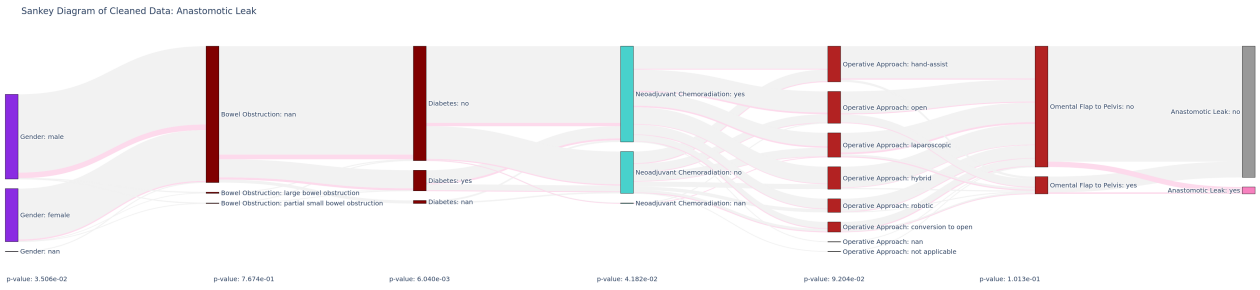

Figure 1: Sankey-based visualization of Anastomotic Leak progression using ML-selected features.

ABSTRACT

Effectively treating rectal cancer requires attentive consideration of hundreds of details about the patient and how each may influce patient outcomes and negative side-effects. The US Rectal Cancer Consortium has complied a one-of-a-kind dataset that contains fine-grain details about individual patient treatment paths from 6 different institutions over a decade. Only statistical analysis for the impact of post-op complications (POCs) on oncologic outcomes has been conducted on this dataset, so the development of visualization and machine learning tools could further progress in rectal cancer treatment. We present a simple data processing approach for the RCC as well as a data-driven event sequence visualization of the incidence of a major POC (anastomotic leakage). This visualization utilizes machine learning feature selection approaches to uncover variables that are powerful predictors of an anastomotic leak, including a novel chained approach for event sequences. A rectal cancer surgeon stakeholder was consulted throughout development and provided feedback on the final tool. The tool is generalizable to other event sequence datasets and has been made publically available at github.com/HAI-lab-UVA/RCC-Project.

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Andrew Balch. 2018. Capstone Technical Report. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Colorectal cancer is one of the most common forms of cancer, and was responsible for 935,173 deaths worldwide in 2020 [1]. As its name suggests, colorectal cancer includes both colon and rectal cancer. Rectal cancer specifically involves the rectum, the area of the digestive tract located within the pelvis. As with most cancers, there are a variety of factors that impact a patient's course of treatment and chances of recovery. For these reasons, it is incredibly important that clinicians analyze as much about the patient and their potential treatment paths as possible, in order to maximize treatment effectiveness and minimize negative complications. This has motivated the creation of the US Rectal Cancer Consortium (RCC), a dataset with over a decade of rectal cancer patient data across 6 institutions [2, 4, 5, 7, 13, 19, 28, 29]. A series of papers has been published that conduct statistical analyses on subsets of features in this dataset, but there is an unrealized opportunity to apply machine learning (ML) to yield more complex and comprehensive insights.

The field of precision medicine is broad, but generally involves using fine-grained data about a patient to tailor a treatment plan to their individual characteristics. This can take multiple forms, such as building models to predict treatment paths and/or outcomes, clustering patients based on shared attributes, mining the data for common patterns, or visualizing a patient's information in a complex and sophisticated dashboard. Such an approach that utilizes a large dataset of graular patient treatment data to provide clinicians

with actionable insights has yet to be taken in the field of rectal cancer treatment.

In this paper, we present the process by which we developed a ML tool for analyzing not just the rectal cancer treatment outcomes, but any event sequence dataset. Our contributions are:

(1) A description of data processing methods for a highly-detailed dataset of rectal cancer patient data.
(2) A data-driven, Sankey diagram-based visualization for event series in a popular Python library (Plotly).
(3) An experiment in using a common feature selection method (Sequential Feature Selection) to reduce visualization complexity.
(4) A proposal of a chained feature selection technique for better representation of events in a sequence.
(5) An open-source implementation of the presented visualization and feature selection approaches, generalizable to other event-series data.

Our work was centered around analyzing anastomotic leakage as a result of a lower anterior resection (LAR), but can easily be generalized to other patient outcomes, or event series outcomes more broadly.

2 BACKGROUND

2.1 Rectal Cancer Treatment and Outcome Analysis

The primary treatment for rectal cancer is surgical resection of the tumor, but the surgical approach, intent, and whether the surgery is preceded and/or followed by chemotherapy (neoadjuvant and adjuvant treatment, respectively) depends on the stage of cancer and the individual patient themselves [1]. Treatments and therapies are not without their risks and drawbacks, they can be a large burden on the patient and cause lasting issues related to normal bodily functioning that could even leave to death. Clinicians try to take as many factors about the patient as possible into account, but this is impossible for a human and naturally some variables come to dominate the clinical decision-making process. Therefore, being able to analyze how oft-considered risk factors and different treatment paths impact clinical outcomes is of critical importance to improve the state of rectal cancer care. The RCC exists for this reason, and a number of papers have been written conducting these sorts of analyses [2, 4, 5, 7, 13, 19, 28, 29]. While these papers are successful in exploring how different subsets of features impact one or more outcomes of interest, they are limited by the use of statistical multivariate analyses. There is a need to assess all 400+ features in this dataset and discover where important features may have been overlooked, how treatment decisions impact subsequent treatments as well as outcomes over time, and convey these insights to clinicians in an intuitive way.

Anastomotic leakage (AL) is a high-stakes complication that can occur during rectal cancer surgery. AL has an incidence rate of up to 21%, and "dramatically increases postoperative mortality, increases the risk for a permanent stoma, and leads to worse oncologic outcomes following resection for colorectal cancer" [2, 8]. A leak can be tested for during surgery and its negative impacts avoided by performing a diverting loop ileostomy (diversion). However, the incidence rate of leakage, the severity of its impact, and inaccurate testing methods means that AL is over-treated in ~50% of patients and only ~30% of patients who later developed AL were treated with a diversion [24]. A paper using the RCC database applied multivariable logistic regression to assess the association between an omental pedicled flap (OPF) and the occurrence of anastomotic leak [2]. This single independent variable was selected because OPF is hypothesized to prevent leakage. The study ultimately found no association among patients who underwent a lower anterior resection (LAR) procedure for rectal cancer. Outside the RCC, three papers have explored the usefulness of traditional machine learning techniques on predicting a leak. Wen et al. [31] built a random forest classifier based on the clinical data of 5,220 patients (20 features), which outperformed the widely-used nomogram (AUC of 0.87 v. 0.724). Shao et al. [24] found that a support vector machine (SVM) classifier trained to predict AL would significantly reduce overtreatment. Shen at al. [26] compared stepwise (AUC: 0.759) and LASSO (AUC: 0.79) classifiers on a cohort of 860 patients (9 features). The features most associated with leakage, as selected by the models, differed between the two studies. A machine learning approach has yet to be taken that considers a larger set of rectal cancer features in predicting anastomotic leak, and presents that information to clinicians in an actionable manner.

2.2 Clinical Decision Support Systems

Clinical decision support systems (CDSS) are increasingly using AI and ML to determine the optimal treatments for clinical conditions. Most of the recent work focuses on sepsis prognosis, applying reinforcement learning (RL) techniques to learn the best treatment dose strategies from time-series ICU data [9, 12, 14, 17]. Conversely, inverse reinforcement learning (IRL) has been used to learn treatment strategies from the techniques of clinicians themselves [17, 34]. In day-to-day cancer treatment, however, a clinician's interest is in what interventions (surgical or therapeutic) are the most appropriate for an individual patient, rather than the dosages of individual drugs or the granular intervals of their administration. Thus, tools that predict outcomes and model the impact of different interventions on these outcomes would be most beneficial. In surgical and cancer-specific CDSS, image analysis of MRIs/CTs and "-omics" data mining receives a lot of attention, while the utility of electronic health records (EHRs) to derive treatment insights remains undervalued [3, 18, 25, 33]. Some such approaches have emerged surrounding rectal cancer: a random forest for radiation therapy treatment planning in prostate cancer [20] as well as a Bayesian network for prognosis prediction in post-radical resection surgical patients [16]. While these are great steps forward, CDSS are still limited by a clinician's hesitance to rely on them [3, 27, 30]. Yang et al. [32] addresses this issue by subtly incorporating ML-based predictions into something clinicians use every day: a patient case slide. For s CDSS to be useful, we must consider a clinician's ability to trust it. Part of this battle is creating more capable and transparent systems, and part of it is presenting information and predictions in a manner that is intuitive.

2.3 Event Sequence Visualization

Visualization techniques can help address the latter half of the CDSS challenge. Event sequences are defined as a series of "discrete events in the time order of occurrence" [6] and are a popular way to model the treatment of a disease over time (e.g. pre-treatment staging to neoadjuvant treatment to surgical intervention for cancer). While time-series data can be cast to an event sequence, time-series data is distinct in that it is continuous. Timelines, hierarchies (e.g. trees), and Sankey diagrams are commonly used to provide an overview of event sequences. In clinical use-cases, these charts are useful to compare different patient cohorts, visualize disease outcomes, and analyze treatment prognoses. Modern implementations make use of complex dashboards to facilitate data selection, event pattern/rule mining, and/or clustering of patients into cohorts. These steps are reflected in the visualization itself, which may even use deep learning (DL) to predict future events in the sequence. Putting all of this together creates a sophisticated tool for exploring dense, time-dependent EHR data, gathering insights, and communicating the most important information to the clinician. The sophisitcation of these tools becomes their limiation, with studies noting a steep learning curve for clinical stakeholders, who prefer a more straightforward, at-a-glance style of communicating patient information.

3 METHODS

3.1 Dataset Summary

The US rectal cancer consortium (RCC) is a dataset that combines demographic, intraoperative, histopathologic, and postoperative outcome data for rectal cancer treatment across six member institutions. Overall, the RCC contains 408 unique data features (63 numeric, 345 categorical) on 1881 patients who underwent a surgical intervention for rectal cancer. These 408 features are broken up 13 into different observational and treatment stages (summarized here). Therefore, we model the RCC as an event sequence. The data was compiled by hand from EHRs, so there is a high variance in its quality and completeness, depending on the institution. While the RCC itself cannot be made public, a cursory documentation of its features, pre and post-processing, can be found here.

3.2 Design Approach

We worked closely with a rectal cancer surgeon stakeholder throughout the course of the project and used an iterative design approach to implement their feedback. We planned each feature of the project, discussed the details of the feature with the stakeholder, implemented it, and then moved on to the next feature. At the end of the project, we conducted a full walkthrough with the stakeholder to understand the strengths and weaknesses of the project from a clinical perspective. Additionally, we kept the implementation of the features as data-agnostic as possible in the hopes of creating a tool that could be reused for other event sequence datasets. The source code and documentation for this paper has been published at GitHub for public use: github.com/HAI-lab-UVA/RCC-Project.

3.3 Data Processing

Since the goal of our project was to develop a clinically-actionable tool for modeling and visualizing event series data, fully cleaning and feature engineering the data was not our priority. Given the size and complexity of the RCC, it would have taken up most of our time and would have had little impact on the design as a whole. Instead, we focused on reaching a minimum viable product that would allow us to experiment with the visualization tool in meaningful ways. The Jupyter Notebook where we conducted these operations can be found here.

First, we found columns that had less than 10 non-null samples or that contributed unimportant information (9 total) and dropped them.

Then, we conducted binning in a few different ways. We tried to extract as much information as possible from free-text features that had a high magnitude of unique values (e.g. surgical or therapeutic complications, organs invaded by the cancer). To do this, we looked at the unique tokens for each column and sorted them into bins (e.g. types of complications or organs). Each bin was represented by a new column in the dataset, where its values were the severity of that bin. Severity was represented as an integer, and increased by 1 each time a token within the same bin was found in a data point. 4 columns were binned in this manner. The process was a little different for chemotherapy regimens, which could have been changed over the course of treatment. We binned common tokens in the same way, but also mined for tokens that indicated a change or reduction in treatment. Our program counted each of these tokens as they appeared in a data point, and made separate columns to indicate the number of changes and whether dosage had to be reduced. Instead of numerical severity, one-hot encoding was used to represent the first and second (if applicable) chemotherapy regimens. This was selected because multiple types could be used together in a 'cocktail' and we do not expect the same type to occur multiple times. 3 columns were binned in this manner. A much simpler binning technique was used to one-hot encode columns that had lots of similar values or co-occurrences, but did not need to account for severity or treatment changes/reductions (e.g. prior abdominal operations). 5 columns were binned in this manner. Two numerical columns could be binned based on a threshold of values.

We continued with some more miscellaneous data cleaning. Text-based data points that were simply 'data unavailable' were made null. Similarly, values of 'not assessed' were made null for select columns (10 total). Columns that contained only 'yes' or 'no' were encoded as binary (8 total). The dataset was combed for 'dirty' data points that were invalid or otherwise inconsistent with the rest of the values in a column, their corresponding samples were dropped (48 total).

Last, we cast any date columns to the proper data type and ensured all other columns fit their respective data types.

3.4 Event Sequence Visualization

With the data properly formatted and potentially-useful features extracted, we shifted gears to develop the visualization technique. We chose a Sankey diagram (sometimes called an alluvial diagram) because it retains the temporal aspect of our data, provides more information at a glance than a timeline, and is much more practical than a hierarchy/tree given the distribution of unique values in our data. A Sankey diagram is akin to a flowchart or a directed, acyclic graph. It models a series of nodes connected to each other

by edges. When used to represent data, each node can represent a value and an edge can represent two values occurring together. The relative width of each node and edge can be used to illustrate the magnitude of the flow through them, usually determined by the frequency of values and their co-occurrence within the data. What makes a Sankey distinct from a tree is that any previous edges are ignored when the next edge in the sequence is defined. This means that edges are not as easy to trace down to a leaf node.

Our formulation of a Sankey diagram considers the different levels (unique values) of a variable together. Each level constitutes a different node, with no edges between levels of the same variable. The nodes for each level are arranged horizontally in a stack, and their relative height is determined by the level's prevalence in the overall dataset. Edges are formed between the nodes of the current (incoming) variable and the nodes of the next (outgoing) variable, where their values exist in the same sample (row) of the data. Like node height, edge width is determined by the total number of samples that have both the incoming and outgoing values. Since many levels are likely to crowd the horizontal space of the diagram, this formulation works best modeling discrete variables without many levels.

We implemented a Python class that defines such a Sankey diagram using the Plot.ly [10] visualization package given a Pandas [21] DataFrame, a sequence of (categorical) features, and a single outcome variable. Plot.ly was chosen because it is one of only a few packages that has a base class for a Sankey diagram. The underlying code iterates over each feature, defining its nodes, their sizes, and locations, then finds any edges between them and the nodes of the next feature, ending at the pre-defined outcome.

We made two important design choices to address the limited ability to 'trace' a Sankey diagram through to an outcome. First, we color-coded the edges to correspond with the levels of the outcome variable. Under the hood, this works by defining a separate edge between nodes for each level of the outcome. Therefore, the number of samples that have the incoming value, the outgoing value, and any particular level of the outcome can be visually indicated. Second, we implemented optional chi-squared statistical testing for associations between levels of a given feature and levels of the outcome. When this toggled, the result of the test for each feature is displayed as a p-value below its nodes.

Furthermore, to maintain the event sequence nature of the RCC data we color-coded the nodes of each feature according to the 'event' they are a part of (e.g. all nodes of features that are demographic data are blue). Our diagram also includes some basic interactivity. Most notably, when a user mouses over a node or edge the frequency of the represented values is quantified as a percent of the total number of samples in the input DataFrame.

3.5 Feature Selection

Our base data-driven Sankey diagram assumes that the user has a set of features they would like to evaluate in the context of the outcome. However, it could be the case that the user does not know which features they are interested. In such a scenario, it would be unrealistic to visualize and explore all 408 features of the RCC. So, we opted to use multiple ML classifiers together with sequential feature selection (SFS) to show this type of user the most important features for their outcome as well as how powerful a potential classifier may be.

Given a base estimator, SFS starts with no features and greedily adds them (forward SFS) or starts with all features and greedily removes them (backward SFS). Both scenarios iterate until a threshold for improvement or total feature count is met [23]. This approach is useful in event sequence cases such as ours becuase it maintains the general temporal relationship between features. We implement two techniques of using SFS with our visualization, both using scikit-learn's [22] SFS implementation and the scikit-learn BaseEstimator class for compatability with a variety of ML models.

3.5.1 Sequential Feature Selection. This first aproach conducts (forward or backward) SFS on the whole input dataset with several ML classifiers. The selected features for each classifier are aggregated, and the final features are determined by whether agreement between classifiers meets a threshold hyperparameter (i.e. 3 of 5 classifiers must agree that 'Bowel Obstruction' improves accuracy). Once the features are selected, they are plotted with our Sankey diagram. The function itself takes a DataFrame, training data, output labels, a dictionary of scikit-learn classifiers, and an agreement threshold. Optional features include the ability to load a previous SFS session from a JSON log file, provide a tree-based model as a baseline, adjust the number of features selected by each SFS run, and evaluate each model against a test set.

3.5.2 Feature Selection Chains. The second approach is an extension of the first. For event sequence data where events have multiple features, SFS may do a poor job at selecting features across different events because of its greedy nature. In testing, we found that SFS was selecting only demographic features and nothing from later events. To address this, we formulate a feature selection chain where SFS is conducted on each event. This ensures that the important features from each event in a sequence are shown to the user and included in the model. While this function is implemented around SFS, the source code could easily be extended for any feature selection technique.

4 RESULTS: CASE STUDY IN ANASTOMOTIC LEAK PREDICTION

To demonstrate the use of the tools we developed, we conducted a case study in AL prediction using the RCC dataset. We presented our progress to the stakeholder throughout the semester, who provided us with feedback. For more information about this case study and how results were obtained, please see this paper's GitHub page.

4.1 Data Processing

The RCC dataset was processed as described in Section 3.3. The inclusion/exclusion criteria from Ali et al. [2] were replicated, which left us with $n = 825$ samples where ~54.06% did not experience AL, ~4.97% did, and ~40.97% were missing data.

4.1.1 Stakeholder Feedback. Our stakeholder recommended that the severity-based binning system for treatment complications be changed to separate major complications from minor ones. Their concern was that a patient's symptoms could appear misleadingly severe if they were to have several "basic" side effects. They provided a list of critical complications that should be noted as major.

For the chemotherapy regimen features, the stakeholder pointed out an error in our binning dictionary. They clarified that many regimens are simply combinations of other regimens. For example, the regimen Folfox is 5FU with Oxaliplatin and Folfiri is 5FU with Irinotecan. Therefore, Folfox, Folfiri, and 5FU are not distinct categories because they all include 5FU. The stakeholder also identified several categorical features that could be encoded into hierarchical, numeric values, such as wound class or pathologic response.

4.2 Event Sequence Visualization

Our initial visualization used the following features, selected based on existing studies predicting or analyzing AL [2, 24, 26, 31]: gender, bowel obstruction, diabetes, neoadjuvant chemoradiation, operative approach, and omental flap to pelvis (OPF). The resulting figure (fig. 1) summarizes these characteristics of the sample and how they are associated with AL. This can lead to some interesting observations. For example, the open, laparoscopic, and conversion to open surgical approaches were the only ones that carried out an OPF and also had an anastomotic leak. Note that OPF is not intended to prevent a leak, just the negative outcomes it causes. We can also visually see that male patients had a much higher rate of AL than female patients.

4.2.1 Stakeholder Feedback. The stakeholder was excited about the visualization, although it did take a few minutes to form a full understanding of the Sankey diagram and what it represented. They had seen and used decision trees before in their work, and thought it was a unique approach. The diagram prompted curiosity from the stakeholder, who began contemplating the reasons for many of the connections, proposed new variables to test in the context of AL, and speculated about how it may be helpful for patients trying to decide on a treatment path for themselves. They thought the addition of quantitative metrics, value proportions and statistical testing, was practical and provided them with additional information. However, they did wish the different levels of the response were more visually distinct. When asked if the visualization would be useful for presenting in a meeting, such as to discuss a patient's treatment plan, the stakeholder responded "yes".

4.3 Feature Selection

4.3.1 Sequential Feature Selection. The following classifiers made up our SFS ensemble: Support Vector Machine, Ridge, Logistic Regression with Stochastic Gradient Decent, Multi-Layer Perceptron, and Adaboost. These were selected because they are some of the most common ML classification models. We ran forward SFS to select 10 features, and set the between-model agreement threshold to 3. To prepare the data for model training, we dropped features that were numeric, date-time, unique identifiers, or that implied a leak. We also decided to impute all missing values for AL as 0 ('no AL'). Features with >4 unique values were one-hot encoded, while all others were target encoded. The transformed dataset was split into a train and test set (80/20 split), stratified by AL. SFS selected the features gender, functional status, and family history of colorectal cancer (fig. 2). Of these, only gender was significant (p-value: 0.035) via the chi-squared test. Using these three features, the strongest classifier was a tie between the SVM, Ridge, and Adaboost. All of these models showed a F1-score of 0.77 for predicting AL and 0.99 for predicting no AL. Classifier results are provided only as a proof-of-concept and are not the main purpose of this project.

4.3.2 Sequential Feature Selection Chain. Our data transformation and model ensemble remained the same as in vanilla SFS. The agreement threshold was adjusted to be a proportion relative to the number of features in an event (5%). The feature chain selected 13 features across the 4 pre-defined events (fig. 3). All the features from SFS were selected again. The models trained to select the features from each event were much poorer predictors of AL, able to predict true negatives very well but often not predicting positive instances at all.

4.3.3 Stakeholder Feedback. The stakeholder was impressed with the general performance of the classifiers, and very interested in the features they selected. Like the connections in the Sankey diagram, it sparked a conversation about why these features in particular could be seen as more important than others. The stakeholder appreciated how the feature chains looked at a much broader snapshot of treatment.

5 DISCUSSION

Overall, our tool was successful in that it provided a novel and useful visualization of RCC data. This is most evident in how our stakeholder met the tool with curiosity, they wanted to explore the data and experiment with different variables and outcomes. While predictive power was not our priority and there are enough biases in our process to not be fully confident in the classification results, our stakeholder was nonetheless impressed with these results.

5.1 Limitations

With this said, our approach was not without its limitations. Our stakeholder already pointed out the most glaring issues with our data processing (Section 4.1.1). Besides these, the RCC still suffers from a lot of missing data that will need to be properly handled before any reliable modeling can be carried out.

Most visualization limitations are inherited from the Sankey diagram itself. First, it is only practical to visualize discrete variables with few levels. Any numerical features would need to be discretized, through which information is lost, or displayed using a parallel coordinates plot, which increases the complexity of both the visual and the underlying code. Next, comparing and contrasting patient cohorts within the same diagram is difficult. Examining multiple subpopulations at the same time would require stacking multiple Sankeys. Lastly, the overall length of the diagram can be detrimental. As more and more features across separate events are added, the plot gets longer and it gets harder to see how features towards the beginning flow into features towards the end.

In feature selection, we have already pointed out one limitation of SFS. Its greedy nature means that it selects only the first best features (or removes the last worst in the backwards case) and is therefore biased towards the beginning of an event sequence (fig. 2). So, it is likely that the whole sequence of events is poorly represented by SFS. Feature chains were proposed to ameliorate this issue. While effective in increasing representation of multiple events, the fact that it still relies on SFS means that the greedy bias

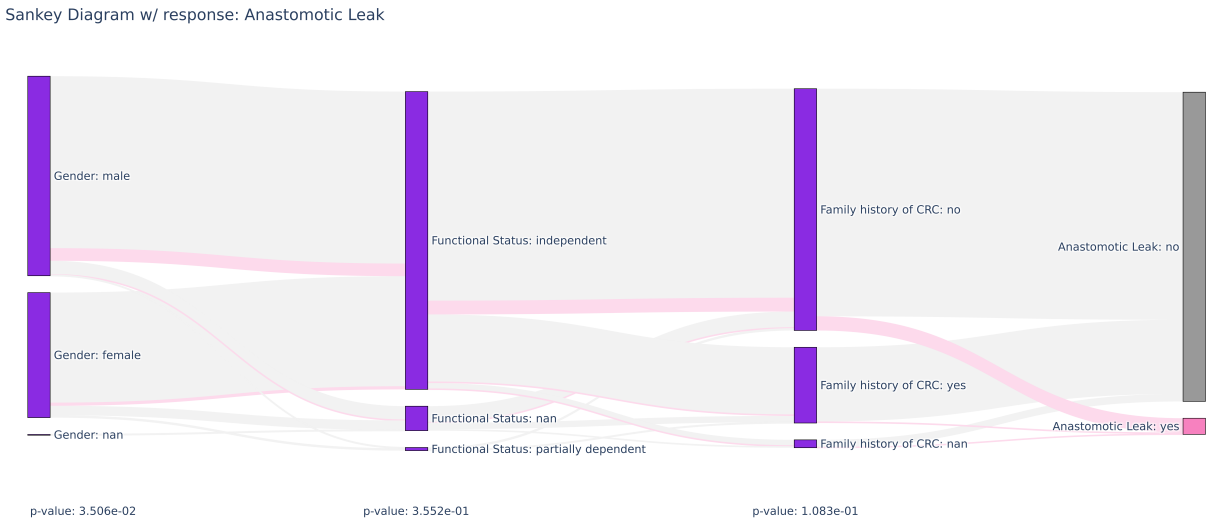

Figure 2: Visualization of features selected by Sequential Feature Selection, note the bias towards the beginning of the event sequence (demographic variables).

persists within events. Feature chains also exacerbate the length issue of Sankey diagrams. As more features across events are selected, the diagram becomes longer and harder to follow.

5.2 Future Work

The biggest hurdle for future work that hopes to leverage the RCC is refining the data processing approach. The stakeholder's suggestions (Section 4.1.1) should be implemented and missing values should be carefully considered. Some of these missing values may hold information, and some may be safely imputed using an appropriate strategy.

Improvements can also be made to the Sankey diagram itself. We have already begun work creating a branching Sankey. Like a decision tree, it splits across the levels of a feature that the user wants to explore more (e.g. whether a patient completed chemotherapy), allowing for easier comparison of patient cohorts. Support for visualizing continuous variables would be greatly beneficial. Statistical analysis could also be refined. Chi-squared tests are not universally applicable and cannot accommodate continuous variables. We may also be able to give an indication of one feature's association with the next feature in the diagram by adding a statistical test. Once the data processing is improved, we can be more confident displaying the results of different classifiers to the user. Interactivity is another consideration. This could be enhanced by allowing the user to filter the sample by clicking on a node or an edge, even going so far as re-training any predictive models based on the new sample. Finally, we could take inspiration from other event sequence modeling papers that use deep learning to forecast prognosis and apply a recurrent neural network (RNN) or RL algorithm to do the same thing for RCC [11, 15].

Feature selection may not be the best technique to reduce the visual complexity of the Sankey diagram while displaying the most important features to the user. Association rule mining (ARM), Sequential Pattern Mining (SPM), and clustering are more popular analyses in the field of event sequence visualization [6]. These approaches are used to uncover common sequences of events within the sequence as well as distinct cohorts within a sample. They could be useful alternatives to SFS and feature selection chains, providing more pertinent information to the clinician especially when combined with a branching Sankey diagram.

6 CONCLUSION

We implemented an open-source, data-driven event sequence visualization tool in Python using the Plotly package [10]. The tool utilizes machine learning classifiers and sequential feature selection techniques to uncover the most important variables for a given outcome and present them to the user. Specifically, we demonstrate the use of sequential feature selection and a novel feature selection chain approach to this problem. Development was carried out with the US Rectal Cancer Consortium dataset in mind. We targeted the problem of anastomotic leakage in rectal cancer surgery, which is a critical treatment complication. Minimal data processing was carried out to facilitate the proper development of the tool. A rectal cancer surgeon stakeholder was consulted throughout the development process, and their feedback was used to evaluate the tool.

We found that the Sankey-based visualization is a unique and informative way to present information about the incidence of anastomotic leakage that has potential in the day-to-day battle against rectal cancer. Our stakeholder met the tool with curiosity as well as a desire to explore other treatment outcomes and cofactors. Although the tool is much less complex than other existing work, there was still a learning curve for the stakeholder. It is also evident that the visualization can quickly get very long with many edges, making its interpretation more difficult. Careful attention must also be paid to how different levels of the response variable are indicated

throughout the chart. On the data processing approach, the stakeholder made several useful suggestions to better represent the data presented in the RCC, and to model cancer treatments as a whole. The feature selection techniques were more difficult to evaluate. Both approaches uncovered interesting features associated with leakage, but the chained method resulted in decreased predictive power within events and a diagram that was more complicated overall.

Future work should invest the time and effort into extracting as much information as possible from the RCC dataset. This is no menial task, but one that could lead the way for transformative innovations in rectal cancer treatment. Potential improvements to the presented visualization technique include a feature that allows the diagram to branch on a particular variable, tighter integration with predictive models, refined statistical metics, among others. The greatest improvement is likely to be found by implementing rule/pattern mining or clustering analysis in place of feature selection. In the future, deep learning models could be applied to enable full forecasts of cancer treatment plans and counterfactual analyses of different treatment options. Such a tool could consider more factors about an individual patient and their response to treatment than any lone clinican, select the important ones to show to a patient or surgeon, and allow them to use their best judgement on how to move forward.

REFERENCES

[1] 2024. Rectal Cancer Treatment (PDQ®) - NCI. https://www.cancer.gov/types/colorectal/hp/rectal-treatment-pdq Archive Location: nciglobal,ncienterprise.

[2] Danish Ali, Maria Syed, Adriana C. Gamboa, Alexander T. Hawkins, Scott E. Regenbogen, Jennifer Holder-Murray, Matthew Silviera, Aslam Ejaz, Glen C. Balch, and Aimal Khan. 2024. Association of omental pedicled flap with anastomotic leak following low anterior resection for rectal cancer. *Journal of Surgical Oncology* (Jan. 2024). https://doi.org/10.1002/jso.27572

[3] Bhavneet Bhinder, Coryandar Gilvary, Neel S. Madhukar, and Olivier Elemento. 2021. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discovery* 11, 4 (April 2021), 900–915. https://doi.org/10.1158/2159-8290.CD-21-0090

[4] Gretchen C. Edwards, Adriana C. Gamboa, Michael P. Feng, Roberta L. Muldoon, Michael B. Hopkins, Sherif Abdel-Misih, Glen C. Balch, Jennifer Holder-Murray, Maryam Mohammed, Scott E. Regenbogen, Matthew L. Silviera, and Alexander T. Hawkins. 2022. What's the magic number? Impact of time to initiation of treatment for rectal cancer. *Surgery* 171, 5 (May 2022), 1185–1192. https://doi.org/10.1016/j.surg.2021.08.032

[5] Adriana C. Gamboa, Rachel M. Lee, Michael K. Turgeon, Christopher Varlamos, Scott E. Regenbogen, Katherine A. Hrebinko, Jennifer Holder-Murray, Jason T. Wiseman, Aslam Ejaz, Michael P. Feng, Alexander T. Hawkins, Philip Bauer, Matthew Silviera, Shishir K. Maithel, and Glen C. Balch. 2021. Impact of Postoperative Complications on Oncologic Outcomes After Rectal Cancer Surgery: An Analysis of the US Rectal Cancer Consortium. *Annals of Surgical Oncology* 28, 3 (March 2021), 1712–1721. https://doi.org/10.1245/s10434-020-08976-8

[6] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. 2022. Survey on Visual Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (Dec. 2022), 5091–5112. https://doi.org/10.1109/TVCG.2021.3100413 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[7] David N. Hanna, Adriana C. Gamboa, Glen C. Balch, Scott E. Regenbogen, Jennifer Holder-Murray, Sherif R. Z. Abdel-Misih, Matthew L. Silviera, Michael P. Feng, Thomas G. Stewart, Li Wang, and Alexander T. Hawkins. 2021. Perioperative Blood Transfusions Are Associated With Worse Overall Survival But Not Disease-Free Survival After Curative Rectal Cancer Resection: A Propensity Score-Matched Analysis. *Diseases of the Colon and Rectum* 64, 8 (Aug. 2021), 946–954. https://doi.org/10.1097/DCR.0000000000002006

[8] Paul T. Hernandez, Raj M. Paspulati, and Skandan Shanmugan. 2021. Diagnosis of Anastomotic Leak. *Clinics in Colon and Rectal Surgery* 34, 6 (Nov. 2021), 391–399. https://doi.org/10.1055/s-0041-1735270

[9] Na Hong, Chun Liu, Jianwei Gao, Lin Han, Fengxiang Chang, Mengchun Gong, and Longxiang Su. 2022. State of the Art of Machine Learning–Enabled Clinical Decision Support in Intensive Care Units: Literature Review. *JMIR Medical Informatics* 10, 3 (March 2022), e28781. https://doi.org/10.2196/28781 Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[10] Plotly Technologies Inc. 2015. Collaborative data science. https://plot.ly Place: Montreal, QC Publisher: Plotly Technologies Inc..

[11] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare* 1, 1 (Jan. 2020), 1–20. https://doi.org/10.1145/3344258

[12] Simi Job, Xiaohui Tao, Lin Li, Haoran Xie, Taotao Cai, Jianming Yong, and Qing Li. 2024. Optimal Treatment Strategies for Critical Patients with Deep Reinforcement Learning. *ACM Transactions on Intelligent Systems and Technology* (Feb. 2024). https://doi.org/10.1145/3643856 Just Accepted.

[13] Jessica M. Keilson, Adriana C. Gamboa, Michael K. Turgeon, Lillias Maguire, Katherine Hrebinko, Jennifer Holder-Murray, Jason T. Wiseman, Aslam Ejaz, Alexander T. Hawkins, Ebunoluwa Otegbeye, Matthew Silviera, Shishir K. Maithel, and Glen C. Balch. 2023. Is There a Role for Adjuvant Chemotherapy in Pathologic Node-Negative Locally Advanced Rectal Cancer After Neoadjuvant Chemoradiation Therapy? *Annals of Surgical Oncology* 30, 1 (Jan. 2023), 224–232. https://doi.org/10.1245/s10434-022-12432-0

[14] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 11 (Nov. 2018), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5 Number: 11 Publisher: Nature Publishing Group.

[15] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 299–309. https://doi.org/10.1109/TVCG.2018.2865027 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[16] Ruikai Li, Chi Zhang, Kunli Du, Hanjun Dan, Ruxin Ding, Zhiqiang Cai, Lili Duan, Zhenyu Xie, Gaozan Zheng, Hongze Wu, Guangming Ren, Xinyu Dou, Fan Feng, and Jianyong Zheng. 2022. Analysis of Prognostic Factors of Rectal Cancer and Construction of a Prognostic Prediction Model Based on Bayesian Network. *Frontiers in Public Health* 10 (2022). https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.842970

[17] Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. 2020. Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. *Journal of Medical Internet Research* 22, 7 (July 2020), e18477. https://doi.org/10.2196/18477 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[18] Tyler J. Loftus, Patrick J. Tighe, Amanda C. Filiberto, Philip A. Efron, Scott C. Brakenridge, Alicia M. Mohr, Parisa Rashidi, Gilbert R. Upchurch, Jr, and Azra Bihorac. 2020. Artificial Intelligence and Surgical Decision-making. *JAMA Surgery* 155, 2 (Feb. 2020), 148–158. https://doi.org/10.1001/jamasurg.2019.4917

[19] Alisha Lussiez, Samantha J. Rivard, Kamren Hollingsworth, Sherif R. Z. Abdel-Misih, Philip S. Bauer, Katherine A. Hrebinko, Glen C. Balch, and Lillias H. Maguire. 2023. Management and Outcomes of Pathologic Upstaging of Clinical Stage I Rectal Cancers: An Exploratory Analysis. *Diseases of the Colon and Rectum* 66, 4 (April 2023), 543–548. https://doi.org/10.1097/DCR.0000000000002225

[20] Chris McIntosh, Leigh Conroy, Michael C. Tjong, Tim Craig, Andrew Bayley, Charles Catton, Mary Gospodarowicz, Joelle Helou, Naghmeh Isfahanian, Vickie Kong, Tony Lam, Srinivas Raman, Padraig Warde, Peter Chung, Alejandro Berlin, and Thomas G. Purdie. 2021. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nature Medicine* 27, 6 (June 2021), 999–1005. https://doi.org/10.1038/s41591-021-01359-w Number: 6 Publisher: Nature Publishing Group.

[21] Wes McKinney and others. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[23] scikit-learn developers. [n. d.]. 1.13. Feature selection. https://scikit-learn/stable/modules/feature_selection.html

[24] Shengli Shao, Yufeng Zhao, Qiyi Lu, Lu Liu, Lei Mu, and Jichao Qin. 2023. Artificial intelligence assists surgeons' decision-making of temporary ileostomy in patients with rectal cancer who have received anterior resection. *European Journal of Surgical Oncology* 49, 2 (Feb. 2023), 433–439. https://doi.org/10.1016/j.ejso.2022.09.020

[25] R. Sharmila and K. Gayathri. 2023. The Transformative Impact of Precision Oncology across Diverse Cancer Types. (Dec. 2023). https://doi.org/10.5281/

ZENODO.10250972 Publisher: [object Object] Version Number: 1.

[26] Yu Shen, Li-Bin Huang, Anqing Lu, Tinghan Yang, Hai-Ning Chen, and Ziqiang Wang. 2024. Prediction of symptomatic anastomotic leak after rectal cancer surgery: A machine learning approach. *Journal of Surgical Oncology* 129, 2 (2024), 264–272. https://doi.org/10.1002/jso.27470 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jso.27470.

[27] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581075

[28] Michael K. Turgeon, Adriana C. Gamboa, Jessica M. Keilson, Jeffrey Maniko, Lillias Maguire, Katherine Hrebinko, Jennifer Holder-Murray, Jason T. Wiseman, Sherif Abdel-Misih, Saif Hamdan, Alexander T. Hawkins, Philip Bauer, Matthew Silviera, Shishir K. Maithel, and Glen C. Balch. 2021. Radiological assessment of persistent retroperitoneal and lateral pelvic lymph nodes after neoadjuvant therapy for rectal cancer: An analysis of the United States Rectal Cancer Consortium. *Journal of Surgical Oncology* 124, 5 (Oct. 2021), 818–828. https://doi.org/10.1002/jso.26600

[29] Michael K. Turgeon, Adriana C. Gamboa, Scott E. Regenbogen, Jennifer Holder-Murray, Sherif R. Z. Abdel-Misih, Alexander T. Hawkins, Matthew L. Silviera, Shishir K. Maithel, and Glen C. Balch. 2021. A US Rectal Cancer Consortium Study of Inferior Mesenteric Artery Versus Superior Rectal Artery Ligation: How High Do We Need to Go? *Diseases of the Colon and Rectum* 64, 10 (Oct. 2021), 1198–1211. https://doi.org/10.1097/DCR.0000000000002052

[30] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445432

[31] Rongbo Wen, Kuo Zheng, Qihang Zhang, Leqi Zhou, Qizhi Liu, Guanyu Yu, Xianhua Gao, Liqiang Hao, Zheng Lou, and Wei Zhang. 2021. Machine learning-based random forest predicts anastomotic leakage after anterior resection for rectal cancer. *Journal of Gastrointestinal Oncology* 12, 3 (June 2021), 921–932. https://doi.org/10.21037/jgo-20-436

[32] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[33] Chaoran Yu and Ernest Johann Helwig. 2022. The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. *Artificial Intelligence Review* 55, 1 (Jan. 2022), 323–343. https://doi.org/10.1007/s10462-021-10034-y

[34] Chao Yu, Jiming Liu, and Hongyi Zhao. 2019. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Medical Informatics and Decision Making* 19, 2 (April 2019), 57. https://doi.org/10.1186/s12911-019-0763-6

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

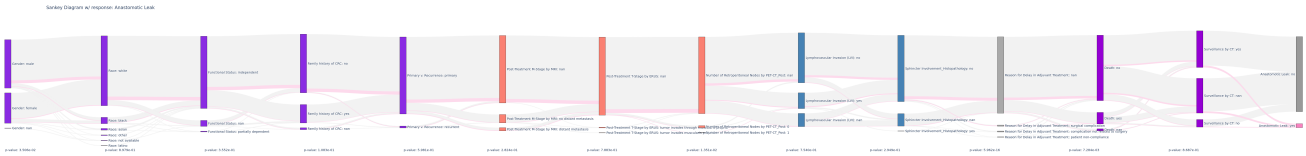

Figure 3: Visualization of features selected by an SFS chain, resulting in a much longer diagram but one that represents more of the event sequence. PS: Couldn't get this to render sideways :(