A Tale of Two Centers: Visual Exploration of Health Disparities in Cancer Care

Sanjana Srabanti*
University of Illinois at Chicago

Michael Tran
University of Illinois at Chicago

Virginie Achim
University of Illinois at Chicago

David Fuller
University of Texas

Guadalupe Canahuate
University of Iowa

Fabio Miranda
University of Illinois at Chicago

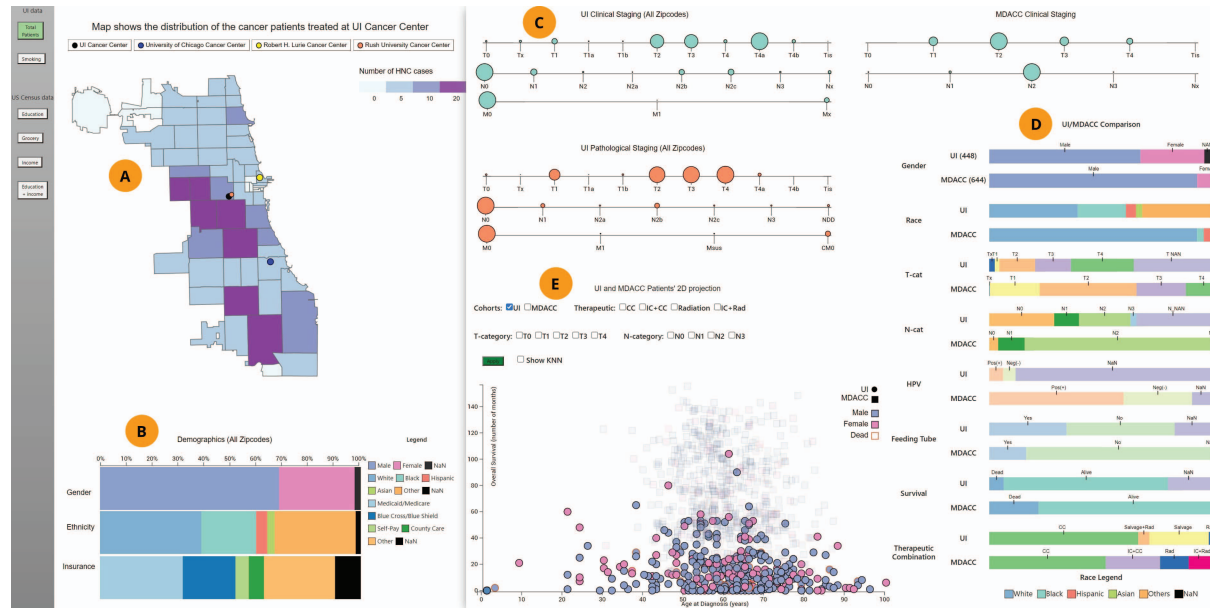G.Elisabeta Marai
University of Illinois at Chicago

Figure 1: A tale of two head and neck cancer centers, UI and MDACC. (A) A geospatial panel shows the number of patients treated at UI, along with the other centers in the area and census data; the data is not normalized by population, to better illustrate catchment. Most patients come from just a few neighborhoods. (B) Demographics pane, showing a diverse population and very few patients without known medical insurance. (C) Disease staging panel for the two populations, showing UI patients report later for diagnosis and treatment. (D) Summarization panel comparing the two populations based on demographic and disease characteristics, treatments, and outcomes. (E) Scatterplot panel comparing patients with respect to their age at diagnosis (x-axis) and overall survival (y-axis). In this view, the MDACC patients are filtered out via transparency, to allow better focus on the UI population.

ABSTRACT

The annual incidence of head and neck cancers (HNC) worldwide is more than 550,000 cases, with around 300,000 deaths each year. However, the incidence rates and disease-characteristics of HNC differ between treatment centers and different populations, due to undetermined reasons, which may or not include socioeconomic factors. The multi-faceted and multi-variate nature of the data in the context of the emerging field of health disparities research makes automated analysis impractical. Hence, we present a visual analysis approach to explore the health disparities in the data of HNC patients from two different cohorts at two cancer care centers. Our approach integrates data from multiple sources, including census data and city data, with custom visual encodings and with a nearest neighbor approach. Our design, created in collaboration with oncology experts, makes it possible to analyze the patients' demographic, disease characteristics, treatments and outcomes, and to make sig-

nificant comparisons of these two cohorts and of individual patients. We evaluate this approach through two case studies performed with domain experts. The results demonstrate that this visual analysis approach successfully accomplishes the goal of comparing two cohorts in terms of different significant factors, and can provide insights into the main source of health disparities between the two centers.

Index Terms: Biomedical and medical visualization—Design studies—Hypothesis forming

1 INTRODUCTION

The incidence of head and neck cancer (HNC), which accounts for more than 330,000 deaths annually [45], is not uniform across populations. Twice as many men are affected by HNC than women [45]. HNC is also often diagnosed among individuals over the age of 50, is associated with the use of tobacco and alcohol [9], and exhibits increasing rates due to the wider prevalence of the human papillomavirus (HPV) in recent years. At the same time, whereas demographic and clinical factors like gender, age, HPV status, and tobacco usage are reasonably well studied in the context of HNC, socioeconomic and cultural factors also affect individual health [40], yet are less understood. These factors disproportionately affect spe-

*e-mail: ssraba2@uic.edu

101

cific ethnic groups. For example, Black and White patients in the US have similar HNC incidence rates of 14.3 and 12.2 per 100,000 people, respectively [16]. In contrast, the five-year survival rate among Black HNC patients is around 29%-31% compared to 55%–59% in White patients. Furthermore, cancer patients with no insurance or low income have in general poorer survival rates than higher income patients [13], although the insurance impact on HNC outcomes is unknown. Additionally, clinicians collect further detailed patient information during diagnosis and treatment, from demographics to disease descriptors related to the cancer stage and disease extent in the body, to the treatment applied, and to the treatment outcomes. And yet, it is not known how these factors, if any, influence patient outcomes, and detailed HNC data has never before been analyzed across multiple cancer-care centers serving significantly different populations.

To help answer some of these questions in the emerging field of health disparities research, we describe the design of an exploratory data visual analysis approach to support the study of HNC patient data across two cancer treatment centers, the University of Illinois Health System (UI), and the University of Texas M.D. Anderson Cancer Center (MDACC). The main contributions of this work are: 1) a description and characterization of health disparities research, with emphasis on the data and activities involved in this field; 2) the multi-site collection and harmonization of two HNC datasets from two different institutions, in remote collaboration with two teams of oncologists; 3) the design of a visual analysis tool for the HNC data from these two centers, including disparities-driven analysis in terms of demographics, disease characteristics, provided treatments and outcomes; 4) the implementation of this approach into a system that integrates data from multiple sources; the system blends custom visual encodings with a nearest neighbors approach, and supports the rapid formulation and testing of hypotheses; and 5) the evaluation of this system with domain experts, and a discussion of the lessons learned through this experience.

2 RELATED WORK AND BACKGROUND

Head and neck cancer background. Previous works have explored the basic HNC demographic and disease characteristics in different populations. Clarke et al. [12] analyzed the connection between race and the rural-urban context in HNC survival. The result of this study reflected that White urban patients have the highest survival rate (67 months) whereas Black rural patients have the lowest (35.1 months). Gourin et al. [20] took additional variables into account, such as race, gender, age, primary tumor site, tumor stage, and smoking status, as well as socioeconomic data provided by the US Census Bureau, and showed that Black patients had significantly worse survival compared with White patients. Statistical data also show that Black patients have more advanced stage disease than White patients at the start of treatment. Moreover, White patients had a more secure socioeconomic status than Black patients. Furthermore, Xiao et al. have sought to measure tumor progression and have shown that increased time to treatment can worsen patient survival [51], although this aspect has not been considered before in HNC health disparities research.

Whereas these previous projects have analyzed tumor progression and how race and rural-urban status can significantly impact HNC survival outcomes, they did not seek to examine a number of other variables that may impact timely diagnosis and effective treatment for disadvantaged populations, such as insurance status, or tumor, node, metastasis (TNM) staging at diagnosis time, which can play an important role in the ultimate treatment of HNC. They also did not seek to analyze treatment and outcomes variability across centers.

Cohort analysis and visualization. Previous works in cohort visual analysis have used choropleth maps and treemaps to show cancer rates in the US [25], while simple bar charts, stacked bar charts, and line charts are typically used to compare demographic

status, gender, and race cancer statistics [10, 24, 52]. Similar representations have been used to visualize patients' treatment process and cohort summary [30, 48], and as a way to compare different cohorts [43]. Although our project does not include temporal data, other visualization projects have dealt with temporal cohort data analysis [17, 27, 28, 53]. Other works have dealt with visualization for clinicians [3, 23, 37, 38, 46, 54], but not with health disparities.

Whereas many works have dealt with visualization for clinical domains, including cohort data visualization, health disparities research features different requirements. In terms of data, we leverage the same intuitive visual representations, but now taking into account not only the patient medical information, but also socioeconomic and geolocation factors. In terms of tasks, we focus on the disparities-driven analysis of different populations.

Multivariate data visualization and comparison. Different strategies have been proposed to visualize multivariate data, including glyph-based techniques [8], animation [42], flexible linked axes [11], and combining parallel coordinates with MDS-based projections [22]. These approaches, however, present two main problems: they require high visual literacy and are constrained to the analysis of a data subset. Our work tackles these challenges by enabling the user to interact with the data through familiar visual encodings.

In terms of data comparison, pie charts are widely used to compare data, although studies disagree on their merit [4, 26]. Other approaches for data comparison include barcharts and stacked bar charts, coordinated multiple views, and focus+context views [29, 32]. Inspired by these works, we use a side-by-side comparison view approach with custom compact encodings.

3 METHODS

3.1 Data Harmonization

Data harmonization seeks to bring together various types, levels, and sources of data, in such a way that they can be made compatible and comparable [18]. Unlike standardization, which is not realistic in collaborative contexts [19], harmonization does not impose a single methodology or norm, but rather permits the pooling of data collected in different ways, thus reducing study heterogeneity [21]. In this project, data harmonization encompasses the efforts to harmonize intra-dataset the entries in that dataset, and to harmonize inter-dataset the two centers with respect to each other.

Data of 448 HNC patients treated at University of Illinois Health System, in Chicago, IL (UI) was collected over several years. Under an existing institutional review board (IRB) protocol, we manually transferred the data, through a laborious process, into Research Electronic Data Capture (REDCap), a web-based application developed by Vanderbilt University to capture data for clinical research and to create databases and projects. The patient data was de-identified and anonymized with dummy IDs, and then exported to Excel.

We faced several challenges during the harmonization process for the UI dataset, because the data had been collected using different methodologies. As a result, first, we encountered irregular document structure in the UI Health medical records, where specific data was sometimes entered in a different box than expected. Second, multiple nurses and doctors inputted data without following a common note template; as a result, some RedCap fields were not filled in the UI Health System. Last, the data featured multiple staging criteria, and sometimes the taxonomy used was not specified (e.g., clinical vs. pathological). Three transcribers completed this transcription process over the course of two months, under expert supervision from an HNC surgical team and our visual computing team.

Data of 644 HNC patients treated at MD Anderson Cancer Center at the University of Texas (MDACC) had also been collected over several years, and made available to us in Excel format via an institutional Material Transfer Agreement. This dataset was also de-identified and anonymized with dummy IDs. We reconciled these
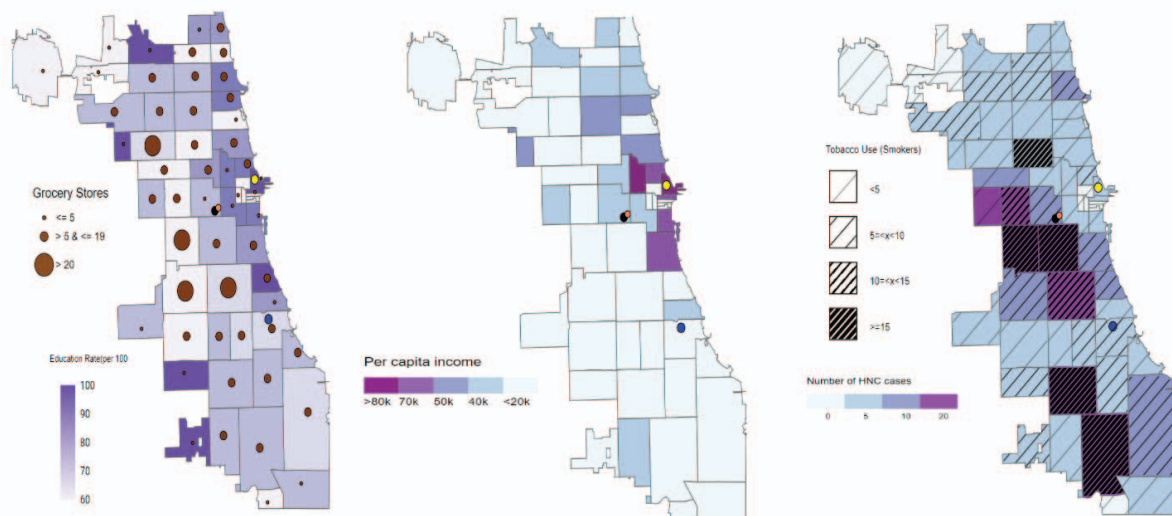
Figure 2: Left: Education rate and grocery store information. Middle: Per capita income of the population. Most patients come from areas where the education rates and per capita income are low. Right: Overlay map showing a strong correlation between the total number of HNC patients and smoking statistics.

entries using a combination of automated (e.g., spell-checking) and manual checks, in collaboration with the MDACC experts.

In order to harmonize the two datasets with each other, and based on our prior experience with this type of data [36, 39, 49], we first labeled the data entries as pertaining either to patient demographics, disease characteristics, treatment characteristics, or outcomes. For each category, we then manually identified and reconciled common data subcategories wherever possible. We marked data subcategories appearing in only one dataset, for example, insurance status which appeared only in the UI dataset, or specific treatment options that appeared only in the MDACC dataset. In some cases we computed new variables from the existing variables, e.g., we calculated UI survival length from recorded calendar dates. Last, we reconciled the units used to report the data. The two resulting multivariate datasets contain detailed information of each patient: demographic data (gender, ethnicity, age); insurance status and zip code (for UI dataset only); disease characteristics established during clinical examination (standard TNM clinical staging [tumor, node, metastasis], tumor site and subsite); disease characteristics as established during surgery/biopsy, i.e., pathological staging (for the UI dataset only); HPV status; smoking status information; as well as the therapeutic combination applied: CC concurrent chemo, IC induction chemo, Rad radiation, Salvage specific treatment to rescue organs (for the UI dataset only), CC + Rad, IC + Rad (MDACC dataset only); and the treatment outcomes recorded, such as survival (in months), known alive/dead status at the end of surveillance, and toxicity.

3.2 Design Process and Health Disparity Research

To design our analysis system, we followed an Activity-Centered-Design (ACD) paradigm, which is an extension of Human-Centered-Design (HCD) with emphasis on the user activities. We followed this approach because ACD has a higher success rate in interdisciplinary collaboration, compared to HCD (63% compared to 25%) [35]. Because this project was the first time this type of HNC data was compared between two centers, as well as the first time the UI dataset was being analyzed, the main user activities were centered around first gaining an understanding of the UI dataset, and then on comparing the UI and MDACC datasets. To implement the ACD paradigm, we completed a series of semi-structured interviews with the UI and with the MDACC healthcare providers. These
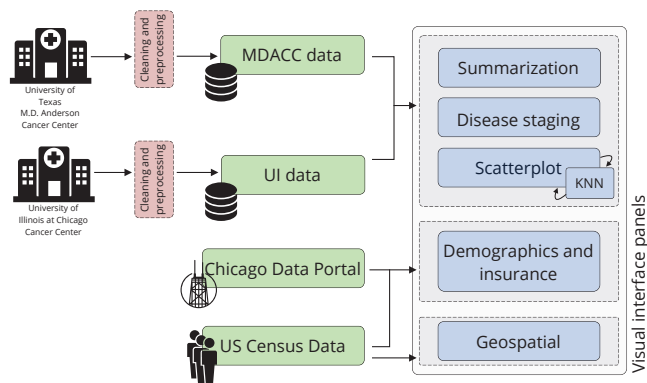

Figure 3: Data sources and architecture of our system. Our solution integrates data from two cancer care centers, MDACC and UI, as well as data from the Chicago Data Portal and from the US Census Bureau. These datasets and their specific aspects relevant to our problem are then visualized by different panels in our system interface.

providers included two senior clinicians at UI, an oral surgeon and a radiation therapist, and their respective groups, and a team of four radiation oncologists, including a senior clinician, at MDACC. We used these interviews to extract information about the project requirements, captured as activity-reflecting functional specifications and as nonfunctional requirements, and to better understand the data.

We furthermore read extensive documentation, and attended webinars and lectures to educate ourselves about the new domain of health disparities research. In brief, many populations, whether defined by race, ethnicity, immigrant status, disability, sex, gender, or geography, experience higher rates of certain diseases and more deaths and suffering from them compared with the general population. Health disparities research seeks to reduce the profound disparity in health status of racial and ethnic minority, rural, low-income, and other underserved populations, through the traditional examination of biological factors, but also through an increased focus on non-biological factors such as socioeconomics, politics, discrimination, culture, and environment in relation to health disparities.

In our particular instance, which focused on HNC at two centers, the main user activities were centered around specific questions related to: viewing the geographical data distribution of the UI patient population (A1); understanding the population demographics at the two centers (A2); exploring the disease characteristics in terms of disease staging between the two centers (A3); exploring the socioeconomic factors potentially affecting the UI population (A4); and then analyzing the therapies applied to the two populations (A5) and the overall outcomes of treatment (A6). Further user activities related to the comparison of the two centers tended to be patient specific and detail-oriented: once the two teams gained an understanding of the overall patient population distribution and of the differences between the two datasets, they tended to focus on individual UI patients and their treatment and outcomes compared to the MDACC dataset (A7). Because the two teams speculated that the MDACC patients were faring better in their outcomes and benefitted from better socioeconomic factors than the UI population, the analysis did not seek to compare individual MDACC patients against the UI dataset. In other words, activity A7 is not bi-directional.

We used parallel prototyping [15] to design the visual analysis system, in order to explore a more divergent and diverse set of ideas, and both low fidelity and high-fidelity prototypes. According to the ACD paradigm, we held frequent meetings with the provider teams to collect feedback, to revise, then refine and evaluate these prototypes. Because of the exploratory nature of the project, the nonfunctional requirements were particularly soft, and were more or less limited to visual encodings and interactions appropriate for viewers who had low visualization literacy.

Our final top-design consists of a set of coordinated multiple views, which can help establish correlations among the different data facets [34], and which can support a variety of workflows, from Overview-First [44] to Details-First [31]. The data sources and datasets we use are illustrated in Fig. 3. The top-design is divided into panels serving the main user activities: 1) a geospatial map showing the UI dataset distribution, along with important socioeconomic factors, and a demographics pane; 2) a disease staging panel comparing disease staging characteristics between the two datasets, using a novel rod-knot encoding; 3) a scatterplot panel comparing the two datasets in terms of disease characteristics, treatment, and outcomes, as well as individual patients; and 4) a summarization panel providing a comprehensive comparison of the two datasets. Below we describe each panel in detail.

Our visual analysis system is built using Javascript with the visualization library D3.js [6]. Whereas similar dashboards could also be built using other environments, their design would have to tackle the same challenges we describe and address in this work, with respect to domain characterization, harmonization, encoding design, implementation, and evaluation.

3.2.1 Geospatial Panel

To support the analysis of the geographical data distribution of the UI patient population, the geospatial panel includes a choropleth map of Chicago and several overlaid filters (Fig. 1.A). The background map is color-coded according to the number of patients treated at UI. To better illustrate the catchment area for UI, four markers indicate the four main cancer centers in the city: the University of Illinois Health System, the University of Chicago Medical Center, the Robert H. Lurie Comprehensive Cancer Center of Northwestern University, and the Rush University Cancer Center. Because the two teams understood the overall population distribution in the city, with higher densities downtown and lower in the suburbs, and the main question was about catchment, not cancer incidence rates, the data was not normalized by zip code population. Details on demand show the total patients for a specific zip code.

To capture socioeconomic factors for this population, we integrate the following data from the US Census [2] and from the Chicago

Data Portal [1], relevant to the main questions asked by our collaborators: education, income, and grocery store availability (Fig. 1.B and Fig. 2). These data are not collected during patient care, so we integrate the census and city portal information based on the patient zip code data. The education and income factors are shown as colored transparent overlays, via darker shade (higher education), respectively a divergent scheme (higher/lower income using $80K, respectively $20K as thresholds). A bivariate colormap indicates education and income simultaneously (Fig. 4). Furthermore, because our collaborators speculated on the potential correlation between a wholesome diet and cancer incidence and outcomes, we incorporated grocery store information from the Chicago Data Portal [1], using scaled glyphs. Finally, because of a known correlation between smoking and HNC [14], a texture overlay shows the number of HNC smokers for each zip code, calculated based on the UI dataset.

A compact demographics pane shows the demographic and insurance characteristics of the UI cohort. In this pane, a rotated stacked bar chart shows the demographic (gender and ethnicity) and insurance status of the UI HNC patients. Depending on the neighborhood selected in the geospatial map, the pane gets updated to show the information specific to that zip code.

3.2.2 Disease Staging Panel

To support further investigation of the patient data, the disease staging panel displays clinical and pathological cancer staging for the UI cohort and clinical staging for the MDACC cohort (Fig. 1.C). The staging is reported using the TNM classification method, which classifies tumors and disease spread by its size, indicated by T0 - Tis, where the higher the T number, the larger the tumor and/or the more it has grown into nearby tissues or spread to other organs; for example, numbers 0 to 4 indicate increasing tumor size and extent (T1 describes a small tumor, and T4 describes a bigger tumor); Tx means unknown, and Tis reports the pre-cancerous changes or early stages of cancer. Degree of the spread through the regional lymph node network is indicated by N0 – Nx (the higher the N number, the greater cancer spread to nearby lymph nodes). N0 signifies no tumor in the lymph nodes, Nx is unknown, and the numbers 1 to 3 in N1 - N3 stand for location and number of affected regional lymph nodes. The T-category and N-category are further divided into subcategories a, b, and c to provide more detailed HNC information. Finally, TNM reports the extent of spread to organs by M0 - Mx (M0 indicates no distant metastases have been found, M1 means there are distant metastases, and Mx is unknown).

To easily support this type of aggregate comparison, we created a custom rod-knot encoding, after exploring and discarding multiple alternative encodings such as bar charts, parallel coordinate plots, scatterplot matrices, glyph-based diagrams [33, 36], and parallel marker plots [50]. In this encoding inspired by bubble plots, each horizontal rod represents the increasing stages of T-category (T0-Tis), N-category (N0-Nx), and M-category (M0-Mx). To represent the TNM staging of the MDACC cohort, we have only used T and N-category, as the M-category data is already reflected in the T and N staging, and so it was not collected separately at MDACC.

Knot markers are placed on each rod at each category represented in the data. The knot size indicates the total number of patients in that category. Color is used to distinguish between pathological and clinical staging. This rod-knot encoding was deemed more effective by our collaborators in terms of comparing the two datasets, relative to the other encodings considered, despite its two-dimensional characteristics. The disease characteristics can also be shown for each zip code of the UI dataset, supporting further detailed analysis.

3.2.3 Scatterplot Panel

A custom scatterplot panel encodes each patient's position as a 2D projection determined by overall survival (number of months) against age at diagnosis (years) (Fig. 1.E and Fig. 5). In the order
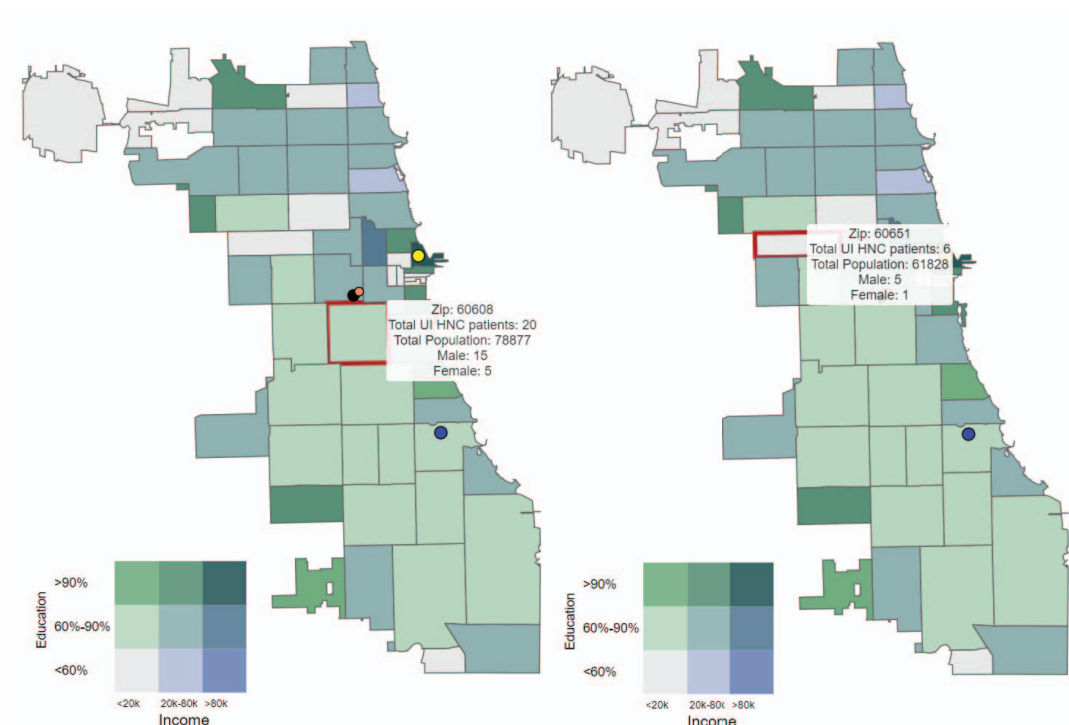
Figure 4: Two neighborhoods in the UI dataset, both within the catchment area of the UI Cancer Center. The neighborhoods feature similar education and income levels (bivariate color map) and population density, yet they have significantly different HNC incidence rates.
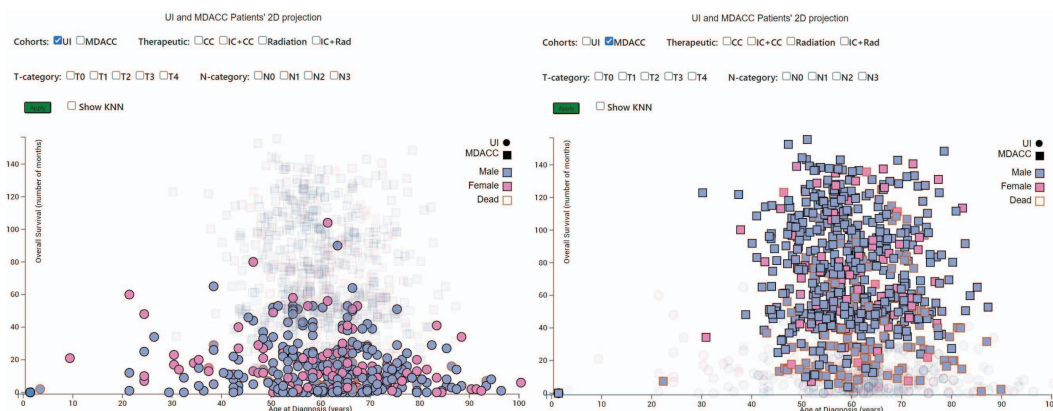

Figure 5: For the same disease characteristics (T-category and N-category) and treatment (CC), UI patients' survival rate (left), and MDACC patients' survival rate (right). We note that the UI cohort features shorter surveillance periods than the MDACC cohort, and a wider age range, including a few very young patients (shown close to the origin).

of further importance to our collaborators, the marker shapes distinguish between the UI and MDACC patients, and colors encode gender. A highlight border indicates deceased patients. Several filters enable filtering based on the dataset, disease characteristics (T and N, as M is already reflected in these two), and the therapy applied (Fig. 5). Data filtered out is rendered transparent, for better context. Additional data about gender, disease, treatment, HPV status, race, and toxicity (i.e., the necessity of inserting surgically a feeding tube) are available as details on demand.

As soon as the clinicians had an understanding of the main differences between the two populations and debunked several hypotheses, they returned to their main focus, which is on providing care to patients in the clinic. Specifically, the clinicians were interested in analyzing a specific UI patient, and determining whether, once di-

agnosed, their treatment and outcomes would differ had they been seen at MDACC instead. To help test this what-if type of scenario, we provide the capability of retrieving, given a specific UI patient, the most similar patients in the MDACC cohort. The similarity is determined based on a subset of the following pre-treatment attributes: demographic (age, gender, and ethnicity), disease characteristics (T-category, and N-category), tumor subsite (base of tongue, tonsil, soft palate, or not otherwise specified NOS), and HPV status if known (positive, negative). Smoking status was not included, due to the significant discrepancies between the two datasets along this attribute. To calculate similarity, we converted all categorical values to numerical values, and then applied the K-nearest neighbor (KNN) algorithm using Hamming distance, as all the patients' attributes except for the age are discrete values having no intrinsic order relation.

105

We decided to use Hamming distance where each attribute has equal weights instead of using variable weights for the attributes because the measure was well suited for this study, and the Euclidean and Manhattan distance metrics are not appropriate in this case. The number of possible values of the categorical variables considered are in the same scale (between 2-4 categories/attributes), so there was also no need to consider measures that normalize the similarity using the number of categories such as Eskin [5]. Likewise, the use of IOF, Goodall, or other measures that consider frequency and/or data distribution as weights for each attribute, would not only increase computation time but also introduce another layer of complexity since the KNN is performed between samples from two different centers and their distribution varies between centers. Their use would make the interpretation of the results more difficult in this case, because the use of the data distribution as weights could bias the results to the given samples. Last, the Gower measure [5] could have been an alternative to combine the categorical attributes and the numerical attribute, however, since only age was numeric, we decided Hamming would be a better metric. The Hamming distance between patients A and B is defined as:

$$Hamming(A,B) = \sum_{i=1}^{|X|} 1 \;\; iff \;\; (A_i \neq B_i)$$

where $|X|$ is the number of attributes, and A_i indicates the value of the ith attribute of patient A. As the Hamming distance is a discrete function with range $[0, |X|]$, the age difference is used as a tiebreaker. To this end, the normalized age difference (difference / range) was added to the Hamming distance, effectively breaking all ties. Since the difference over age is never greater than 1, its addition never supersedes the Hamming distance ranking computed over the categorical attributes; it just ranks the ties. The five patients with the smallest distance measures are selected as the nearest neighbors.

3.2.4 Summarization Panel

The comparison between UI and MDACC cohorts is further summarized using a compact set of rotated stacked bar charts (Fig. 1.D). We selected stacked bar charts instead of parallel coordinates plots, composite glyphs, pie charts, or donut charts because our practitioners had relatively low visual literacy, and the bars yield good comparison accuracy. Considering the space constraints, the stacked bar charts were rotated to maximize information density. We selected multiple qualitative color schemes from Color Brewer [7] to distinguish between the attribute values, and between attributes. We added explicit labels to the rotated stacked bar charts in order to make them more legible. To avoid the overlap of multiple informative labels, in the race chart we have repositioned the label legend below. Details on demand are available when hovering.

4 EVALUATION

Although ACD has a higher success rate than HCD in interdisciplinary settings [35], no design approach is failproof. Therefore we report here the results of two case studies we completed together with a group of domain experts, which included four radiation oncologists with clinical experience, an oral oncology surgeon, a data mining expert, and three visual computing researchers. Because the teams were in different locations and due to the COVID-19 restrictions, the evaluation process was conducted remotely with screen sharing and note-taking. The first author piloted the visual analysis interface under the direction of the oncologists and data mining expert.

4.1 Disease Characteristics and Treatment Investigation

The group started by exploring the top elements of the summarization panel, in order to get a high-level understanding of the two cohort demographics (activity A2). From this panel, it was evident

Figure 6: Comparison between UI and MDACC cohort based on demographic (Gender, Race, HPV status), disease characteristics (T-category, N-category), provided treatments and outcomes (survival and toxicity).
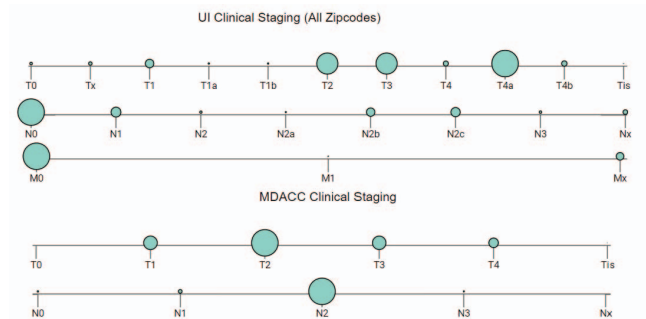

Figure 7: Clinical TNM staging of UI and MDACC HNC patients. UI patients receive treatment later than MDACC patients.

that the UI cohort included higher female representation (30.8%) than the MDACC cohort (12%). One of the oncologists immediately wondered whether this higher female representation was related to cancer incidence across ethnicities. The UI cohort was also strikingly more ethnically diverse than the MDACC one (Fig. 6).

Reaching the disease-stage summarization component in the panel (third row), the teams were surprised to notice a higher per-

centage of advanced cancer (T4) patients in the UI cohort than in the MDACC cohort (which was expected), but not larger spread to the lymph nodes (N-stage): there was a much larger number of MDACC patients with N2 stage compared to the UI cohort. This observation was confirmed beyond doubt by the rod-knot panel. The teams were furthermore surprised at the large proportion of UI patients with undetermined HPV status (summarization panel, fourth row). The group quickly confirmed that the UI patients had adequate medical insurance using the demographics pane, and, using the scatterplot panel, that most of both the MDACC and the UI HNC cohort were indeed older than 65, when US citizens become eligible for the federal medical insurance program. At this point, the investigation was paused, and a vivid discussion ensued.

Returning to the Summarization panel, a visual computing researcher commented on the further differences between the two cohorts in terms of Feeding Tube and Therapeutic Combination. Counter to a prior comment from the UI team noting that "At MDACC, they insert feeding tubes in all patients", the data shown was not supporting this statement: fewer MDACC patients had feeding tubes. This observation prompted an a-ha insight on the side of the oncologists: they speculated, and we later confirmed, that the UI cohort featured predominantly patients with oral HNC, which had been seen by the oral surgeon, whereas the MDACC cohort featured a wider distribution of HNC tumor locations, including oral, but also pharyngeal, laryngeal locations and so on. This observation helped explain to a certain extent the discrepancies in terms of the disease characteristics (T-staging and N-staging), HPV testing (not as relevant for oral cancer treatment), as well as in terms of the treatment applied between the two cohorts (more induction chemo, concurrent chemo, and radiation therapy at MDACC versus more surgery and more feeding tubes at UI), and helped motivate the KNN filtering approach based on tumor subsite described below.

4.2 Socioeconomic Factor Investigation

In this case study, the teams wished to investigate the role of various socioeconomic factors in HNC incidence and treatment outcomes. The investigation started with the geospatial panel to view the geographical data distribution of the UI patient population (A1) in relation to various factors (A4). The investigation confirmed that the catchment area of the UI center is primarily in a set of racially diverse, under-served neighborhoods adjacent to the center. The exploration (Fig. 4) then revealed higher HNC patient counts in neighborhoods that have low education rates in addition to low income. However, the UI team noted that some adjacent neighborhoods also featured combined low education and low income rates, yet did not seem to feature such high HNC incidence. Based on this observation, the two teams hypothesized that cigarette smoking increases the cancer rate, which they quickly confirmed: Fig. 2 shows that increased rates of cigarette smoking are directly proportional to HNC rates.

Encouraged by the previous findings, the analysis then moved promptly to the rod-knot plots of clinical staging of UI and MDACC cohorts to examine the disease characteristics (A3). The panel suggests that, beyond the tumor subsite location, MDACC HNC patients tend to seek medical attention in the earlier tumor stage (T2) compared to UI patients (T4) (Fig. 7). The MDACC team was confident that their patients had medical insurance coverage, most often through the federal insurance program for people over the age of 65. Indeed, most MDACC patients have T2 category tumors, whereas a large share of the UI dataset has T3 and T4 category tumors, although they also do not lack medical insurance.

Upon investigation of specific zipcodes with different ethnic distributions, it furthermore became apparent that the pathological UI assessment was, fortunately, less negative than the initial UI clinical assessment (Fig. 8). One neighborhood (60612) has a higher Black component than the other (60608), which features more Hispanics

and a higher incidence of unknown medical insurance coverage. Patients in 60612 seek treatment earlier in the disease course (stage T1 or T2) compared to patients in zip code 60608 (stage T2 or T3). Furthermore, pathology analysis (orange knots) on patients in 60612 indicates primarily no nodal involvement (N0) and thus no metastasis (M0), and so, fortunately, does not confirm the initial clinical analysis of the lymph node spread (cyan knots). The pathology analysis on patients in 60608 is also more conservative than the initial clinical estimate. Overall, we found a common trend in which the clinical staging was higher than the pathological staging. In other words, the clinicians initially believed the disease was more severe than the actual surgical findings.

In the next stage, the analysis focused on the treatment outcomes and the possible determining factors. The teams noticed, in the Summarization panel, an apparently higher mortality rate among the MDACC patient population, which however they knew to be benefitting from better socioeconomic factors and earlier care. Indeed, overall, the MDACC population seemed to be doing better in terms of survival, based on a Scatterplot Panel analysis (Fig. 5). And yet, based on the UI tumor subsite location, the MDACC team was skeptical of both observations: they did not expect particularly high mortality among the UI oral cancer cohort, and they did not expect higher mortality among the MDACC HNC cohort either. Indeed, when analyzing the number of dead patients in the scatterplot for each cohort, it became apparent that the main difference between the two groups was a longer surveillance period for MDACC patients. In other words, MDACC patients return to the center for continued care and surveillance after the initial treatment, whereas UI patients are treated, but are not followed up on. Therefore, UI does not have long-term survival data, a problem known as left-censoring: Left-censoring occurs when we cannot observe the time when the event of interest (in this case, the death of the patient) occurred, due to the patient no longer being followed up. In fact, the UI cohort includes a large number of patients with unknown survival status (see Summarization panel), which may influence the mortality report.

Given the lack of further survival data, the teams then examined the grocery store data, in view of its potential to lead to better treatment outcomes, and were again surprised to see the UI cohort came from zip codes with significant availability of grocery stores. However, both the MDACC and the UI team were skeptical of this finding, in particular given the neighborhoods' correlation with smoking status. They speculated that the Chicago Portal report may have included in their count also convenience stores, which sell cigarettes and alcohol but have limited fresh produce offerings, and marked this hypothesis for future data collection and investigation.

Finally, the analysis focused on the individual patient care (A7). Using the KNN filter, the team investigated several specific UI patients, and determined that overall, once accounting for the tumor subsite, the standard of care was remarkably similar between the two centers. Table 1 and Table 2 show the detailed analysis of two example UI patients. The first patient, a 54 year old male Black patient with advanced (T4 and N2) cancer, does not have a large set of matches in the MDACC dataset. The most similar patient is the one labeled MDACC Patient2 in the Table, although there are several worse, but still reasonable MDACC matches in terms of patient and disease characteristics (remaining columns in the Table). With respect to this most similar patient, however, the treatment and outcome recorded are virtually indistinguishable between the two centers. Table 2 shows a second example, this time for a UI patient with reasonably good matches in the MDACC dataset, but no match in terms of race. Although the treatment is marked as Unknown, the Feeding Tube toxicity and survival outcomes indicate the patient must have been prescribed a similar treatment with the MDACC patients, who feature uniform treatment and outcomes.

Given that the standard of care appeared to be very similar between the two centers, the teams speculated that better patient out-
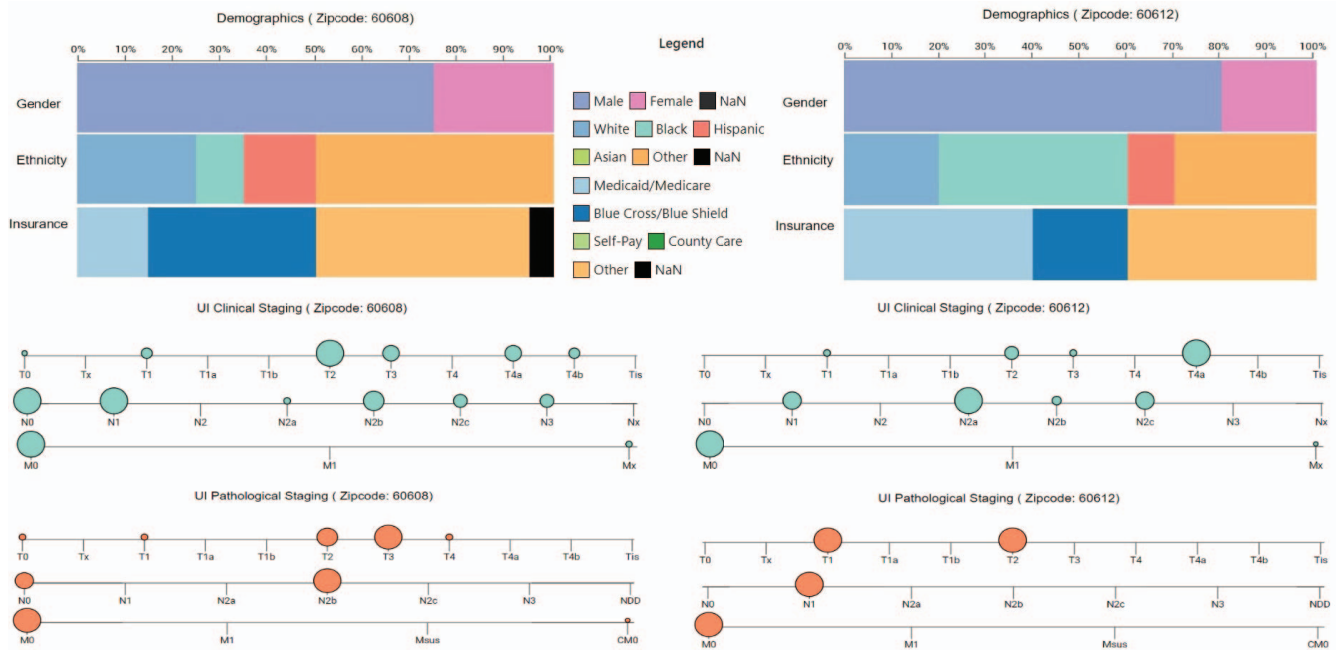
Figure 8: Comparison between two zip codes in terms of patient demographics and disease staging. In the top demographics pane, notice the different ethnicity distribution between the two zip codes, one with a higher Black component (60612) than the other (60608), which however has a higher proportion of Hispanics and a higher incidence of unknown medical insurance coverage. In the bottom rod-knot encoding, notice that patients in zip code 60612 seek treatment earlier in the disease course (stage T1 or T2) compared to patients in zip code 60608 (stage T2 or T3). Pathology analysis (orange knots) on patients in 60612 indicates primarily no nodal involvement (N0) and thus no metastasis (M0), and so, fortunately, does not confirm the initial clinical analysis of the lymph node spread (cyan knots). The pathology analysis on patients in 60608 is also more conservative than the initial clinical estimate.

Table 1: Attributes of a UI patient without a large set of good matches in the MDACC dataset, and the top 5 most similar MDACC patients.

Attributes	UI Patient	MDACC Patient1	MDACC Patient2	MDACC Patient3	MDACC Patient4	MDACC Patient5
Age	54	53.59	54.2	54.63	54.33	51.48
Gender	Male	Male	Male	Male	Male	Male
Race	Black	Black	Black	Black	Black	White
T-category	T4	T2	T4	T3	T2	T4
N-category	N2	N2	N2	N2	N2	N2
HPV Status	Unknown	Unknown	Negative	Positive	Positive	Negative
Treatment	CC	IC+Radiation alone	CC	IC+CC	CC	IC+CC
Feeding Tube	Yes	No	Yes	No	No	No
Survival Months	22	133.033	16.43	54.5	59.7	56.13
Dead/Alive	Alive	Alive	Alive	Alive	Alive	Alive

reach and education at the UI Center may help the center's HNC patients report earlier for diagnosis and treatment, and also help these patients return for continued care for longer periods. In turn, this could lead to improved quality of life, for example, by reducing the necessity for feeding tube insertions.

4.3 Expert Feedback

In addition to the system being successfully used for these case studies by domain experts, we have demonstrated the system to a group of 30 senior health researchers at the UI Cancer Center. Just like our close collaborators, these domain experts expressed enthusiastic appreciation for the system and its abilities, in particular with respect to its capacity to show so many different aspects of the data in a compact design, and the ability to integrate multiple data sources like the census and the city data portal, without overwhelming the viewer. Although we did not pursue specific inquiries with respect to encodings and interactions, which are by design simple and familiar, the experts were mesmerized by the data views. Whereas our MDACC collaborators posed a brief calibration question with re-

spect to the Chicago map population density and patient distribution, the wider UI group were immediately able to delve into scientific hypothesis forming, and to agree with our collaborators' insights. This constituted valuable testimony to the utility of these encodings. We note that the experts spent inordinate amounts of time examining and pointing to the data encodings, as opposed to interacting with the system, and the discussions were lively and the stream of hypothesis forming and testing extremely fast. All requests were related to further deployment of the system with other datasets.

5 DISCUSSION AND CONCLUSION

The case studies we report indicate that our integrated approach, which blends multiple data sources and datasets with legible visual encodings and easy interactions and with similarity-computing capabilities, can assist users not only in exploring these complex data, but also in generating novel hypotheses and insights. The domain experts were able to focus on the interesting aspects of the datasets, and quickly form and test hypotheses related to medical insurance coverage, the prevalence and importance of HPV negative status

Table 2: Attributes of a UI patient with good matches in the MDACC dataset, and the top 5 most similar MDACC patients.

Attributes	UI Patient	MDACC Patient1	MDACC Patient2	MDACC Patient3	MDACC Patient4	MDACC Patient5
Age	68	68.04	64.12	68.175	66.63	64.94
Gender	Male	Male	Male	Male	Male	Male
Race	Hispanic	White	White	White	White	White
T-category	T4	T4	T4	T4	T4	T2
N-category	N2	N2	N2	N2	N2	N2
HPV Status	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Treatment	Unknown	CC	CC	CC	IC+CC	IC+CC
Feeding Tube	Yes	Yes	No	Yes	Yes	Yes
Survival Months	23	13.97	21.5	85.133	95	124.8
Dead/Alive	Alive	Alive	Alive	Alive	Alive	Alive

in under-represented groups for this type of cancer, the type and breadth of treatment options and standard of care available at one center versus another, survival rates, and patient toxicity (feeding tube) outcomes. Our collaborators were even able to detect subtle differences in the data, like the oral focus of the UI dataset as opposed to the more general oro-pharyngeal focus of the MDACC dataset, although the initial data did not distinguish between tumor subsites. These operations of hypothesis forming and testing were made possible by our coordinated multi-view design and use of layering, and by our careful design of compact summarization panels, which allowed us to maximize the use of screen area on a variety of displays. In this implementation of the "eyes beat memory" principle, and of "placing the knowledge in the world", our collaborators had immediate access to the many facets of the datasets.

Because our approach did not rely on participatory design due to limited availability of the domain experts, and so our collaborators could not benefit from visual scaffolding [34] via regular interactions, we had extremely limited opportunities for complex visual encoding design. However, we had success with our design of a custom novel encoding for the disease staging data, in the form of the rod-knot plot. We speculate that the reason our collaborators took immediately to this encoding is its similarity to the popular two-dimensional bubble plots, our careful design of the encoding axes, and our use of clear, direct labeling. As a result, the encodings provided an effective comparative view of the TNM staging of the two cohorts.

Last but not least, in terms of our design experience, we note that, to a certain extent, our project appeared to be data-driven. Initially, the domain experts had few questions about specific aspects of the data, in line with a data-first design approach [41]. However, as we integrated more aspects of the datasets and additional datasets, a rapid loop of hypothesis generation and testing occurred, and in the end, our collaborators converged towards a typical scientist-driven workflow and specific activities, which tend to focus on the details of a specific datapoint [31]—in this case, treatment and outcomes for a specific patient. With this observation in mind, we would recommend the deliberate design of interfaces that allocate significant screen space to detailed datapoint analysis, not only to overviews in the form of aggregation or summarization, even in situations where the problem seems to be data-first motivated.

In terms of scalability, our summarization plots and the rod-knot plots scale well with larger datasets, up to any size, as long as there are no extremely under-represented groups included in the data. Such groups would require logarithmic scaling or a completely different type of representation in order to remain visible at any scale. The choropleth map overlays can also scale to arbitrary numbers of variables through our use of layering and separation. The custom scatterplot can also handle large datasets, although overplotting could become an issue for extremely large datasets. Our use of transparent filtering can alleviate some of these overplotting issues.

Last, in terms of generalizability and further deployment, our approach can extend to multiple patient datasets beyond HNC or cancer care, and to other population disparities problems with a spatial component, for example in fair transit access or fair access

to energy resources [47], as long as those problems and datasets feature similar categorical and quantitative data mixtures. With respect to incorporating other HNC datasets and further deployment, the existing design would support well the pairwise comparison of the UI cohort with other cohorts, assuming the other datasets could be standardized to the MDACC cohort. Harmonization of other datasets with respect to either cohort would entail the same challenges as described in this paper. Other cohorts could also be compared pairwise against the MDACC cohort, assuming insurance, geospatial, census and grocery store data are available for those cohorts. Simultaneous comparison of more than two HNC cohorts at a time, or the spatial comparison of multiple cohorts may require more modifications to the design, as well as identifying specific interesting questions where global visualization would help.

In conclusion, in this work we describe a visualization approach for a novel problem which incorporates new datasets and domain-specific questions and hypotheses in the realm of health disparities research. These new problem and questions pose significant design challenges due to a shift in the questions asked, from "show me data" initially to "which factors influence outcomes" and finally to "help me understand the precise details of this particular datapoint" in later stages of the scientific investigation. The multi-faceted and multi-variate nature of the data makes automated analysis impractical for such problems. We address this new problem through the activity-centered design of a system which integrates multiple datasets and data sources, with custom visual encodings and with a nearest neighbor approach. Created in collaboration with oncology experts, our resulting design supports both population-level analysis of the data, as well as detailed analyses of similar patients. Overall, our visual approach supports rapid, collaborative hypothesis formation and testing, and can help give insights into the complex problem of health disparities.

REFERENCES

[1] Chicago Data Portal. Available: `https://data.cityofchicago.org/`. Accessed on: August 17, 2020.

[2] US Census. Available: `https://www.census.gov/`. Accessed on: August 17, 2020.

[3] J. Bernard, D. Sessler, T. May, et al. A visual-interactive system for prostate cancer cohort analysis. *IEEE Comp. Graph. and Appl.*, 35(3):44–55, 2015.

[4] E. Bertini, N. Elmqvist, and T. Wischgoll. Judgment error in pie chart variations. In *Proc. Eurograph. Conf. on Vis.*, pp. 91–95, 2016.

[5] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proc. SIAM Intern. Conf. Data Mining*, pp. 243–254. SIAM, 2008.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *IEEE Trans. Vis. Comp. Graph.*, 17(12):2301–2309, 2011.

[7] C. A. Brewer. *Color Brewer*, October 2021.

[8] J. M. Chambers, W. S. Cleveland, B. Kleiner, et al. *Graphical methods for data analysis*. Chapman and Hall/CRC, 2018.

[9] L. Q. Chow. Head & neck cancer. *NE J. of Med.*, 382(1):60–72, 2020.

[10] K. K. Chui, J. B. Wenger, S. A. Cohen, et al. Visual analytics for epidemiologists: understanding the interactions between age, time, and disease with multi-panel graphs. *PLOS One*, 6(2):e14683, 2011.

[11] J. Claessen and J. Van Wijk. Flexible linked axes for multivariate data visualization. *IEEE Trans. Vis. Comp. Graph.*, 17(12):2310–16, 2011.

[12] J. A. Clarke, A. M. Despotis, R. J. Ramirez, et al. Head and neck cancer survival disparities by race and rural–urban context. *Cancer Epidem. Prevent. Biomarkers*, 29(10):1955–1961, 2020.

[13] P. Daraei and C. E. Moore. Racial disparity among the head and neck cancer population. *J. Cancer Ed.*, 30(3):546–551, 2015.

[14] H. S. Delivered. How cigarette smoke makes head & neck cancer more aggressive. *Oncology Times*, 2019.

[15] S. P. Dow, A. Glassco, J. Kass, et al. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. on Computer-Human Interaction*, 17(4):1–24, 2010.

[16] C. Fakhry, M. Krapcho, D. W. Eisele, et al. Head and neck squamous cell cancers in the united states are rare and the risk now is higher among white individuals compared with black individuals. *Cancer*, 124(10):2125–2133, 2018.

[17] C. Floricel, N. Nipu, M. Biggs, et al. THALIS: Human-machine analysis of longitudinal symptoms in cancer therapy. *IEEE Trans. Vis. Comp. Graph.*, 28(1):151–161, 2021.

[18] I. Fortier. *Data Harmonization: Basic principles and best practices*. Maelstrom Research, 2010.

[19] I. Fortier, P. R. Burton, P. J. Robson, et al. Quality, quantity and harmony: the datashaper approach to integrating data across bioclinical studies. *Int. Journal of Epidemiology*, 39(5):1383–1393, 2010.

[20] C. G. Gourin and R. H. Podolsky. Racial disparities in patients with head and neck squamous cell carcinoma. *The Laryngoscope*, 116(7):1093–1106, 2006.

[21] P. Granda and E. Blasczyk. Data harmonization. *Guidelines for Best Practice in Cross-Cultural Surveys, Uni. of Michigan*, 2011.

[22] H. Guo, H. Xiao, and X. Yuan. Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates. *IEEE Trans. Vis. Comp. Graph.*, 18(9):1397–1410, 2012.

[23] R. Guo, T. Fujiwara, Y. Li, et al. Comparative visual analytics for assessing medical records with sequence embedding. *Vis. Informatics*, 4(2):72–85, 2020.

[24] P.-J. Huang, H.-H. Lin, C.-C. Lee, et al. Comutplotter: a web tool for visual summary of mutations in cancer cohorts. *BMC Medical Genomics*, 12(5):1–7, 2019.

[25] N. Johnson and H. Amarasinghe. Epidemiology and aetiology of head and neck cancers. In *Head & neck cancer*, pp. 1–57. Springer, 2016.

[26] R. Kosara. Circular part-to-whole charts using the area visual cue. In *EuroVis*, pp. 13–17, 2019.

[27] J. Krause, A. Perer, and H. Stavropoulos. Supporting iterative cohort construction with visual temporal queries. *IEEE Trans. Vis. Comp. Graph.*, 22(1):91–100, 2015.

[28] B. C. Kwon, V. Anand, K. A. Severson, et al. DPVis: Visual analytics with hidden markov models for disease progression pathways. *IEEE Trans. Vis. Comp. Graph.*, 2020.

[29] A. Lex, M. Streit, C. Partl, et al. Comparative analysis of multidimensional, quantitative data. *IEEE Trans. Vis. Comp. Graph.*, 16(6):1027–1035, 2010.

[30] M. H. Loorak, C. Perin, N. Kamal, et al. Timespan: Using visualization to explore temporal multi-dimensional data of stroke patients. *IEEE Trans. Vis. Comp. Graph.*, 22(1):409–418, 2015.

[31] T. Luciani, A. Burks, et al. Details-first, show context, overview last: supporting exploration of viscous fingers in large-scale ensemble simulations. *IEEE Trans. Vis. Comp. Graph.*, 25(1):1225–1235, 2018.

[32] T. B. Luciani, B. Cherinka, D. Oliphant, et al. Large-scale overlays and trends: Visually mining, panning and zooming the observable universe. *IEEE Trans. Vis. Comp. Graph.*, 20(7):1048–1061, 2014.

[33] C. Ma, F. Pellolio, D. A. Llano, et al. Rembrain: Exploring dynamic biospatial networks with mosaic matrices and mirror glyphs. *Electronic Imaging*, 2018(1):060404–1, 2018.

[34] G. E. Marai. Visual Scaffolding in Integrated Spatial and Nonspatial Analysis. In *EuroVis Workshop Vis. Analyt. (EuroVA)*. The Eurographics Association, 2015.

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

[36] G. E. Marai, C. Ma, A. T. Burks, et al. Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots. *IEEE Trans. Vis. Comp. Graph.*, 25(4):1732–1745, 2019.

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

[38] L. Mica, C. Niggli, P. Bak, et al. Development of a visual analytics tool for polytrauma patients: proof of concept for a new assessment tool using a multiple layer sankey diagram in a single-center database. *World journal of surgery*, 44(3):764–772, 2020.

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

[40] National Institutes of Health. *National Institute on Minority Health and Health Disparities*, June 2021.

[41] M. Oppermann and T. Munzner. Data-first visualization design studies. In *IEEE BELIV Workshop*, pp. 74–80. IEEE, 2020.

[42] P. Ordóñez, M. DesJardins, M. Lombardi, et al. An animated multivariate visualization for physiological and clinical data in the icu. In *Proc. 1st ACM International Health Informatics Symp.*, pp. 771–779, 2010.

[43] J. Rogers, N. Spina, A. Neese, et al. Composer—visual cohort analysis of patient outcomes. *Appl. Clin. Inform.*, 10(02):278–285, 2019.

[44] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of Information Visualization*, pp. 364–371. Elsevier, 2003.

[45] K. M. Stenson, B. E. Brockstein, and M. E. Ross. Epidemiology and risk factors for head and neck cancer. *UpToDate*, 2014.

[46] M. M. Thomas, T. Kannampallil, et al. Echo: A large display interactive visualization of icu data for effective care handoffs. In *IEEE Worksh. Vis. Analyt. Healthcare (VAHC)*, pp. 47–54. IEEE, 2017.

[47] J. Trelles, D. Lee, S. Derrible, et al. Visual analysis of a smart city's energy consumption. *Multimodal Technol. Interact.*, 3(2):30, 2019.

[48] T. D. Wang, C. Plaisant, B. Shneiderman, et al. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Trans. Vis. Comp. Graph.*, 15(6):1049–1056, 2009.

[49] A. Wentzel, P. Hanula, et al. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. *Radiother. & Onco.*, 148:245–251, 2020.

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

[51] R. Xiao, M. C. Ward, et al. Increased pathologic upstaging with rising time to treatment initiation for head and neck cancer: a mechanism for increased mortality. *Cancer*, 124(7):1400–1414, 2018.

[52] Z. Yuan, S. Finan, J. Warner, et al. Interactive exploration of longitudinal cancer patient histories extracted from clinical text. *JCO Clinical Cancer Informatics*, 4:412–420, 2020.

[53] T. Zhang, T. H. McCoy, R. H. Perlis, et al. Interactive cohort analysis and hypothesis discovery by exploring temporal patterns in population-level health records. In *IEEE Workshop on Vis. Analy. in Healthcare*, pp. 14–18. IEEE, 2021.

[54] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Info. Vis.*, 14(4):289–307, 2015.