Data Visualization Tools for Investigating Health Services Utilization Among Cancer Patients

DRAFT CHAPTER for ONCOLOGY INFORMATICS

Eberechukwu Onukwugha, MS, PhD*

Catherine Plaisant, PhD

Ben Shneiderman, PhD

*Corresponding Author:

Eberechukwu Onukwugha

e-mail: eonukwug@rx.umaryland.edu

telephone: 410-706-8981

fax: 410-706-5394

Abstract word count: 244

Body text word count: 8,509

Figures:10

References: 61

1 Abstract

The era of "big data" promises more information for health practitioners, patients, researchers, and policy makers. For big data resources to be more than larger haystacks in which to find precious needles, stakeholders will have to aim higher than increasing computing power and producing faster, nimbler machines. We will have to develop tools for visualizing information; generating insight; and creating actionable, on-demand knowledge for clinical decision making. This chapter has three objectives: 1) to review the data visualization tools that are currently available and their use in oncology; 2) to discuss implications for research, practice, and decision making in oncology; and 3) to illustrate the possibilities for generating insight and actionable evidence using targeted case studies. A few innovative applications of data visualization are available from the clinical and research settings. We highlight some of these applications and discuss the implications for evidence generation and clinical practice. In addition, we develop two case studies to illustrate the possibilities for generating insight from the strategic application of data visualization tools where the interoperability problem is solved. Using linked cancer registry and Medicare claims data available from the National Cancer Institute, we illustrate how data visualization tools unlock insights from temporal event sequences represented in large, population-based datasets. We show that the information gained from the application of visualization tools such as EventFlow can define questions, refine measures, and formulate testable hypotheses for the investigation of cancer-related clinical and process outcomes.

Keywords: Data Visualization, Temporal Analytic Techniques, Big Data, Human-System Integration, Human Computer Interaction

2 Contents

Data Visualization Tools for Investigating Health Services Utilization among Cancer Patients

"The greatest value of a picture is when it forces us to notice what we never expected to see."

—John Tukey, American Mathematician(1)

1 Introduction

1.1 Background

One of the promises of an informatics-infused health care system is the ability to extract meaning from large volumes of data for the purposes of improving the quality of care delivery and for generating new knowledge. Schilsky and Miller illustrate this case aptly in the first section of this book, as they described a vision for how to leverage informatics data from oncology practices into focused feedback for quality improvement. Likewise, Penberthy, Winn, and Scott presented a vision in their chapter for how electronic health record (EHR) data could be used to complement electronic pathology reports and other types of cancer registry data to offer a more complete view of cancer incidence and progression in the general population.

Unlocking the knowledge embedded within these massively distributed data streams in cancer, however, will require continual progress within the interdisciplinary scientific area of data visualization. Specialists in oncology informatics can benefit from advances in data visualization to make decision making more efficient, to improve systemic outcomes within hospitals and their communities, to engage patients more effectively in their own care, and to facilitate exploration of patterns and trends for hypothesis generation in research. Fortunately, advances in our understanding of how the human perceptual system works (see the chapter by Horowitz and Rensink), combined with advances in our understanding of how to construct more

efficient computer interfaces to support those processes, will put informaticists in good stead as we prepare for active participation in the "learning oncology system." (2).

1.2 Purpose of the Chapter

This chapter investigates the possibilities for generating insight and evidence from the strategic application of data visualization tools. While the era of "big data" promises more information for practitioners, patients, researchers, and policy makers, there is limited guidance for analysts about how to leverage the availability of such data. A few key questions must be addressed in order to turn the data into evidence: "How well do we extract insights from the information that is currently available?" "Are we prepared to gain insight directly from the information that is available in massive data sets?" "How well do we leverage longitudinal information that is available?" For big data resources to be more than larger haystacks in which to find precious needles, stakeholders will have to aim higher than increasing computing power and producing faster, nimbler machines. We will have to develop tools for visualizing information; generating insight; and creating actionable, on-demand knowledge for clinical decision making.

The White House press release (March 29, 2012) on the national Big Data Initiative identified two challenges, one of which we address directly in this chapter: 1) Develop algorithms for processing massive, but imperfect data and 2) Create effective human-computer interaction tools for visual reasoning. These well-crafted challenges position data visualization solidly on the national agenda. Three roles of data visualization address the White House challenges and clarify human participation:

1) Cleaning the often error-laden data. Consider the case in which statistical analyses of 6,300 emergency room admission records had failed to account for the eight patients who were entered into the EHR system as being 999 years old. Information specialists will

recognize this as a code for "age unknown," but the programs that calculated ages of admitted and discharged patients accepted this as a normal value, thereby distorting the results. A simple bar chart of the ages would have led any viewer to gasp with surprise. This example illustrates a proof of concept and there are an unlimited number of errors that may be missed by algorithms but spotted by experts, such as the patient who was admitted to the emergency room 14 times but discharged only twice. A quick glance at an appropriate visual display enables analysts to confirm the expected and detect the unexpected, especially errors.

2) Supporting exploration and discovery. Analysts typically begin with questions about their data, leading them to choose a particular visualization, such as line charts, size and color-coded scattergrams, maps, networks, and more sophisticated strategies. These analysts may immediately spot surprises or errors, but typically they split a data set to see men or women in separate displays, then group by age or race, and maybe focus on patients diagnosed at a later stage of cancer. Insights can lead to bold decisions regarding cancer-directed treatment receipt, treatment initiation, treatment continuation, and management of comorbid conditions.

3) Presenting results. In many cases, the results will be of interest to national leaders, health industry decision makers, and news media viewers. The more critical challenge with big data is to distill millions of health care data into a few cogent visualizations to guide proximal decision makers including clinicians, patients, and the patients' caregivers.

1.3 Human Systems Integration

The perspectives and work presented in this chapter are guided by a collaborative working relationship between the University of Maryland School of Pharmacy's Department of

Pharmaceutical Health Services Research and the Human Computer Interaction Laboratory (HCIL). Work in the HCIL, in turn, represents an interdisciplinary approach to system development based on contributions from the College of Computer, Mathematical, and Natural Sciences and the College of Information Studies at University of Maryland College Park. The purpose of this overall collaborative relationship is to bring a "human-systems integration" approach to the practice of informatics-supported medicine (3, 4). That is, the purpose is to design systems – in the case of this chapter, data visualization tools – that use computing power to augment and enhance the highly trained expertise of cancer epidemiologists, oncology care teams, health services researchers, and biomedical scientists. It builds on one of the core principles of the National Research Council's 2009 report on "Computational Technology for Effective Health Care," which was to design systems that provide improved cognitive support to care teams, administrators, patients, and their caregivers for the purpose of enhancing outcomes. Within the context of this chapter, the human system integration approach facilitates the development of tools designed for parsing data to generate insight and actionable knowledge, particularly when paired with well-articulated, clinically-motivated questions. As data availability grows, it becomes more important to develop methods and approaches for connecting humans with these data sources and systems.

The current health information technology systems are ill-suited to establish and maintain these connections and continuously inform patient and provider decision making. As noted by Dimitropoulos, health care systems should be data-driven, patient-centered, and continuously improving (5). For health care systems to effectively inform, influence, and interact with patients, it will require integrating systems that are not currently or widely blended such as hospital, outpatient, pharmacy, and dental systems. As research highlights the importance of holistic cancer care, the role of psychosocial cancer care, a link between

comorbidity (chronic disease) and cancer outcomes, as well as a link between dental health and chronic disease, we can no longer afford to deliver cancer care using fragmented care systems. Throughout this section, we emphasize the importance of leveraging big data and making full use of the longitudinal information available in these data to develop actionable evidence based on human interactions with these data based on review, analysis, synthesis, and discussion.

1.4 Chapter Objectives

This chapter section has three objectives: 1) to review the data visualization tools that are currently available and their use in oncology; 2) to discuss implications for research, practice, and decision making in the field of oncology; and 3) to illustrate the possibilities for generating insight and actionable evidence using targeted case studies. The case studies investigated here illustrate the possibilities for research and clinical decision making in situations where the interoperability problem is solved.

2 Methods and Data Visualization Tools

There are several different techniques and tools that may be applied to visualizing various types of data sets. This section reviews available techniques and discusses their strengths, weaknesses, and applications to cancer research and practice. The case studies in Section 4 focus on the use of two different prototypes of control panels (6) as visualization tools for generating insights from observational data sets including information about individuals diagnosed with cancer.

2.1 Techniques

One of the most common techniques for visualizing statistical data within the cancer epidemiological context is the use of Geographic Information Systems (GIS) to portray the distribution of a measured variable on top of an identifiable map, and one of the most common uses of GIS is to create choropleth maps. Brewer described choropleth mapping as a way to

visualize data relating to regional geographical locations and divisions such as state lines and zip codes (7). Choropleth maps provide policy makers and health officials with a situational awareness of the disease processes that may be at play within their jurisdictions. In some cases, a high incidence of certain types of cancer among a group of people within an identifiable geographical area may signal a public health emergency, known as a "cancer cluster," and might therefor require immediate environmental investigation. In other cases, a high incidence of late stage disease within certain areas may imply that vulnerable populations are falling outside of the reach of recommended public health primary and secondary prevention measures. Choropleth mapping can also give cancer control planners insight into how certain policies, such as cigarette taxes or indoor smoking prohibitions, may be associated with decreases in preventable disease, such as decreases in lung and bronchus cancers.

Symbols, colors, proportional symbols, icons, and textboxes are all commonly used elements within choropleth mapping. Basic cartographic symbols, such as solid lines depicting geopolitical boundaries or icons depicting identifiable landmarks, can provide a sense of consistency for analysts and an anchor for interpreting the underlying data patterns. Patterns or colors within the geographic units portray levels of the mapped data, which may represent an epidemiologic variable such as prevalence or mortality, or it may portray some type of demographic characteristic. Colors may be utilized to represent data and hierarchy based on hues or lightness (7), or they may also be used to stress extremes in data, drawing attention away from more average results colored in white (8). For example, darker hues generally suggest a higher frequency, percentage, or magnitude of the underlying variable in a choropleth map. Looking at the mortality maps presented at the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) website, readers can see how darker hues depict higher incidences of cancer-related mortality across cancer sites. When considering color choices,

however, investigators must keep the medium of their visualization in mind as the color's appearance and impact may vary between print, Internet, or presentations (8).

An alternative to using colors or patterns within a choropleth map is to use proportional symbols. Proportional symbols are usually geometric shapes, such as circles or triangles, which vary in size according to the magnitude of the underlying variable. Symbols of a larger size generally depict a greater underlying quantity than symbols of a smaller proportion. Proportional symbols may also be used outside of the context of GIS display to juxtapose magnitudes in more of a categorical sense. Figure 1 illustrates how the Centers for Disease Control and Prevention (CDC) used proportional circles to portray the relationship between some infectious agents and some cancers. Icons (with appropriately crafted legends) and text boxes are often needed to complete the reader's interpretation of the data and to facilitate general sense-making when working with interactive graphs. Because a user's gaze generally orients to the center of a graphic, it is often useful to place icons and textboxes in a central location (9).

<Insert Figure 1 About Here>

For some purposes, choropleth maps and other data visualization tools may need to portray values from more than one variable. To represent multiple variables, Brewer suggested using one or a combination of the following techniques: overlaying symbols, overlaying patterns, creating series, or combining variables (7). Colors, bands, and customizable icons are often used to represent categorical data, while numerical data is often shown with line plots, point plots, and bar charts. Other visualization techniques may also be applied to data to allow investigators to see trends more clearly such as pan and zoom; animation; filtering; brushing; and linking of different views of the same data, matrices (10), and rate smoothing (11).

Another technique that is commonly used in data visualization is data stream clustering; that is, using a clustering algorithm to reduce the dimensionality or noise associated with high frequency data streams. Chauhan, Kaur, and Alam explained how data clustering may be applied to identify and explore data patterns (12). Hierarchical algorithms are often employed, with either an agglomerative (i.e., bottom-up, starting with data points and building clusters) or divisive (i.e., top down, beginning with one overarching cluster and then dividing) approach, and clusters may either be density-based or grid-based. Density-based clustering techniques, such as Density Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS), and Clustering Based on Density Distribution Function (DENCLUE), form clusters from density distributions directly on databases. On the other hand, grid-based techniques, such as Sting, Wave Cluster, and Clique, cluster statistical data on a uniform grid. Data classing, or use of class breaks, is another common method used for developing choropleth or GIS maps (7). Data are typically grouped by quantiles, standard deviation, size, equal intervals, or natural breaks (8).

There are several other techniques that may be utilized for data visualization. Vellido et al. touched upon the use of directed graphs, which allow for the visualization of covariates and their relationships, and hierarchical visualizations, which provide detailed information about relationships between and for different hierarchical levels (13). Map projections are utilized to represent a given geographical area, taking into account the spherical curve of the earth (7). Neural networks, such as Self-Organizing Map (SOM), may be used for non-linear projects to "project high dimensional, time-varying information in 2D maps that correlate with diagnostic features" (13), while proximity networks form links between molecular information, pathways, and graphs. Community Health Map allows researchers to easily explore and visualize state and

county health patterns (14). GIS maps are important data visualization tools, as they allow participants' behaviors or characteristics to be linked with particular geographic factors (15).

2.2 Software Systems

Incorporating various combinations of the techniques above, many systems have been developed to visualize data. The Hierarchical Clustering Explorer (HCE) has a rank-by-feature framework that allows researchers to choose ranking criteria and visualize results in one-dimensional (histograms) or two-dimensional (scatterplots) projections (16). A software system called Caregiver is a tool that assists with therapy-related decisions through visualizations of general patient overview, patient cohorts, and individual patients (10). InfoZoom is a system that ensures displays of data sets will always fit on the selected screen in the form of compressed tables (10, 17). VisPap incorporates both medical images and laboratory data into its scatter plots and parallel coordinate plots, and the Cube uses EHRs to interactively identify and analyze of patterns with two-dimensional parallel planes in a three-dimensional cube display (10, 17).

Many systems have been developed to generate and/or base analyses on temporal abstractions. Moskovitch and Shahar described the KarmaLegoSification (KLS) framework that allows for the analyses of multivariate time series through temporal abstraction, time intervals mining, and pattern classifications (18). The Medical Information Visualization Assistant (MIVA) provides a visualization of the numerical value progression of point plots over a period of time (10). Interactive Parallel Bar Charts (IPBC) is an interactive system that simultaneously analyzes time-series and its associated values from multiple patients as well as sessions (10, 17). Lifelines is a system that illustrates historical data and events from EHRS and allows for aggregation of sets of events (10). Lifelines2 allows researchers to specify queries with event operators and align records by events (10, 19). The Similan system can also align records, but uses a similarity measure to take into account "addition, removal transposition of events and

temporal differences" (10). Building upon these point-event data projects, EventFlow allows researchers and clinicians to interactively visualize and analyze patterns of medication use from health systems' EHRs (19).

Another type of system is KNAVE II, which utilizes knowledge-based temporal abstraction and allows researchers to interactively visualize and explore temporal abstractions and patterns for single or small sets of electronic health data (10, 17). Building upon KNAVE II, Klimov et al. developed the VISualization of Time-Oriented recordS (VISITORS) system used for the visualization of multiple patient records. It is able to search and aggregate numerical and categorical data from both raw and abstracted data (17).

Several proposed systems are currently being developed. Goovaerts described a geostatistical simulation that uses Poisson kriging, p-field simulation, and local clustering to generate risk maps that are more realistic than those formed using solely smoothing methods (20). West, Borland, and Hammond developed two additional prototypes to explore and analyze large data sets through visualization: 1) the radial-coordinates visualization tool incorporates many techniques including colors, lines spreading, parallel and radial coordinates, histograms, and scatter plots, to allow investigators to visualize clusters, data distributions and individual data sets; and 2) the force-directed network visualization tool uses proportional symbols, links, and nodes to explore queries made from particular elements of EHRs (21).

Harford, Edwards, Nandakumar, Ndom, Capocaccia, and Coleman described a "cancer atlas" system derived from India's Internet-based registry that brings many of the techniques discussed at the beginning of this section into use within a global cancer control context (22). The system, referred to as *GDB Compare*, allows international users to make comparisons between countries on the global disease burden. As illustrated in Figure 2, the system allows international users to select filtering options on the left side of the screen for two coordinated

visualizations. The treemap at the top of the screen represents a type of hierarchical clustering technique while the map at the bottom represents a choropleth mapping technique based on countries as geographical units. The treemap clusters based on all causes of death, with the size of the rectangle corresponding to the number of deaths and the color hue signifying change over time. Darker hues signify a worsening changing for the cause while lighter hues suggest an improving condition. The overarching color palette for the treemap breaks causes into chronic conditions (blue), infectious disease (brown), and injury (green). The color spectrum used for the choropleth map at the bottom of the screen ranges from dark blue to dark red, with blue signifying low numbers of death and red signifying high numbers.

<Insert Figure 2 About Here >

2.3 Strengths and Weaknesses of Available Tools

2.3.1 Strengths of Available Tools

The methods described above have varying advantages. Similar to how different hues allow observers to easily distinguish variance, map series allow users to recognize patterns easier through contrasts between maps. Overlay allows observers to readily identify unreliable data in a particular region, while map projections provide appropriate views of the geographic distributions of diseases (7). The rate smoothing technique Kafadar described not only mitigates problems that come with utilizing multiple sources, but it also allows investigators to temporarily overlook certain patient characteristics to better visualize patterns (11). When considering the advantages of density-based clustering, DBSCAN does not require significant information to identify inputs, and OPTICS is able to automatically determine the necessary number of clusters from data sets. On the other hand, grid-based clustering allows for fast processing time (12).

Furthermore, temporal abstractions overcome challenges such as varied frequencies, gaps in data, and working with both time points and intervals (18). EventFlow allows for easy-to-use interactive visualization, event overlapping, and pattern identification (19). InfoZoom enables researchers to identify hidden knowledge, and the Lifelines system provides ease of use and access, along with the ability to zoom in and out (10). The VISITORS system has several strengths. First, it captures data for multiple patients and allows for time and value analysis of clinical data. VISITORS also allows researchers to quickly and accurately answer clinical questions utilizing these temporal abstractions and clinical information (17).

One of the strengths of the radial visualization system prototype designed by West et al. is its ability to utilize numerous techniques to clearly organize data and clusters without muddling the visualization (21). It is also able to display many different data distributions through multiple axes.

2.3.2 Weaknesses of available tools

It is also important to recognize the weaknesses that visualization methods may have. Utilizing colors becomes a disadvantage when considering individuals with color blindness or color distinguishing deficiencies. Brewer also explained that maps, in general, have several weaknesses such as misleading titles, technological difficulties with sharing maps, and skewed judgement of densities on map projections (7). Bhowmick, Griffin, MacEachren, Kluhsman, and Lengerich found that cancer researchers often face limited data, difficulty in merging data, time-consuming steps, and overly complex software when employing GIS or other spatial analysis software (23).

Additionally, West et al. explained that different patterns and interpretations may be concluded from alternate views of the same forced-directed network visualization (21). Likewise, James et al. pointed out that analyses and spatial outcomes may vary greatly,

depending on the different techniques investigators may choose to adopt (8). For example, there are several approaches to establishing cut points; applying Jenks algorithm (24) would cluster data based on their natural breaks, while standard deviations would result in clusters that may be more sparsely dispersed. Clustering techniques have several other disadvantages. It may not take into account the uncertainty associated with predicted risk (20), and different approaches to organizing clusters may result in varied interpretations (13). Density-based clustering systems are flawed as well. For example, DBSCAN may not be entirely sensitive to all inputs, making it difficult to recognize clusters that are closely related. On the other hand, grid-based clustering is not ideal for irregularly distributed data, as it may not be able to fully capture the cluster quality or time (12).

There are drawbacks to other systems as well. The Caregiver system does not follow patients' development over time (10). Users need a degree of statistical knowledge to easily and successfully use the HCE system (16). Vellido et al. discerned that a disadvantage with the Growing Hierarchical Self-Organizing Map (GHSOM) is that investigators would not be able to visualize information from each hierarchal level at the same time (13). Other techniques, such as directed graphs and proximity networks, have not been well developed. Similarly, several visualization tools are still just developing prototypes (21) or theoretical systems (13, 23).

Of the remaining systems previously discussed, Klimov et al. noted that KNAVE II is not an ideal system for large data sets (17). Because the TimeFinder system is based on time-oriented data, it is not able to focus on a specific set of subjects. Contrarily, Spotfire, SimVis, and Lifelines lack the ability to incorporate or produce high level abstractions such as those focused on time (25). Lifelines2 and Similan are based on point events rather than time intervals; do not distinguish between data from tests, diagnoses, or treatments; and are not able to display individual record details (10).

3 Applications of data visualization in the cancer setting

3.1 Basic Cancer Science

Visualization has long been used to complement algorithmic analysis in the basic sciences underlying cancer research. This has been especially true in areas such as genomics in which the amount of raw data to explore for hypothesis generation is simply too large and cumbersome to portray through individual vectors of raw values. The expansive genomic data space lends itself to an exploration of relationships through data visualization. Figure 3, for example, depicts a visualization of whole-genome rearrangements using the *Circos* software package for visualizing data relationships in a circular layout. Circos was developed to give scientists the ability to explore relationships between objects, such as chromosomes and other genomic elements, their size, and orientation in relationship to each other (26). In Figure 3, the outer ring of the circular graph depicts chromosomes arranged in sequential order from end to end, while the inner ring displays copy-number data in green and interchromosomal translocations in purple for two different tumors. The Circos data visualization package can produce charts with high "data to ink" ratios (27), making the format a highly efficient mechanism to explore relationships in a big data context.

<Insert Figure 3 About Here>

3.2 Population Statistics

Aside from using advanced techniques for research purposes, another compelling reason to create data visualization tools is to make the complex incidence and prevalence statistics associated the national surveillance of cancer trends accessible to journalists, policy makers, and the public (9). For example, the American Cancer Society (ACS) collaborates with the CDC and NCI to publish an annual compilation of "Cancer Facts and Figures" (28) as a report card on the nation's collective progress against cancer. The report breaks out data from the cancer registries

and other surveillance mechanisms to enumerate trends over time, to explore prevalence and mortality as broken out by sociodemographic groupings, and to make distinctions in progress between variants of the disease. These visualizations have employed some of the standard variants of charts and graphs already familiar to most audiences – such as the elements associated with line charts, bar graphs, and pie charts – but more recently have employed new innovations such as the graphical depiction of quantities and numerical trends.

A more recent innovation in communicating to the public, made feasible by diffusion of dynamic HTML/web technologies, is the use of publicly facing informatics tools to present interactive data displays for local customization and exploration. Figure 4 presents an image of the U.S. Cancer Statistics Interactive Atlas website hosted by the CDC. This data visualization tool allows analysts to interact with the control box on the left to filter data based on cancer event (e.g., incident rate, death rate); cancer site (e.g., lung and bronchus, colon and rectum); gender; race/ethnicity; year; and classifying statistic (e.g., quintiles). Results are portrayed on a choropleth map at the top center of the screen. A choropleth map uses shading or patterning to fill in geographic areas on a map (e.g., states or counties) according to levels of an analytic variable. In this case, the absence of coloring within states indicates an absence of reportable data. Lighter shading indicates a lower value on the outcome variable, while darker shades indicate higher values. Clicking on a state will indicate the ranking of its values within the context of all states' values portrayed graphically within the box at the bottom of the page. The precise numeric values with accompanying confidence intervals are listed in a table on the right, while a player bar in the upper right allows the user to explore trends over time.

< Insert Figure 4 about Here >

More generally, GIS systems are used in the cancer setting to examine data quality (23) and to investigate the association of cancer with socioeconomic, genetic, or environmental

factors (7), as they may play a role in the development of cancer. For example, Finney-Rutten et al. (29) explored the use of isopleth maps to investigate the distribution of cultural norms and behaviors related to smoking cessation using nationally available data from NCI's Health Information National Trends Survey (HINTS) (30). Unlike choropleth maps, which display data by filling in geographic units such states or counties with the same shade of color or patterning, isopleth maps portray gradual patterns of change across predefined borders. Weather maps and topographic maps are good examples of isopleth mapping techniques. The isopleth maps for the behaviorally oriented HINTS data illustrated for cancer control planners how beliefs and their concomitant actions can cluster in geographic communities. These maps illustrated how beliefs in the scientific linkage between smoking and cancer were weakest along the Appalachian ridge, which when juxtaposed against the SEER choropleth maps for cancer incidence and mortality corresponded to high cancer mortality rates from lung and bronchus cancer.

Similarly, Chauhan et al. describe the use of DBSCAN, OPTICS, and DENCLUE to visualize cancer clusters using data from two large databases: GLOBOCAN from the International Agency for Research on Cancer and SEER from NCI (12). SimVis interactively classifies and clusters data from clinical trials and examinations, and visualizations such as caMATCH have been used to identify potential clinical trial patients (14, 15). Spurred by examples such as these, the White House initiated a government-wide effort to make health data from all of the national surveillance programs available to data scientists for the development of usable, transparent interfaces for community planning. On July 10, 2014, the U.S. Department of Health and Human Services included open access to large scale, health-related databases as an integral part of its Open Government Plan. Examples of open-access data sets, and the data visualization tools being created to access them, can be found at HealthData.gov.

To understand how these new data visualization tools are being utilized in the cancer space, Bhowmick et al. interviewed cancer researchers to identify what aspects of spatial analysis they often employ or consider most useful and suggest features useful for cancer data visualization (23). The authors observed that cancer control researchers proceed methodically through three phases: 1) a pre-analytic phase in exploring and repairing attributes of a given data set; 2) a conceptually exploratory stage, in which scientists explore the nature of preliminary associations; and 3) an analytic phase, in which population estimates are generated, spurious associations are appropriately controlled statistically, and specific conclusions are drawn. What is produced in the analysis phase is then readied for publication. From their interviews, the authors noted that tables and maps are used both in the early exploratory phases of cancer research as well as in the later publication process.

3.3 Clinical Applications

As EHR systems become more powerful and greater attention is given to optimizing the use of data for predictive, preemptive, personalized, and participative care (31), then the use of data visualizations within the EHR interface will become more important for allowing analysts to quickly assimilate large amounts of data for clinical purposes. Figure 5 shows a sample screen of a urology EHR system, summarizing the record of a patient with prostate cancer, and using a design similar to early research on Lifelines (32). In this example, the attending clinical team is given the ability to view the rise and fall of Prostate Specific Antigen (PSA) levels before and after treatment. The approach typifies an area of human system integration research aimed at using informatics tools to create better visualizations of temporal patterns to track the course of treatment over time (33, 34), and to reduce discontinuities in care from missed prescriptions (19) or laboratory results (35). Visualization techniques can also be used at the individual patient level to improve the effectiveness and efficiency of medication reconciliation tasks (36).

< Insert Figure 5 about Here >

When large collections of cancer patient records are available, looking for temporal patterns of treatment, side effects or outcomes become possible, and visualization can reveal possible linkage to population attributes such as age or gender (Figure 6). Systems such as EventFlow (see case study of Section 4) or VISITORS (17) may be used to quickly answer clinical questions. The program VisCareTrails has been to analyze cancer case studies using EHR data. The Cube extracts data from EHR, and VisPap utilizes medical images and laboratory data to interactively visualize patterns (10). Simpao and colleagues demonstrated how a visual analytics dashboard in a pediatric hospital's EHR system can be used to optimize drug-drug interaction alerts. (37)

< Insert Figure 6 about Here >

Looking at systems that are currently in place to extract cancer data from EHR and pathology reports, Forman et al. discussed E-path, caBIG's Cancer Text Information Extraction System (caTIES), and MediClass(15). Information from these databases, in conjunction with data visualization tools described in Section 1 may then be used to explore and analyze cancer trends.

Several approaches may not have been applied in the cancer setting yet, but have been effectively used to visualize data in other similarly complex situations. Augmented Interactive Starfield Display uses point plots to display blood glucose readings, while the web-based interactive visualization system (WBIVS) uses data from home monitoring systems to display lung transplant patients' data. Used in intensive care settings, Midgaard "integrates the display of numerical data with graphical representations of medical treatment plans" (10).

Moving forward, visualization systems and programs continue to be developed and incorporated into the cancer setting.

4 Case Studies

Case studies provide an ideal framework for illustrating the insights that are possible with these tools. Via targeted case studies, we investigate the utility of two tools, EventFlow and Cohort Comparison (CoCo) that are ideal for investigating longitudinal event sequences. The case studies illustrate the purposeful integration of data visualization and observational data to address questions that are relevant for clinical practice. Using linked cancer registry and health care claims data, we investigate the timing of treatment initiation and health services utilization following the diagnosis of late-stage cancer.

4.1 Introduction to EventFlow and CoCo

EventFlow (Figure 7) allows analysts to understand the temporal features and prevalence of the patterns found in a cohort of patients. Figure 7 illustrates dummy data representing 29 men diagnosed with cancer. We use a small sample for clarity of presentation. On the right the timeline shows details of individual records. Triangles represent events. The records have been aligned by the cancer diagnosis date (green event). Users would need to scroll to see all 29 records. In the center, the overview aggregates groups of records with the same sequence of events into horizontal (gray) block stripes that include colored vertical bars representing each event. Within each horizontal block stripe, the height of the vertical bar is determined by the number of patients in the group and the horizontal gap between events is proportional to the average time between events. Reading from the left we can see that all records start with a cancer diagnosis. We can then see the different sequences of treatment with luteinizing hormone-releasing hormone (LHRH) (purple) and radiation therapy (brown). The most common first treatment is the LHRH. The second most common is radiation therapy and we can see that it occurs earlier on average than LHRH as the distance from green to brown is shorter than the distance from green to purple.

<Insert Figure 7 about here>

The two views (overview and timeline) are coordinated so that when users select records in one view they are highlighted in the other view. The timeline shows the sequencing and timing of therapy for individual patients. EventFlow also includes two separate search interfaces including an advanced graphical user interface that makes it possible for analysts to specify complex temporal queries including temporal constraints and the absence of events (38) (e.g., men who did not receive LHRH within 6 months of diagnosis), or search and replace (25). The combination of those techniques (39) allows analysts to sharpen the focus of an analysis on records exhibiting particular event sequences of interest, e.g. considering skeletal complications, analysts could investigate the occurrence of pathological fracture followed by bone surgery then palliative radiation to the bone (RtB).

The second tool, CoCo, (see Figure 8) facilitates the identification of salient differences between the temporal patterns found in two separate cohorts of men diagnosed with prostate cancer and identified from the SEER registry data linked with Medicare claims data. In Figure 8, we compare a cohort of 474 stage IV M0 prostate cancer records to a cohort of 2,470 stage IV M1b (bone metastatic) prostate cancer records in the 3 months following diagnosis. In pilot work, we proposed an initial taxonomy of metrics (such as differences in the prevalence of events, sequences or subsequences of consecutive events, co-occurrence of events, duration of gaps between events, event attributes, etc.—to be refined during the study) (40). For each metric, CoCo computes a series of statistical tests and presents the results using an interactive user interface. This is a novel approach that combines both statistical methods and a visual representation of the results and encourages rapid hypothesis generation. Users are provided with a set of metrics they can choose from (bottom left), and then review the results of the

visualization in the bottom right. Based on our early user tests we find that analysts can usefully incorporate insights from CoCo and EventFlow (41).

<Insert Figure 8 about Here>

4.2 Application #1: Algorithm Development Using Claims Data

This case study illustrates an approach that combines the billing information found in claims data with key longitudinal information regarding the timing of health services utilization to isolate probable radiation to the bone. Clinical providers and researchers need reliable measures in order to identify treatment receipt and its consequences. Claims data are used to identify treatment and associated consequences in large populations. However, claims-based algorithms used to infer conditions and treatments are error-prone, unless validated, and better algorithms are needed. Research on the development of claims-based algorithms relies on the ability to unlock the rich but incomplete data found in the temporal sequences of events that are available in claims data. However, research on algorithm development has been limited by the lack of a clearly-defined approach for unlocking the rich data in temporal patterns and sequences that are available in claims data. These patterns can be first order (e.g., interval between events) or second order (e.g., patterns of intervals over time for *each* patient and *across* patients) in nature. The first-order events are easily summarized and analyzed using standard statistical methods while the second-order events, as we found out when using standard statistical analysis software, cannot be summarized using standard statistical tools. Another challenge is selecting from competing alternatives to identify the temporal components that are most useful in the algorithms.

Studies using claims data have documented increased mortality and costs associated with bone metastasis (BM) and BM-related complications (42-46). However, their utility is limited by the fact that the claims algorithms are not validated, are differential (47) (e.g., misclassification

of metastasis using claims data varies with patient characteristics that also associate with survival), inaccurate (48-50), and can lead to biased conclusions regarding survival (47). Cooper et al. (51) examined the use of Medicare claims data for identifying the stage of prostate cancer, and found that billing codes had a 78.2% and 72.8% positive predictive value (PPV) for regional and distant prostate cancer, respectively when compared to medical records. Hassett et al. (48) studied billing codes as indicators for recurrence of prostate cancer after definitive local therapy and reported a maximum PPV of 19%. Results are not unique to prostate cancer. Chawla et al. (47) reported that claims data had a PPV of 65.8% for identifying a diagnosis of distant breast cancer compared to SEER registry data.

Previous studies have identified patients with BM based on the presence of a diagnosis of "secondary malignant neoplasm of bone and bone marrow" (International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) 198.5) in claims data. These claims-based algorithms differ in their use of the ICD-9-CM codes. Several studies have defined BM patients as persons with two or more claims encounters including 198.5 anytime on or after the date of the first claim with a diagnosis of cancer (45, 46, 52). Other studies have defined BM patients as persons with at least one inpatient claim with the 198.5 code, at least one outpatient claim with the 198.5 code paired with a code for procedures used to diagnose/treat BM, or at least one outpatient physician evaluation and management claim with the 198.5 code (43, 44). Our published results (53) indicate that the approach to measuring BM can impact validity.

Reliable identification of BM is critical for identification of the appropriate clinical subpopulation to study BM complications including RtB. Billing codes available in claims data do not provide information regarding the anatomic site that was treated with radiation therapy. In the absence of these codes, researchers use the BM ICD-9 diagnosis code to identify a BM diagnosis based on claims and then define RtB based on radiation claims occurring after the BM

claim (43-45). The validity of this approach depends on the validity of using BM ICD-9 diagnosis codes to identify a BM diagnosis, a practice which is likely to be unreliable given prior results (48-50) regarding the low sensitivity and PPV of claims-based algorithms to identifying metastasis. In our work, we have found that the duration of radiation therapy can be useful for distinguishing between RtB and radiation to other sites for cancer treatment. As part of the Choosing Wisely campaign, the American Society for Radiation Oncology discouraged routine use of extended fractionation schemes (>10 fractions) for palliation of BM (54), since single fractionation schemes are more convenient for patients and provide comparable pain relief for uncomplicated BM, further indicating the potential utility of duration of radiation therapy for identifying RtB. We investigate the length of therapy and the presence of BM coding on the radiation claim using Eventflow and CoCo.

In Figure 8, the selected metric is the most differentiating event and we could use this information to identify components of an algorithm for identifying radiation to the bone separately from radiation to the prostate gland. For example, we see that "Bmv2" (blue rectangle) representing a BM diagnosis code on the health care claim, is found in 40.9% of the M0 records, and 88.2% of the M1b (bone metastatic) records, with a difference of 47%. We also see a difference in the next two most differentiating events: Death and "Rad_b_a3.". The former variable, *Death*, represents all-cause mortality while the latter variable, *Rad_b_a3*, represents health care claims for short-course (i.e., less than 4 weeks) radiation therapy. We expect that all-cause death and short-course radiation therapy (likely RtB) will be more common in the incident M1b compared to the M0 group within 3 months following diagnosis of prostate cancer. Via this case study using EventFlow output, we illustrate that the presence of a BM code on the radiation claim and the length of radiation therapy may be important for identifying radiation to the bone (separately from radiation to the prostate gland) using health care claims data.

4.3 Application #2: Patient Comorbidity and Health Services Utilization

Among men diagnosed with prostate cancer, it is commonly stated that they are more likely to die from underlying comorbid conditions (e.g., heart failure) than they are to die from the prostate cancer. Much of the research on comorbidity has been conducted among men diagnosed with low or intermediate risk disease (55-58). Compared to men diagnosed with non-metastatic cancer, men diagnosed with metastatic prostate cancer are more likely to die from the prostate cancer. It is important that these men receive cancer-directed therapy as soon as possible following diagnosis of late-stage disease. In this application, we investigate whether patient comorbidity status impacts the timing of receipt of cancer-directed therapy and use of other health services including hospital, skilled nursing facility, and hospice services. We focus on a particularly vulnerable group of cases: men diagnosed with incident bone metastatic disease as identified by the American Joint Committee on Cancer (AJCC) staging information available from the SEER registry. Categories that represented too few patients (i.e., N<11) were suppressed in the EventFlow graphic, per the requirements of the SEER-Medicare Data Use Agreement. Specifically, we suppressed the category of CCI=0 in Figure 9 and suppressed the indicator for a hospice admission in Figure 10.

The time to receipt of cancer-directed treatment is plotted in Figure 9 for four groups of men, defined based on their Charlson Comorbidity Index (CCI) score at the time of diagnosis with bone metastatic disease. The data represented in Figure 9 are based on a stratified random sample of 200 men diagnosed with stage IV M1b (incident bone metastasis) prostate cancer between 2005 and 2009 and with at least 1 year of follow-up information following prostate cancer diagnosis. Fifty men were randomly selected from each of the CCI subgroups. We grouped patients based on information in the Medicare claims data from the 12 months prior to the diagnosis of incident M1b prostate cancer. Patients were categorized into groups: missing, 0,

1, and >=2. Of note, the CCI score was categorized as "missing" when no claims were observed during the time 12 months prior to cancer diagnosis.

<Insert Figure 9 about here>

Given that the patients in this data set were diagnosed with incident M1b prostate cancer, the group with missing CCI score is of particular interest because there were no observed claims for receipt of health services for 12 months prior to the diagnosis of M1b disease and for use in calculating the CCI score. In our prior work (59), we found that these patients are less likely to visit an urologist for a follow-up visit following diagnosis. We found that patients with CCI score coded as "missing" should be studied as a separate group as opposed to combining the "missing" group with the group with CCI score = 0. Our results here are consistent in that we find that the proportion of men who receive treatment is lowest (60%) among the group with CCI coded as "missing," suggesting that the absence of engagement with the health care system prior to diagnosis persists following diagnosis, despite the diagnosis of severe disease (i.e., M1b prostate cancer). By comparison, the proportion of men who received treatment was 68% among men with CCI score greater than or equal 2, 76% in the men with CCI score = 1 and greater than 76% in the men with CCI score = 0.

The EventFlow graphic in Figure 9 plots time to receipt of any of the following: orchiectomy, radical prostatectomy, radiation therapy, LHRH agonist, anti-androgen, chemotherapy, and radiopharmaceutical. Figure 9 provides information that is immediately useful for understanding treatment receipt in this sample of men diagnosed with incident M1b prostate cancer. Reading along the y-axis, the height of the light gray panels is informative (or, alternatively, the height of the negative space) for providing information on the proportion of men (not) receiving treatment in a given subgroup. Within each stratum, the height of the *light gray* panel provides information on the proportion receiving cancer-directed treatment, which

can then be compared across strata. The height of the white (i.e., negative) space provides information on the proportion who did not receive cancer-directed treatment within the timeframe of the follow-up. From Figure 9, we can see that the height of the gray panel is largest among those with CCI score = 0, i.e., the healthiest subgroup. The height of the gray panel is smallest among those with CCI = "missing" (those with no health claims for calculating the CCI score during the 12 months prior to their cancer diagnosis). EventFlow also provides information regarding the timing of treatment receipt, via the length of the light gray panel. Comparing across the CCI strata, we see that the time to treatment initiation is shortest among those with CCI score = 0 and longest among those with CCI score = "missing".

The ordering of the light gray and white space is also informative since EventFlow plots the most common events first. Thus, within a given stratum, if treatment receipt is more common than no treatment, the light gray panel will be ordered first (reading top to bottom along the y axis), followed by the white space. If 'no treatment' is more common, the negative space will appear first. In Figure 9, we immediately see that the light gray panel (representing treatment receipt) occurs first among the individuals with few or no comorbidities (CCI=1 or CCI=0), indicating that the probability of treatment receipt is higher among the healthier subgroups.

Together, the results from Figure 9 regarding the probability of treatment receipt and timing of treatment initiation suggest that:

1. The group of men with CCI=2+ are the most vulnerable group among those with non-missing CCI scores. Compared to individuals with CCI=0, they are less likely to receive treatment and more likely to receive it in a delayed fashion. The results indicate that comorbidity impacts disease management among those with late-stage prostate cancer, in this case, in terms of their likelihood of receiving critical cancer-directed therapies.

2. The group of men with CCI=missing is also a vulnerable group, and *more vulnerable than the group with the highest comorbidity burden*. They are least likely to receive treatment and, when they do receive treatment, exhibit the longest delay in initiating treatment.

Figure 10 provides information regarding time from prostate cancer diagnosis to first hospitalization, skilled nursing facility (SNF) stay or hospice, stratified by pre-diagnosis CCI score (0, 1, >=2, or missing). The figure is based on 200 men diagnosed with stage IV M1b (incident bone metastasis) prostate cancer between 2005 and 2009 and with at least 1 year of follow-up information following prostate cancer diagnosis. Fifty men were selected from each of the CCI subgroups. The events of interest included time to: all-cause hospitalization, SNF admission, and hospice admission. The absence of a shaded light gray area (i.e., negative space) indicates that none of the events of interest were observed during the 1 year follow-up period post-diagnosis of incident bone metastatic prostate cancer.

<Insert Figure 10 about here>

The figure reflecting the proportion and timing of hospitalizations and SNF admissions (Figure 10) indicates that:

1. Hospitalizations (green) are more common than SNF admissions (blue) in the year post-diagnosis in this stratified random sample of 200 men.
2. SNF admissions are most likely in the group with CCI score=2+.
3. Hospitalizations are more common in the group with CCI=missing and CCI=2+. The group of patients with a hospitalization is ordered first, followed by the group of patients with no hospitalization events.

4. Hospitalizations are less common in the group with CCI=0 and CCI=1. The group of patients with no hospitalization events appears first, followed by the group with hospitalization events.

Note that we have not examined characteristics of the hospitalizations. These characteristics (urgent vs. routine admission, length of stay, disease severity index at admission, clinical diagnosis at admission) can be incorporated in Eventflow as attributes in order to provide additional information regarding the hospitalization. As illustrated in these targeted case studies, visualization provides an efficient and intuitive approach to conduct exploratory data analysis of timing and sequencing of events. When supported by a population-based sample of men, these insights from EventFlow can be used to develop formal testable hypotheses (e.g., a higher comorbidity index score is associated with a lower probability of treatment receipt) and determine what variables to include (e.g., an indicator for treatment receipt, time to treatment receipt). The information provided regarding event sequences for patient groups can assist with refining measures, answering questions, and formulating hypotheses for the investigation of cancer-related clinical outcomes.

5 Conclusion

The easier production of high quality static graphics, animated weather maps, video presentations and interactive websites has lowered the barriers to entry into the data visualization product market. However, we are just at the early stages of broadening visual literacy and training a new generation of researchers and decision makers. If data visualization tools that integrate powerful statistical techniques are made commonly available, the benefits could be as potent as the use of graphical user interfaces.

Glossary

Algorithm. A step-by-step procedure for solving a problem or accomplishing some end especially by a computer. Source: http://www.merriam-webster.com/dictionary/algorithm

Choropleth Map. A thematic map that uses graded differences in shading or color or the placing of symbols inside pre-defined, aggregated units (or areas) on a map in order to indicate differences in the average values of some measure in those areas. Sources: http://www.ncgia.ucsb.edu/cctp/units/unit47/html/mas_form.html and http://www.thefreedictionary.com/choropleth+map

Distant disease. The tumor has spread beyond the original site, traveled to other parts of the body and begun to grow in the new location(s). Source: http://training.seer.cancer.gov/ss2k/staging/categories/distant.html

Localized disease. The tumor has extended beyond the original site but has not spread to other organs and begun to grow in the new location(s). Source: http://training.seer.cancer.gov/ss2k/staging/categories/regional.html

Prostate-specific antigen. A protease that is secreted by the epithelial cells of the prostate gland. Source: http://www.merriam-webster.com/dictionary/prostate-specific%20antigen

Treemap. Treemap, coined by Dr. Ben Shneiderman, is the name for a space-constrained visualization of hierarchical structures that splits the screen into rectangles in alternating horizontal and vertical directions as you traverse the screen from top to bottom. Source: http://www.cs.umd.edu/hcil/treemap-history/

6 Acronyms and Abbreviations

ACS American Cancer Society

ACA Affordable Care Act

AJCC American Joint Committee on Cancer

AMIA The American Medical Informatics Association

BM Bone metastasis

CCI Charlson Comorbidity Index

CDC Centers for Disease Control and Prevention

CMS Centers for Medicare and Medicaid Services

EHR Electronic Health Record

EMR Electronic Medical Record

FDA Food and Drug Administration

GIS Geographic Information Systems

HCIL Human Computer Interaction Laboratory

HINTS Health Information National Trends Survey

HIT Health Information Technology

HITECH Health Information Technology for Economic and Clinical Health Act

HTML HyperText Markup Language

IOM Institute of Medicine

ICD-9-CM International Classification of Diseases, Ninth Revision, Clinical Modification

LHRH Luteinizing hormone-releasing hormone

NCD Noncommunicable diseases

NCI National Cancer Institute

NIH National Institutes of Health

NRC National Research Council

ONC The Office of the National Coordinator for Health Information Technology

PCAST President's Council of Advisors on Science and Technology

PPV Positive predictive value

PSA Prostate-specific antigen

RtB Radiation to the bone

SNF Skilled Nursing Facility

SHARP Strategic Health IT Advanced Research Projects

SEER Surveillance Epidemiology and End Results system

WHO World Health Organization

References

1. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley; 1977.

2. American Society of Clinical Oncology. Accelerating progress against cancer: ASCO's blueprint for transforming clinical and translational cancer research. Alexandria, VA: American Society of Clinical Oncology2011.

3. Hesse BW, Shneiderman B. eHealth research from the user's perspective. American journal of preventive medicine. 2007 May;32(5 Suppl):S97-103.

4. Shneiderman B, Plaisant C. Designing the user interface : strategies for effective human-computer interaction. 5th ed. Boston: Addison-Wesley; 2010.

5. Dimitropoulos L. Health IT research priorities to support the health care delivery system of the future. (Prepared for the Agency for Healthcare Research and Quality under Contract No290200900023-I. 2014;AHRQ Publication No. 14-0072-EF.

6. Heer JS, B. Interactive dynamics for visual analytics. Communications of the ACM. 2012;55(4):45-54.

7. Brewer CA. Basic mapping principles for visualizing cancer data using Geographic Information Systems (GIS). American journal of preventive medicine. 2006 Feb;30(2 Suppl):S25-36.

8. James WL, Cossman RE, Cossman JS, Campbell C, Blanchard T. A brief visual primer for the mapping of mortality trend data. International journal of health geographics. 2004 Apr 8;3(1):7.

9. Nelson DE, Hesse BW, Croyle RT. Making Data Talk: Communicating health data to the public, policy, and the press. New York, NY: Oxford; 2009.

10. Rind A, Wang TD, Aigner W, Miksch S, Wongsuphasawat K, Plaisant C, et al. Interactive information visualization to explore and query electronic health records. Foundations and Trends in Human-Computer Interaction. 2013;5(3):207-98.

11. Kafadar K. Geographic trends in prostate cancer mortality: an application of spatial smoothers and the need for adjustment. Annals of epidemiology. [Research Support, U.S. Gov't, Non-P.H.S.]. 1997 Jan;7(1):35-45.

12. Chauhan R, Kaur, H., & Alam, M. A. Data clustering method for discovering clusters in spatial cancer databases. International Journal of Computer Applications. 2010;(0975–8887) Volume.

13. Vellido A, Martin, J.D., Rossi, F., Lisboa, P.J. Seeing is believing: the importance of visualization in real-world machine learning applications. 19th European Symposium on Artificial Neural Networks, Computational Intelligence, and Machine Learning. 2011.

14. Shneiderman B, Plaisant, C., & Hesse, B.W. Improving health and healthcare with interactive visualization methods. IEEE Computer- Special issue on challenges in information visualization. 2013;46(5):58-66.

15. Forman MR, Greene SM, Avis NE, Taplin SH, Courtney P, Schad PA, et al. Bioinformatics: Tools to accelerate population science and disease control research. American journal of preventive medicine. 2010 Jun;38(6):646-51.

16. Seo J, Shneiderman B. Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework. IEEE transactions on visualization and computer graphics. 2006 May-Jun;12(3):311-22.

17. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artificial intelligence in medicine. 2010 May;49(1):11-31.

18. Moskovitch R, Shahar, Y. Classification of multivariate time series via temporal

abstraction and time intervals mining. Knowl Inf Syst. 2014;Accepted: 4 September 2014.

19. Monroe M, Meyer, T.E., Plaisant, C., Lan, R., Wongusphasawat, K., Coster T.S., Gold, S., Millstein, J., Shneiderman, B. Visualizing patterns of drug prescriptions with EventFlow: A pilot study of asthma medications in the Military Health System. HCIL Tech Report. 2013.

20. Goovaerts P. Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. International journal of health geographics. 2006;5:7.

21. West V, Borland, D., Hammond, W.E. Visualization of EHR and health related data for information discovery. Workshop on Visual Analytics in Healthcare. 2013.

22. Harford JB, Edwards BK, Nandakumar A, Ndom P, Capocaccia R, Coleman MP, et al. Cancer control-planning and monitoring population-based systems. Tumori. 2009 Sep-Oct;95(5):568-78.

23. Bhowmick T, Griffin, A. L., MacEachren, A. M., Kluhsman, B. C., & Lengerich, E. J. . Understanding the process of cancer data exploration and analysis. Health & Place. 2008;14(3):576-607.

24. Jenks GF. The data model concept in statistical mapping: George Philip; 1967.

25. Monroe M, Lan R, Morales J, Shneiderman B, Plaisant C, Millstein J. The challenges of specifying intervals and absences in temporal queries: a graphical language approach. Proc. ACM CHI 2013, ACM, New York (April 2013), 2349-2358.

26. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009 Sep;19(9):1639-45.

27. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire, Conn.: Graphics Press; 2001.

28. American Cancer Society. Cancer facts & figures 2015. Atlanta, GA: American Cancer Society2015.

29. Finney Rutten LJ, Augustson EM, Moser RP, Beckjord EB, Hesse BW. Smoking knowledge and behavior in the United States: Sociodemographic, smoking status, and geographic patterns. Nicotine Tob Res. 2008 Oct;10(10):1559-70.

30. Nelson DE, Kreps GL, Hesse BW, Croyle RT, Willis G, Arora NK, et al. The health information national trends survey (HINTS): development, design, and dissemination. J Health Commun. 2004 Sep-Oct;9(5):443-60; discussion 81-4.

31. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW. Collaborative biomedicine in the age of big data: the case of cancer. Journal of medical Internet research. 2014;16(4):e101.

32. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: Using visualization to enhance navigation and analysis of patient records. American Medical Informatics Association 1998 Annual Fall Symposium 1998.

33. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. Proceedings / AMIA Annual Symposium AMIA Symposium. 1998:76-80.

34. Plaisant C, Lam S, Shneiderman B, Smith MS, Roseman D, Marchand G, et al. Searching electronic health records for temporal patterns in patient histories: a case study with microsoft amalga. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:601-5.

35. Tarkan S, Plaisant C, Shneiderman B, Hettinger AZ. Reducing missed laboratory results: defining temporal responsibility, generating user interfaces for test process tracking, and

retrospective analyses to identify problems. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:1382-91.

36. Plaisant C, Wu J, Hettinger AZ, Powsner S, Shneiderman B. Novel user interface design for medication reconciliation: an evaluation of Twinlist. Journal of the American Medical Informatics Association : JAMIA. 2015 Mar;22(2):340-9.

37. Simpao AF, Ahumada LM, Desai BR, Bonafide CP, Galvez JA, Rehman MA, et al. Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. J Am Med Inform Assoc. Mar;22(2):361-9.

38. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal event sequence simplification. IEEE transactions on visualization and computer graphics. 2013 Dec;19(12):2227-36.

39. Shneiderman B, Plaisant C. Sharpening analytic focus to cope with big data volume and variety: Ten strategies for data focusing with temporal event sequences, Visualization Viewpoint, IEEE Computer Graphics and Applications 35, 3 (May/June 2015), 10-14.

40. Malik S, Du F, Monroe M, Onukwugha E, Plaisant C, Shneiderman B. Comparing cohorts of event sequences with balanced integration of analytics and statistics, Proc. ACM 20th International Conference on Intelligent User Interfaces, ACM Press, New York (2015), 38-49.

41. Malik S, Du F, Monroe M, Onukwugha E, Plaisant C, Shneiderman B. An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences. EHRVis Workshop on Visualizing Electronic Health Record Data 2014; Paris, France.

42. Norgaard M, Jensen AO, Jacobsen JB, Cetin K, Fryzek JP, Sorensen HT. Skeletal related events, bone metastasis and survival of prostate cancer: a population based cohort study in Denmark (1999 to 2007). J Urol. 2010 Jul;184(1):162-7.

43. Sathiakumar N, Delzell E, Morrisey MA, Falkson C, Yong M, Chia V, et al. Mortality following bone metastasis and skeletal-related events among men with prostate cancer: a population-based analysis of US Medicare beneficiaries, 1999-2006. Prostate Cancer Prostatic Dis. 2011 Jun;14(2):177-83.

44. Sathiakumar N, Delzell E, Morrisey MA, Falkson C, Yong M, Chia V, et al. Mortality following bone metastasis and skeletal-related events among women with breast cancer: a population-based analysis of U.S. Medicare beneficiaries, 1999-2006. Breast Cancer Res Treat. 2012 Jan;131(1):231-8.

45. Lage MJ, Barber BL, Harrison DJ, Jun S. The cost of treating skeletal-related events in patients with prostate cancer. Am J Manag Care. 2008 May;14(5):317-22.

46. Delea T, McKiernan J, Brandman J, Edelsberg J, Sung J, Raut M, et al. Retrospective study of the effect of skeletal complications on total medical care costs in patients with bone metastases of breast cancer seen in typical clinical practice. J Support Oncol. 2006 Jul-Aug;4(7):341-7.

47. Chawla N, Yabroff KR, Mariotto A, McNeel TS, Schrag D, Warren JL. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. Annals of epidemiology. 2014 Sep;24(9):666-72, 72 e1-2.

48. Hassett MJ, Ritzwoller DP, Taback N, Carroll N, Cronin AM, Ting GV, et al. Validating Billing/Encounter Codes as Indicators of Lung, Colorectal, Breast, and Prostate Cancer Recurrence Using 2 Large Contemporary Cohorts. Med Care. 2012 Dec 6.

49. Nordstrom BL, Whyte JL, Stolar M, Mercaldi C, Kallich JD. Identification of metastatic cancer in claims data. Pharmacoepidemiol Drug Saf. 2012 May;21 Suppl 2:21-8.

50. Thomas SK, Brooks SE, Mullins CD, Baquet CR, Merchant S. Use of ICD-9 coding as a proxy for stage of disease in lung cancer. Pharmacoepidemiol Drug Saf. 2002 Dec;11(8):709-13.

51. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. Med Care. 1999 May;37(5):436-44.

52. Delea TE, McKiernan J, Brandman J, Edelsberg J, Sung J, Raut M, et al. Impact of skeletal complications on total medical care costs among patients with bone metastases of lung cancer. J Thorac Oncol. 2006 Jul;1(6):571-6.

53. Onukwugha E, Yong C, Hussain A, Seal B, Mullins CD. Concordance between administrative claims and registry data for identifying metastasis to the bone: an exploratory analysis in prostate cancer. BMC Med Res Methodol. 2014;14:1.

54. Hahn C, Kavanagh B, Bhatnagar A, Jacobson G, Lutz S, Patton C, et al. Choosing Wisely: The American Society for Radiation Oncology's Top 5 list. Pract Radiat Oncol. 2014;4(6):349-55.

55. Albertsen PC, Moore DF, Shih W, Lin Y, Li H, Lu-Yao GL. Impact of comorbidity on survival among men with localized prostate cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2011 Apr 1;29(10):1335-41.

56. Stattin P. Mortality in older men with low-risk prostate cancer and high comorbidity. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. [Letter]. 2015 Mar 20;33(9):1086-7.

57. Daskivich TJ, Chamie K, Kwan L, Labo J, Dash A, Greenfield S, et al. Comorbidity and competing risks for mortality in men with prostate cancer. Cancer. 2011 Oct 15;117(20):4642-50.

58. Daskivich TJ, Fan KH, Koyama T, Albertsen PC, Goodman M, Hamilton AS, et al. Effect of age, tumor risk, and comorbidity on competing risks for survival in a U.S. population-based cohort of men with prostate cancer. Ann Intern Med. 2013 May 21;158(10):709-17.

59. Onukwugha E, Osteen P, Jayasekera J, Mullins CD, Mair CA, Hussain A. Racial disparities in urologist visits among elderly men with prostate cancer: a cohort analysis of patient-related and county of residence-related factors. Cancer. 2014 Nov 1;120(21):3385-92.

60. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.

61. Bernard J, Sessler, D., May, T., Schlomm, T., Pehrke, D., & Kohlhammer, J. A Visual-interactive System for Prostate Cancer Stratifications. VIS 2014 Workshop on Visualization of Electronic Health Records; November 9, 2014; Paris2014.

Figure legends:

Figure 1. Using visualization to inform the public (http://canceratlas.cancer.org/risk-factors/), this particular graphic utilizes proportional symbols to illustrate the relative proportion of cancer cases accounted for by infections.

Figure 2. GBD Compare, based on the Global Burden of Disease. At the top, a treemap shows all the causes of deaths. The size of the box is proportional to the number of deaths, and the color indicates the change over time (light for improving, dark for worsening). Neoplasms are selected, and the map below shows where the disease is most prevalent. http://viz.healthmetricsandevaluation.org/gbd-compare/

Figure 3. Visualization of whole-genome rearrangement. Two different tumors are being compared using Circos plots (26) of whole-genome sequence data, showing gene duplications and chromosome rearrangements. The outer ring depicts chromosomes arranged end to end. The inner ring displays copy-number data in green and inter-chromosomal translocations in purple. Source:(60) .

Figure 4. U.S. Cancer Statistics Interactive Atlas of the CDC. http://nccd.cdc.gov/DCPC_INCA/ .

Figure 5. Visualization of a patient electronic health record for clinical urology care from IntrinsiQ.

Figure 6. A visualization of prostate cancer patient records. At the center, the overview of three main stages of the disease are color coded green, yellow, and red. On the side, the distributions of static patient attributes are shown allowing for the selection of subsets of the population and providing insight into differences between groups(61).

Figure 7. Illustration of temporal patterns in health care claims data using EventFlow.

Figure 8. A screenshot of an early prototype of CoCo, comparing two prostate cancer cohorts: AJCC stage M0 and AJCC stage M1b.

Figure 9. Time from prostate cancer diagnosis to first treatment in the year following cancer diagnosis, stratified by pre-diagnosis Charlson Comorbidity Index score (1, >=2, or missing) (CCI = zero was suppressed due to a small sample size, per the Data Use Agreement).

Figure 10. Time from prostate cancer diagnosis to first hospitalization (green) or skilled nursing facility stay (blue), stratified by pre-diagnosis Charlson Comorbidity Index score (0, 1, >=2, or missing) (The indicator for a hospice admission was suppressed due to the small sample size, per the Data Use Agreement).

Figure 1. Using visualization to inform the public (http://canceratlas.cancer.org/risk-factors/) this particular graphic utilizes proportional symbols to illustrate the relative proportion of cancer cases accounted for by infections.
Permission info: Figure grabbed on August 12, 2015 from http://canceratlas.cancer.org/risk-factors/. Webpage provides contact information as an email: canceratlas@cancer.org

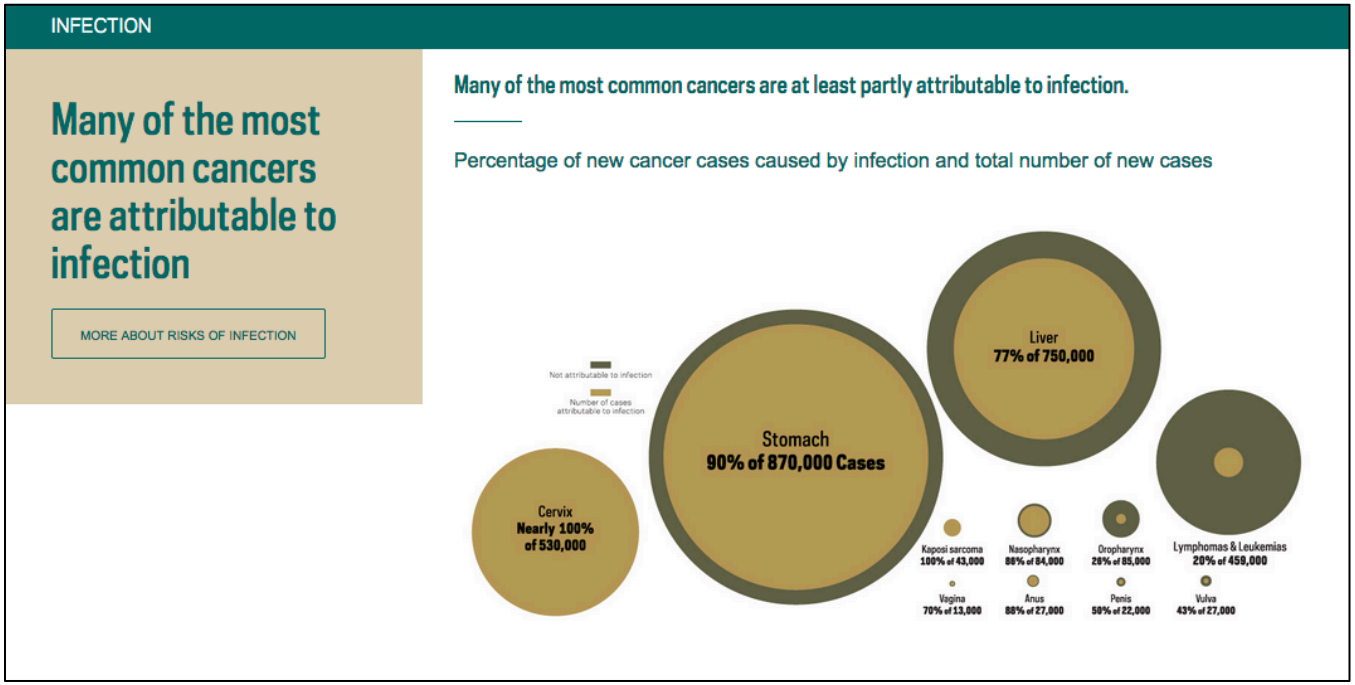

Figure 2. GBD Compare, based on the Global Burden of Disease. At the top a treemap shows all the causes of deaths. The size of the box is proportional to the number of deaths, and the color indicates the change over time (light for improving, dark for worsening). Neoplasms are selected, and the map below shows where the disease is most prevalent. http://viz.healthmetricsandevaluation.org/gbd-compare/

Permission info: Figure grabbed on August 12, 2015 from http://viz.healthmetricsandevaluation.org/gbd-compare/ with neoplasm selected. Webpage provides contact information as an email: http://viz.healthmetricsandevaluation.org/gbd-compare/. May also contact our personal contact: Rhonda Stewart stewartr@uw.edu (University of Washington's Institute for Health Metrics and Evaluation)

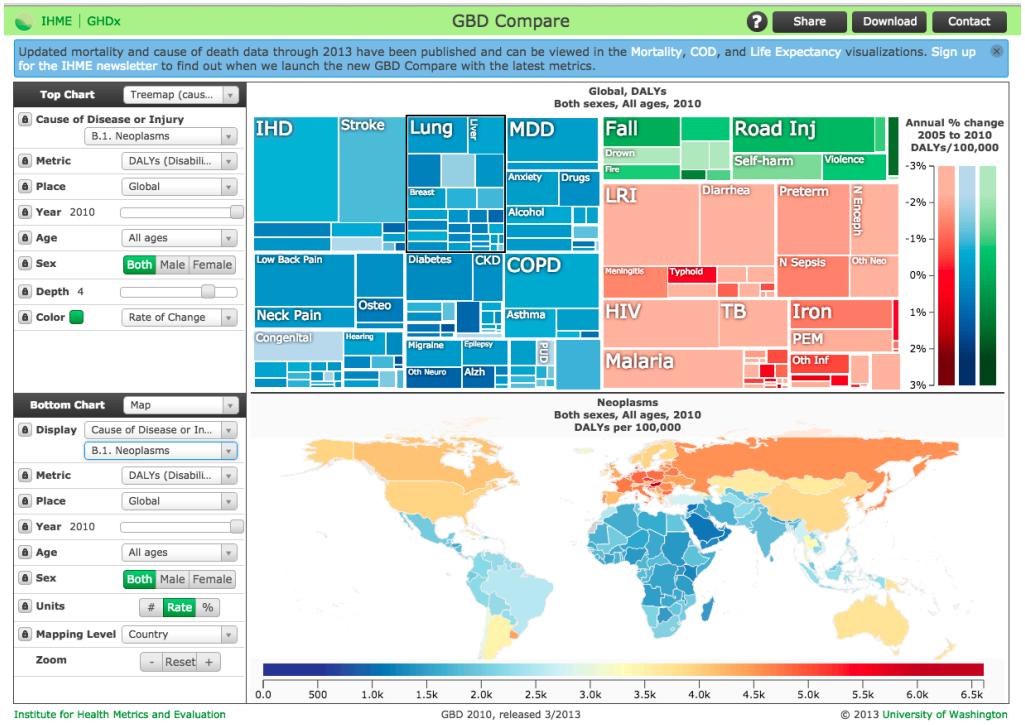

Figure 3. Visualization of whole-genome rearrangement. Two different tumors are being compared using Circos plots (Krzywinski, 2009) of whole-genome sequence data, showing gene duplications and chromosome rearrangements. The outer ring depicts chromosomes arranged end to end. The inner ring displays copy-number data in green and interchromosomal translocations in purple. From: Imielinski, M. 2012

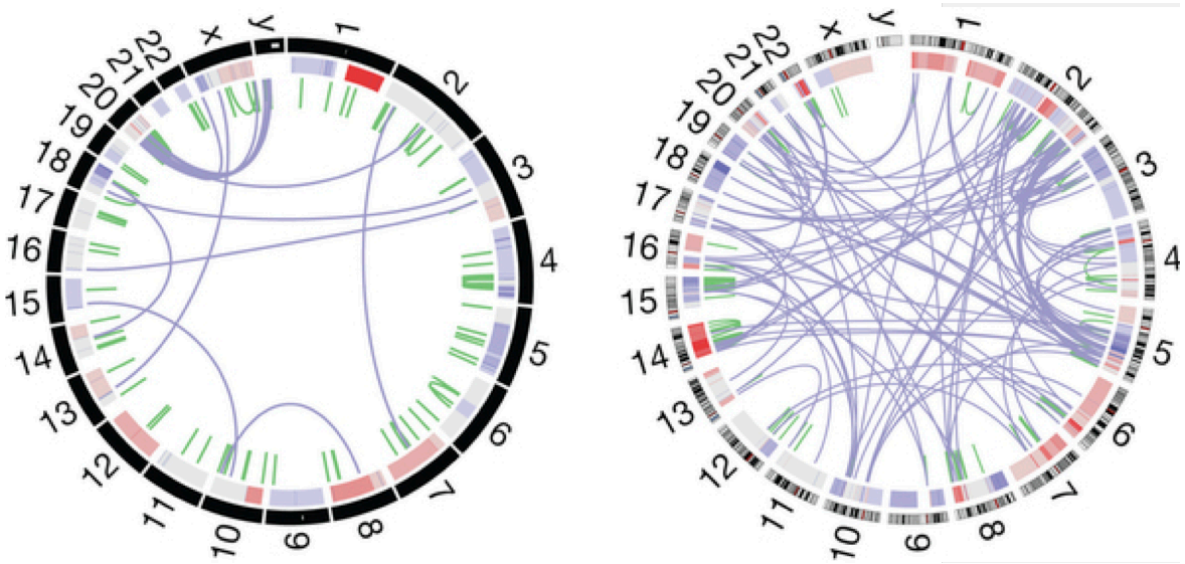

Figure 4. U.S. Cancer Statistics Interactive Atlas of the CDC. http://nccd.cdc.gov/DCPC_INCA/

Permission info: the figure was grabbed on August 12, 2015 from the URL http://nccd.cdc.gov/DCPC_INCA/
This is a government website i.e. CDC
Contact info provided in the site is: Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30329 USA
800-CDC-INFO | (800-232-4636) | TTY: (888) 232-6348, and a form is provided to ask questions at: Contact CDC-INFO

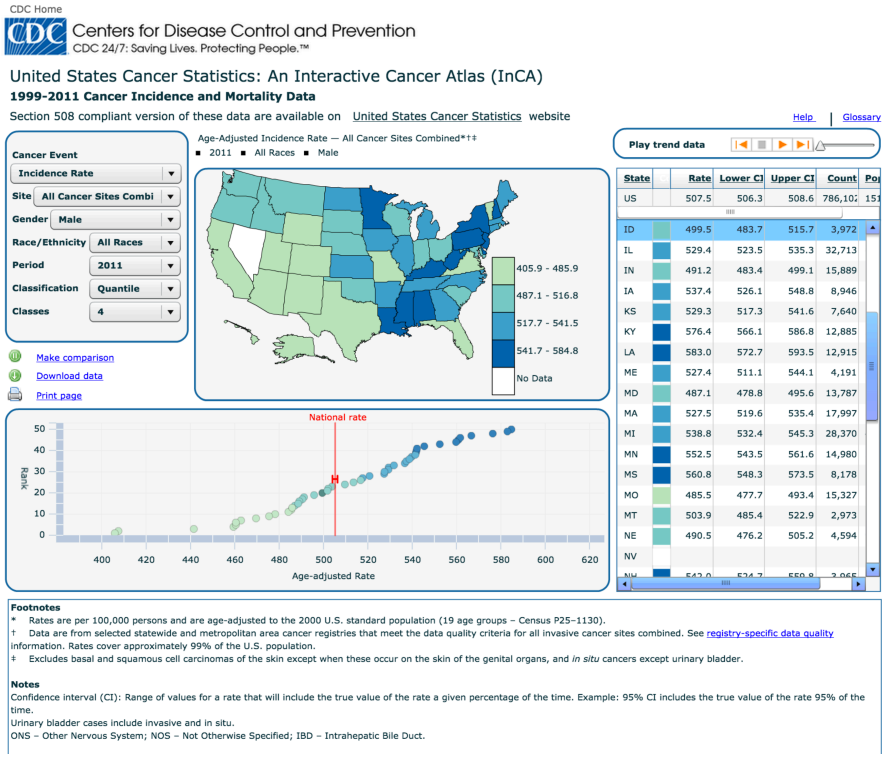

Figure 5. Visualization of a patient electronic health record for clinical urology care from IntrinsiQ

Permission information: This image was found on the HealthTronics (UroChartEHR) website in 2014 but the company seems to have changed. I can see the same figure at http://www.intrinsiq.com/IntrinsiQSoftware/UroChart (+ click on "key advantages").

Contact info for IntrinsiQ is 877-570-8721 or emailinfo@intrinsiq.com

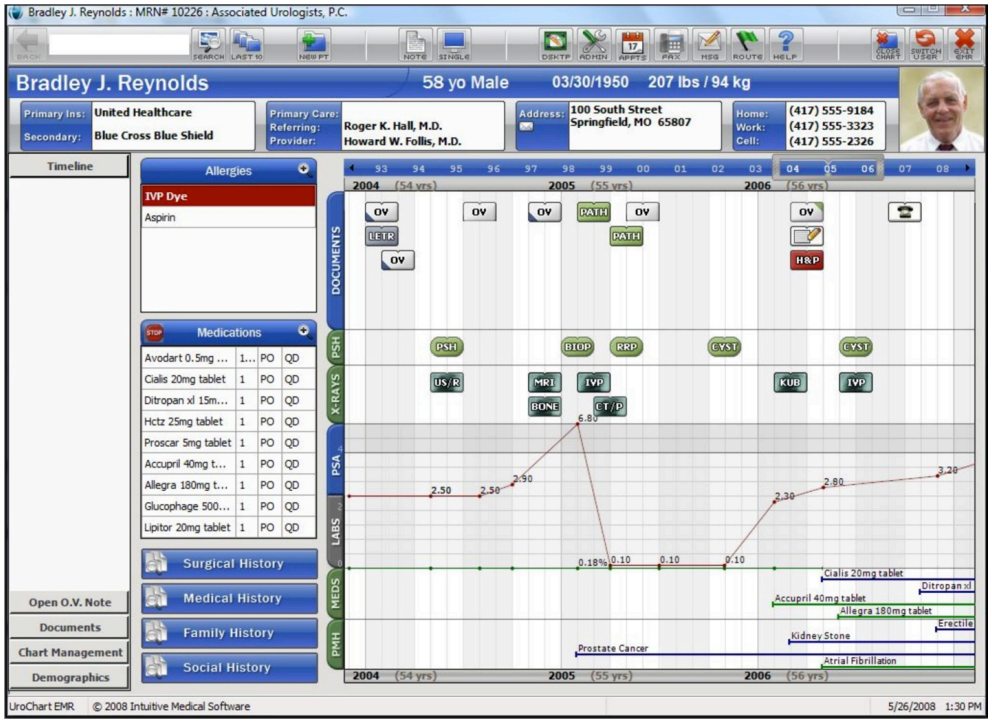

Figure 6. A visualization of prostate cancer patient records. At the center, the overview of three main stages of the disease are color coded green, yellow, and red. On the side the distributions of static patient attributes are shown allowing for the selection of subsets of the population and providing insight into differences between groups (Bernard, 2014)

Permission information: This image was grabbed from the following webpage: http://www.cs.umd.edu/hcil/parisehrvis/papers/prostate_cancer.pdf
This is a workshop paper and the © remains with the authors.
The 1st author of the paper and personal contact is Jurgen Bernard juergen.bernard@igd.fraunhofer.de

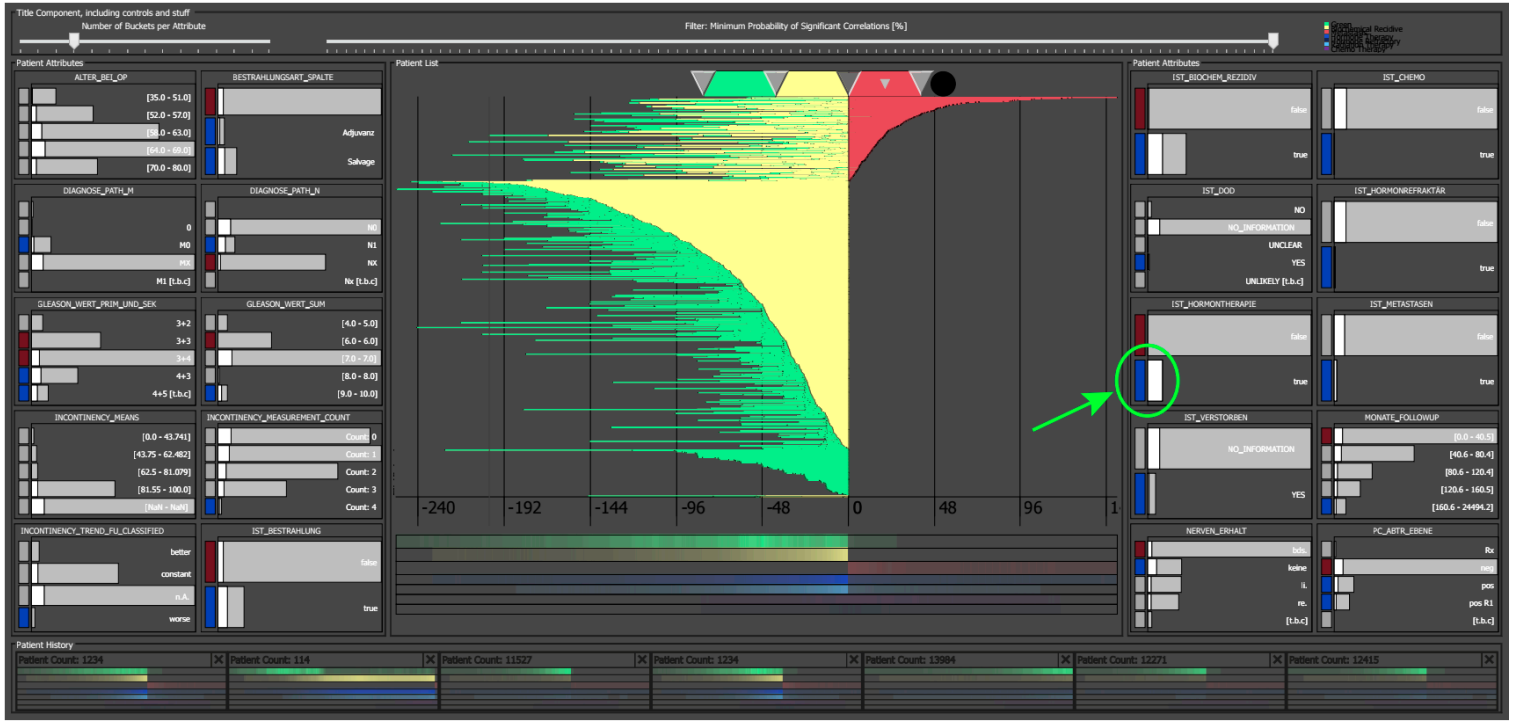

Figure 7. Illustration of temporal patterns in health care claims data using EventFlow.
Permission info N/A: work of the authors

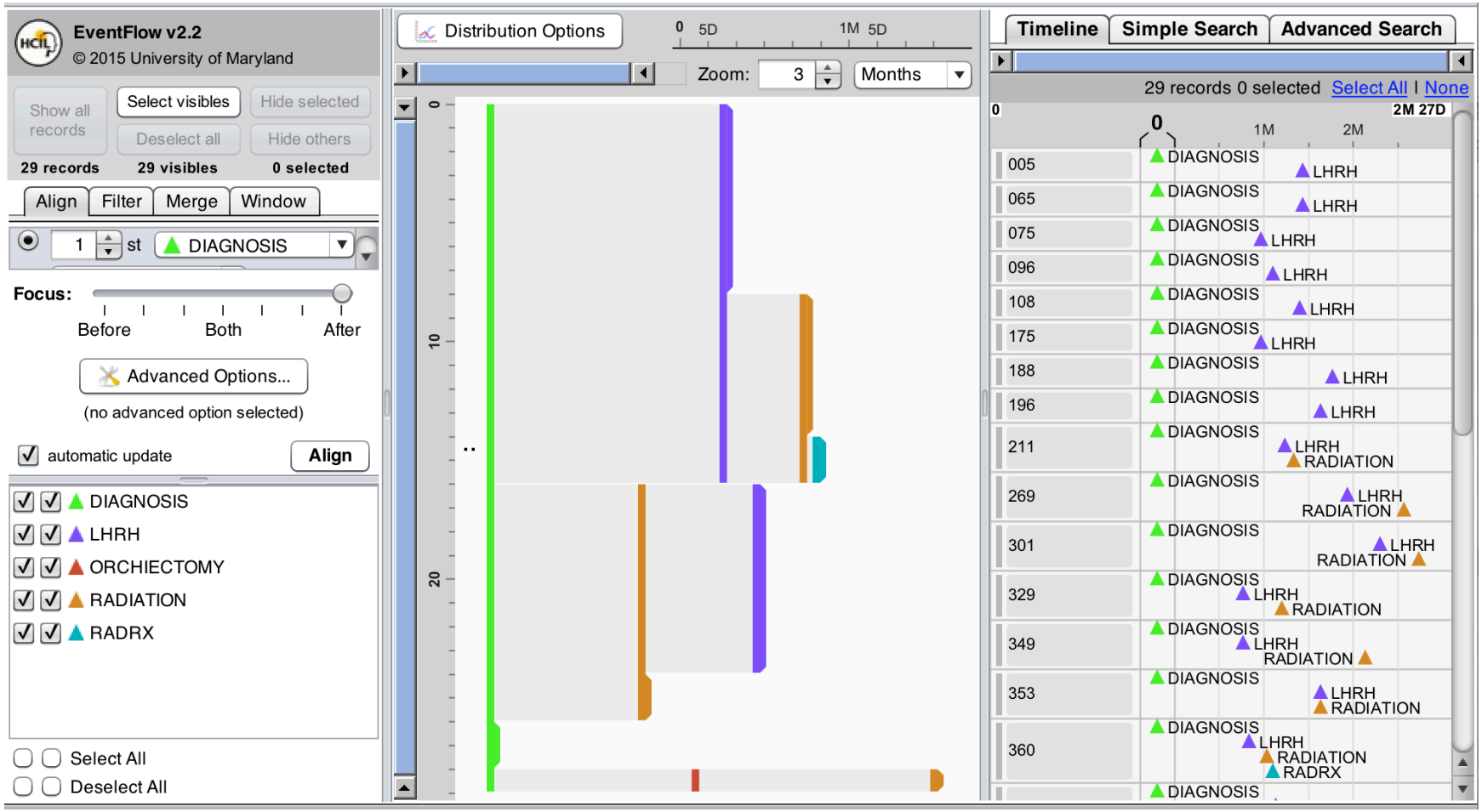

Figure 8. A screenshot of an early prototype of CoCo, comparing two prostate cancer cohorts: AJCC stage M0 and AJCC stage M1b.

Permission info N/A: work of the authors

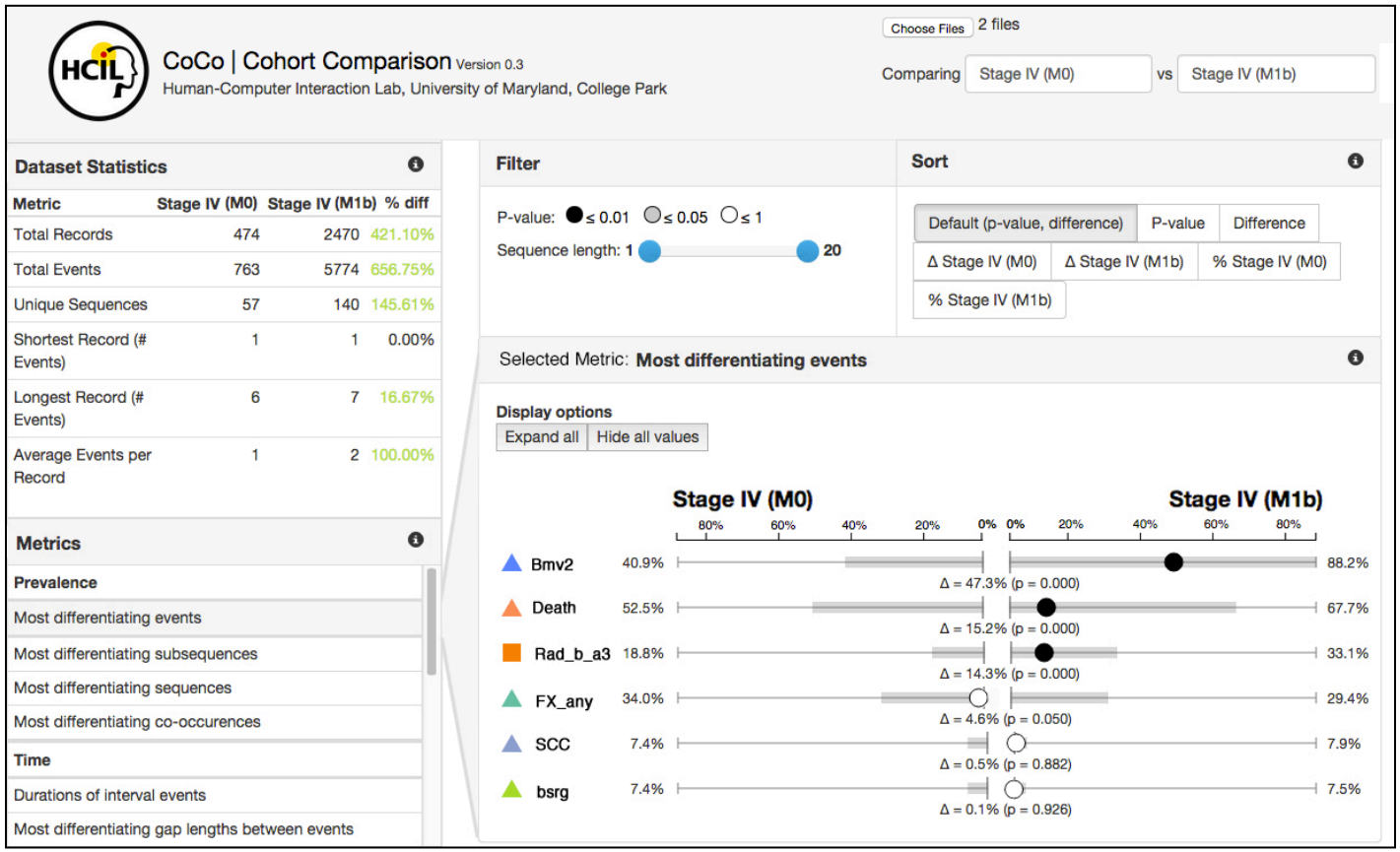

Figure 9. Time from prostate cancer diagnosis to first treatment in the year following cancer diagnosis, stratified by pre-diagnosis Charlson Comorbidity Index score (1, >=2, or missing) (CCI = zero was suppressed due to a small sample size, per the Data Use Agreement).

Permission info N/A: work of the authors

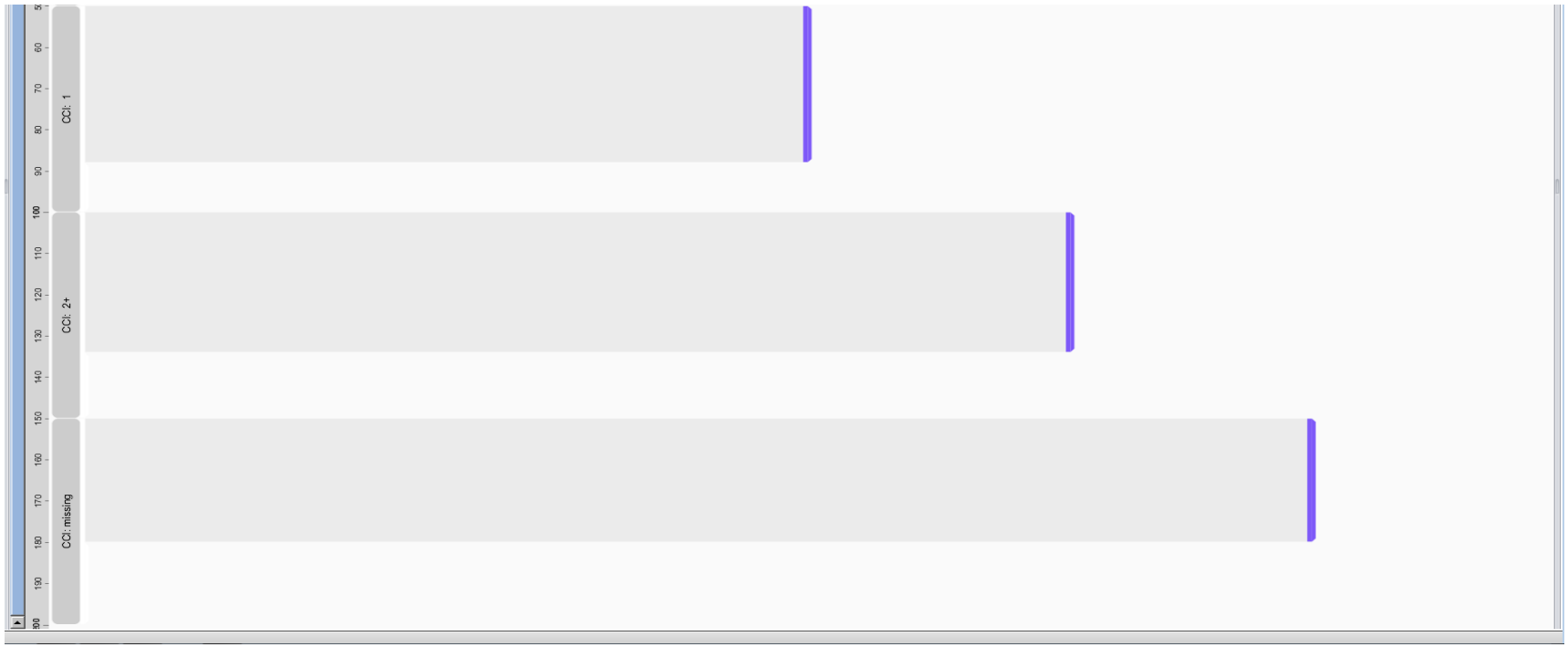

Figure 10. Time from prostate cancer diagnosis to first hospitalization (green) or skilled nursing facility stay (blue), stratified by pre-diagnosis Charlson Comorbidity Index score (0, 1, >=2, or missing) (The indicator for a hospice admission was suppressed due to the small sample size, per the Data Use Agreement).

Permission info N/A: work of the authors