Adaptive Contextualization Methods for Combating Selection Bias during High-Dimensional Visualization

DAVID GOTZ and SHUN SUN, University of North Carolina at Chapel Hill
NAN CAO, Tong Ji University
RITA KUNDU and ANNE-MARIE MEYER, University of North Carolina at Chapel Hill

Large and high-dimensional real-world datasets are being gathered across a wide range of application disciplines to enable data-driven decision making. Interactive data visualization can play a critical role in allowing domain experts to select and analyze data from these large collections. However, there is a critical mismatch between the very large number of dimensions in complex real-world datasets and the much smaller number of dimensions that can be concurrently visualized using modern techniques. This gap in dimensionality can result in high levels of selection bias that go unnoticed by users. The bias can in turn threaten the very validity of any subsequent insights. This article describes Adaptive Contextualization (AC), a novel approach to interactive visual data selection that is specifically designed to combat the invisible introduction of selection bias. The AC approach (1) monitors and models a user's visual data selection activity, (2) computes metrics over that model to quantify the amount of selection bias after each step, (3) visualizes the metric results, and (4) provides interactive tools that help users assess and avoid bias-related problems. This article expands on an earlier article presented at ACM IUI 2016 [16] by providing a more detailed review of the AC methodology and additional evaluation results.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Visual analytics**; *Empirical studies in visualization*; • **Applied computing** → *Health informatics*;

Additional Key Words and Phrases: Visualization, visual analytics, exploratory analysis, intelligent visual interfaces, selection bias

ACM Reference format:
David Gotz, Shun Sun, Nan Cao, Rita Kundu, and Anne-Marie Meyer. 2017. Adaptive Contextualization Methods for Combating Selection Bias during High-Dimensional Visualization. *ACM Trans. Interact. Intell. Syst.* 7, 4, Article 17 (November 2017), 23 pages.
https://doi.org/10.1145/3009973

1 INTRODUCTION

Large and complex datasets are being gathered across a wide range of application disciplines to support data-driven decision making. From healthcare to advertising to business intelligence,

This work was supported in part by a Data Fellow award from the National Consortium for Data Science (NCDS). This material is based upon work supported by the National Science Foundation under Grant No. 1704018.
Authors' addresses: D. Gotz, S. Sun, and R. Kundu, School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive, Chapel Hill, NC 27599; N. Cao, College of Design and Innovation, Tong Ji University, 1239 Siping Road, Shanghai, P.R. China; A.-M. Meyer, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, 450 West Drive, Chapel Hill, NC 27514. A.-M. Meyer (Current Address), IQVIA; email: AnneMarie.Meyer@iqvia.com.

https://doi.org/10.1145/3009973

these datasets are often gathered "in the wild" with large numbers of heterogeneous and diverse variables.

Unlike data for traditional *prospective* studies—which is gathered narrowly according to a study design determined a priori with the goal of supporting a specific analytical question—these real-world datasets capture a vast and diverse sample of data from the system under investigation without knowing in advance the types of questions that will be asked. The aim in these settings is to gather data with enough variation and richness that a wide range of ad hoc, targeted analyses can be performed *retrospectively*—after data has been collected— to quickly provide precision data-driven evidence to decision makers or investigators.

As just one example, there is growing interest within the healthcare domain in using "Big Data" to help personalize care and support precision treatment decisions [28]. The so-called Learning Health System [23] concept is receiving heavy investment, with the aim of creating methods and tools that enable a data-driven environment in which evidence that informs medical treatment can be obtained via analysis of large populations of real-world patient data. The subsequent outcomes from those decisions could then be added to the population-based dataset, with new analyses of the updated data producing a powerful learning effect [1, 12, 13].

These sorts of large-scale data-driven analyses, regardless of domain, require analysts to select subsets of data that can be further analyzed to answer a specific question of interest. This *data selection* process is often accomplished using visual analysis technologies [41] that are designed to leverage interactive visualization algorithms to help user quickly and intuitively navigate complex datasets. For example, visual analysis techniques have been applied to problems in the healthcare domain, allowing users to quickly identify and select a cohort of patients for further analysis during population-based studies [44]. This form of visual data selection can be highly effective, supporting an analysis workflow that is both intuitive and high speed.

However, there is a critical limitation that must be overcome when this approach is applied to high-dimensional datasets. More specifically, there is a dramatic mismatch between the relatively small number of dimensions displayed simultaneously even in "multi-dimensional" visualization methods (e.g., most often less than 20) compared to the very high-dimensional nature of many real-world datasets (e.g., tens of thousands of features in real-world medical data).

This significant difference in dimensionality means that users performing visual data selection must apply filters with an exceptionally narrow view of the dataset they are manipulating. Unfortunately, because the variables in many real-world, high-dimensional datasets are highly interdependent, the filters applied during visual data selection can produce large amounts of selection bias that—for the vast majority of variables that are omitted from the visualization—can occur invisibly and go undetected. As a result, the bias that is introduced can be a silent yet critically limiting factor that undermines the quality of all subsequent analyses of the selected data. For example, the results of a medical study conducted using a heavily biased data sample could unknowingly encourage poor but seemingly evidenced-based treatment decisions. This is a recognized challenge for data-driven techniques in the health domain [14] and is equally problematic in other fields.

In this article, we present Adaptive Contextualization (AC), a novel approach to interactive visual data selection that is specifically designed to combat the invisible introduction of selection bias. Our approach (1) monitors and models a user's visual data selection activity, (2) computes metrics over the model to quantify the degree of selection bias introduced during the process, (3) visually represents the results for contextual awareness, and (4) provides interactive tools that help users assess bias when it is discovered and revert problematic filters to explore alternative selections. Specifically, our work offers the following contributions:

- **Algorithms for capturing and assessing the introduction of bias during data selection.** We define a context model that captures the sequence of steps performed by a

user during the visual data selection process. We then define a multi-dimensional distance metric that quantifies the shift in variable distributions between any pair of steps in our context model. The metric is recomputed automatically as the model changes in response to additional user activity.

- **Visual representations and interaction techniques.** We describe an interactive visual representation for the results produced by the algorithms outlined earlier. The visualization is updated automatically as users explore a dataset and provides interactive tools that allow users to investigate the bias metrics and revise their data selection process.
- **An evaluation of AC's effectiveness.** We include results and analysis from a 18-person study evaluating AC as applied to a medical domain problem. The results show that a data selection tool with AC (when compared to a baseline version without AC) improves awareness of variable distribution changes within a dataset during selection and reduces the time required to estimate levels of selection bias. We also share results from a qualitative evaluation conducted with population health researchers.

This article is an extended version of the first article to describe AC, which was published at the 2016 ACM International Conference on Intelligent User Interfaces [16]. In addition to a revised and expanded presentation of the AC methodology, this article complements the previously published user study findings with new evaluation results derived from qualitative interviews with a team of population health researchers (see Section 6). Finally, supplemental material is provided via the ACM Digital Library. This material includes (1) a video figure demonstrating the system in action and (2) an appendix documenting details of the evaluation process.

2 BACKGROUND AND RELATED WORK

The context-aware visualization methods proposed in this article are related to several different areas of research including high-dimensional visualization methods, provenance modeling, and previous approaches to intelligent visualization.

2.1 High-Dimensional Visualization

The application of interactive data visualization techniques to high-dimensional datasets has been a focus and challenge for the research community since the mid-1990s [7]. The challenge is rooted in constraints that derive from the very fundamentals of the data visualization concept. Visualization uses a relatively small number of visual variables (position, shape, size, brightness, color, orientation, texture, and motion) [43] to encode complex information and relies on humans' visual perception to interpret the resulting graphics to derive insight.

Using clever graphical arrangements (e.g., parallel coordinates [22] and scatterplot-matrices [9]) as well as multiple coordinated views [42], a single visual variable can be leveraged to encode more than one data variable at the same time. However, even advanced applications of these techniques, as shown in the a recent survey of state-of-the-art parallel coordinates techniques [19] are able to concurrently visualize only a relatively small number of dimensions (most often less than 20).

Given this restriction on the dimensionality of visual representations, research in this area often focuses on data summarization, projection, or ranking. This includes traditional projection methods like principle component analysis (PCA) and multi-dimensional scaling (MDS), visual clustering (e.g., Refs. [3, 8, 10, 34]), and a variety of hierarchical summarization methods [11]. Optimization-based techniques can then support algorithmic configuration of these approaches based on specific quality criteria [6]. However, summarization methods result in loss of information due to the reduction in dimensionality. Ranking-based methods, meanwhile, can prioritize dimensions for viewing but do not overcome the limited number of visualized dimensions. As a result, a majority of dimensions can remain invisible.

2.2 Provenance Modeling

Visual queries are a key benefit of many interactive visualization systems [35]. User interface controls allow users to change query constraints, while visualization is used to interactively depict the updated dataset. Visual queries can be combined with direct manipulation of the visualization's graphical objects [20] to make exploratory data selection fast and intuitive. Together, these methods can support an exploratory selection process through which users can quickly and interactively focus a visual analysis on a data subset of interest (e.g., healthcare domain examples [15, 32, 44]).

In recognition of the exploratory and ad hoc nature of these tools, researchers have developed a variety of visual provenance models. These models are designed to capture and record the often complex chain of visual data transformations that users can apply as they explore a dataset [18, 24, 27, 29, 39]. In these most basic form, these models capture trails of user activity to document the origin of a dataset [25] or to allow re-use of a previously saved sequence of analysis operations [4]. This article adopts a similar approach to monitoring and capturing user activity but uses these data interactively to actively contextualize a user's ongoing exploratory data selection process.

2.3 Intelligent Visualization

The provenance models described earlier are often captured as evidence documenting how specific visualizations were constructed or how insights were discovered. However, the same models capture detailed information about a user's analytic activity that can support a wide range of intelligent visualization algorithms. For example, algorithms have been designed to compare a user's currently visualized dataset with a representation of his/her visualization history. These have enabled, for example, user interfaces that recommend past visualizations that are most relevant to a user's current activity [37]. A similar approach has allowed for the ranking and recommendation of relevant notes captured by a user in a visualization notebook [38].

Using sequences of steps along a provenance model can also support intelligent visualization applications. For example, behavior-driven visualization recommendation [17] is a technique that analyzes user activity as it is performed to detect patterns that suggest user intent. Alternative visualizations are then recommended with the aim of better supporting a user's analytic needs. In addition, models can be collected and indexed for subsequent searching, supporting collaboration, and re-use of visualization-based data selection procedures [27]. At a high level, the AC approach outlined in this article is perhaps most similar to these intelligent visualization techniques. However, the goal is quite distinct given AC's focus on data quality and bias as introduced during high-dimensional data selection.

3 MOTIVATING SCENARIO AND BASELINE SYSTEM

As a motivating example for the challenges of high-dimensional data selection, consider the Integrated Cancer Information and Surveillance System (ICISS) managed by the UNC Lineberger Comprehensive Cancer Hospital [26]. Like similar "Big Data" resources in other domains, ICISS gathers large volumes of complex real-world data from multiple sources to build a detailed repository for retrospective analysis. In particular, ICISS integrates the North Carolina Central Cancer Registry (containing a nearly complete sample of all cancer cases in the state) with administrative and claims data for roughly 60% of "general population" patients from across the state. In total, ICISS contains electronic health data for more than six million patients.

ICISS contains a wide variety of data about these patients including demographic data, insurance information, and longitudinal medical data (including diagnoses, lab tests, medications, and procedures). Other variables include census and environmental data, behavioral data, and economic

data. All together, this results in a very high dimensional dataset, with the number of distinct variables easily exceeding 100,000.

Researchers hoping to use this data resource to retrospectively study the impact of various cancer treatments or interventions must begin with a critical but challenging first step: they must select—from this complex pool of over six million patients—a representative cohort of patients to study. This is generally accomplished by specifying a series of inclusion/exclusion criteria to whittle the population down to a group that is (1) a manageable size, (2) appropriate for given clinical question, and (3) representative of the larger population being studied.

Typically, this involves a long and extracted process where investigators attempt to communicate their data needs to a staff of technologists who then work to identify an appropriate cohort for a given study. This repeats iteratively as the clinical investigators (1) wait for the technical team to extract and refine a cohort from the database, (2) vet the results to see if the resulting cohort meets their needs, (3) make judgements about how criteria should be modified to improve the results, and (4) repeat the process until satisfied.

For example, to support one recent study at ICISS, researchers studying breast cancer narrowed in on a study cohort by along 12 dimensions, resulting in a study cohort of just 2,640 patients (from the over six million overall). Using traditional methods (without the visual selection tools described later), this type of iterative selection process can take months of effort, require high levels of technical staff support, and result in a large number of complex custom-built SQL queries and statistical analyses.

3.1 Baseline Visual Data Selection System

To support the scenario earlier in the text, a visualization-based data selection system was developed with a design similar to the recently introduced DecisionFlow system from Gotz and Stavropoulos [15]. Both DecisionFlow and our prototype adopt a design specifically created to allow ad hoc exploration of high-dimensional temporal event data, making it a good match for the electronic health data. Our baseline system, however, provides more capabilities than DecisionFlow. In particular, our baseline system allows users to iteratively apply inclusion/exclusion criteria, more closely matching the workflow outlined in the motivating scenario.

The baseline visualization system, shown in Figure 1, combines (a) a visual query panel for retrieving an initial cohort of patients from a large-scale database, (b) a visual breadcrumb showing the sequence of inclusion/exclusion constraints added during the selection process, (c) an interactive temporal event visualization panel that supports direct manipulation for defining new constraints, and (d) an interactive patient event/demographic panel that visualizes a variety of basic statistics to help users identify interesting variables within the high-dimensional data.

This baseline visualization design allows users to quickly and intuitively select focused cohorts for subsequent analysis. This approach promises to dramatically speed the cohort selection process described in our motivating scenario.

However, the increased visibility and use of correlations made possible by these visual methods means that selection bias an even more salient concern as users identify cohorts of interest. For example, filtering to include only patients with emergency admissions will skew data toward certain diseases with acute manifestations, while filtering to include patients with certain medications can result in a strong age bias. These changes are often hidden from user's view, however, because—due to the high dimensionality—they are not included in the visualization. In this way, the iterative application of multiple filters, as is typical in the motivating scenario, can produce a final cohort that is dramatically—and invisibly—different than the original. This is a recognized challenge [14] and one that the AC methods in this article are designed to help address. Moreover,

Fig. 1. The baseline visualization-based data selection system with four panels: (a) a query/constraint panel, (b) a visual breadcrumb showing steps in the data selection process, (c) an interactive visualization panel allowing user-driven patient subgrouping and inclusion/exclusion constraint definition, and (d) a panel visualizing demographic and clinical event statistics for the selected cohort.

this motivates the evaluation design presented in Section 5, which compares the baseline visual cohort selection system described here with an AC-enabled version of similar design.

4 ADAPTIVE CONTEXTUALIZATION METHODS

This section provides a detailed description of AC. It begins with a definition of a provenance model designed to model user's behavior and the evolving chain of datasets that are created during user interaction. A metric is then defined over this model to quantify differences in variable distributions across the high-dimensional space. Intelligent user interaction capabilities are then adopted automatically update the model, recompute the metrics, and surface the results for visual inspection and manipulation. Together, this AC approach provides users with clear and actionable feedback about the location and degree of bias introduced in response to their interactive data selection activity.

4.1 Provenance Model

At the core of the AC method is a data structure designed to capture the provenance behind the dataset currently being visualized by the user. This *provenance model* must capture each of the datasets visited by a user over the course of the data selection process with sufficient detail to support the metric defined later in this section.

As shown in Figure 2, the provenance model is represented as a sequential chain of datasets d_i linked by filters f_i. Each filter includes one or more constraints defined by users' interaction with the visualization. The very first dataset visualized by a user, d_0, is the dataset returned by the user's initial query. The final dataset in the chain, noted as d_{active}, corresponds to the user's currently visualized dataset. All other d_i represent datasets created as intermediate steps by the user as part of the interactive data selection process. We note that this structure represents the minimally required provenance model for AC. This representation can be extended as required to support additional provenance-based user interaction capabilities.

Given this basic representation, a visualization system can be instrumented to build and maintain the provenance model in response to user interaction. As new constraints are applied via

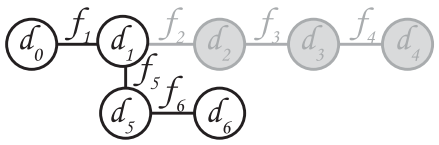

Fig. 2. The provenance model captures the sequence of user activity that led to the current dataset. It captures both the datasets at each step (d_i) as well as the filters (f_i) used to arrive at those datasets. When users revert to earlier steps to undo filters, the model is pruned (gray circles). A new branch is then created to reflect any subsequent filters. The dataset at the end of the model (in this case, d_6) is the active dataset, d_{active}.

direct manipulation with the visualization, new datasets are defined and added to the end of the chain. The model supports *reversion*, allowing users to undo one or more filters from the end of the chain. In response to a reversion, the chain of datasets is first pruned back to the selected point while new filters are extended along a new branch within the model. This is illustrated in Figure 2 which shows a model that was reverted to d_1 (resulting in the pruning of d_2, d_3, and d_4) before being extended with two new filters to arrive at the active dataset d_6.

Except for small datasets, it can be impractical to store a complete copy of each dataset along the chain. Fortunately, AC does not require that the actual data be stored in full at each node in the provenance model. Efficient visualization requires d_{active}, and intermediate datasets can be reproduced as needed by applying the corresponding chain of filters to d_0. AC does require, however, a detailed characterization of the distribution of values for each variable in each dataset d_i. These distributions are used as input to the metric defined in the next section. Therefore, each time d_{active} is updated in response to user interaction, a process runs to compute a detailed high-dimensional *variable distribution vector* for the newly created dataset, which we note as $\vec{v}_i = \{v_i^1, v_i^2, \ldots, v_i^n\}$. This vector contains a discrete probability distribution v_i^j for each individual variable j in the n dimensional dataset d_i.

4.2 Pairwise Dataset Comparison Metric

The provenance model documents each of the steps in the data selection process by which a user transforms d_0 to d_{active}. To quantify the amount of selection bias introduced, we construct a pairwise dataset comparison metric, $\delta(d_j, d_k)$, which varies from zero (to indicate that two datasets have identical variable distributions) to 1 (indicting maximally different variable distributions between datasets).

The δ metric builds on the Hellinger distance [31, 40], a statistical measure designed to quantify the similarity between two probability distributions. For discrete datasets, such as those found in typical visualization applications, a discreet probability version of the Hellinger distance can be computed as follows. For two discrete probability distributions $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_n)$:

$$H(A, B) = \sqrt{\frac{1}{2} \sum_{i=1}^{n} (\sqrt{a_i} - \sqrt{b_i})^2}, \tag{1}$$

where n is the number of discrete values for A and B.

Conceptually, this measure will be used in our algorithm to quantify the difference between the distribution of values observed for the same variable in one dataset versus another dataset (e.g., gender distribution in d_0 versus d_{active}). The uni-variate H evaluates to zero when A and B are identical and produces a value of one when A and B are maximally different.

There are many potential distance or similarity measures that could be used to quantify changes in distributions. We chose to adopt the Hellinger distance as the basis for our approach for several reasons. First, it is an established measure for comparing distributions, used widely within the statistics community. Second, the Hellinger distance provides a normalized value in the range of [0, 1], which supports its application across heterogeneous variable types. Third, the Hellinger distance is symmetric, meaning that $H(A, B) = H(B, A)$. Finally, the Hellinger distance can be computed very efficiently for discrete probability distributions. The speed of calculation is critical, because, as described later in this section, it is computed a very large number of times in response to a user's interaction with the visualization system.

Within our AC algorithm, this discrete form of the Hellinger distance can be applied directly to categorical and ordinal values. Ratio variables, meanwhile, should typically be binned to convert them to ordinal measures before computing the discrete probability distribution. This binning step can prevent n values that are relatively large in comparison to the dataset size. As n grows larger, there is risk of over-sensitivity to small changes in variable distributions.

While H provides a uni-variate measure of similarity, the datasets in our work are high dimensional in nature, often containing tens of thousands of unique dimensions. We therefore define the multi-variate distance measure δ using a weighted average of the uni-variate Hellinger distances across all dimensions m in our dataset,

$$\delta(d_j, d_k) = \frac{\sum_{i=1}^{m} w_i * H(v_j^i, v_k^i)}{\sum_{i=1}^{m} w_i}, \tag{2}$$

where v_j^i and v_k^i are the discrete probability distributions for the ith variable datasets d_j and d_k; and $w_i \in \{0, 1\}$ is the weight for the ith variable. This produces a measure which, like the traditional uni-variate Hellinger distance, ranges from zero (for datasets with identical variable distributions) to 1 (for datasets whose variable distributions are maximally different).

The weights in this measure are used to ignore distances contributed by dimensions that have been explicitly filtered by the user. More specifically, variables that have been explicitly constrained within of filters f_i found prior to the datasets d_j and d_k within the provenance model are assigned weights of zero. All other variables are assigned weights of 1. In this way, the δ measure only considers differences in variable distributions that occur implicitly as a confounding side effect.

For example, consider the motivating medical scenario. If a dataset containing patient medical data was filtered by a user to contain only men, then the gender variable would be assigned a weight of zero when comparing the two datasets (all patients vs. men only). This would ensure that the expected differences in gender distribution would not contribute to the result of the distance measure. However, hidden differences in correlated variables (e.g., differences in the prevalence of maternity-related procedures) would be detected.

4.3 Metric Visualization and Interaction

As users go about the data selection process using an AC-enabled visualization system, the AC algorithm monitors user interaction and dynamically updates the data provenance model after each filter. As d_{active} changes, new δ values (Equation (2)) are computed for each new pair of datasets in the model. The δ values, along with the individual H values computed for each variable (Equation (1)), are then made available via the user interface to highlight for users where the largest biases have been introduced.

The δ and H metrics are then used to highlight and prioritize areas of emerging selection bias for the user. To present this information, the baseline user interface shown in Figure 1 is expanded to include two new intelligent visualization capabilities. First, highlighted in Figure 3(a), a contextualized breadcrumb view provides additional contextual information compared to the baseline

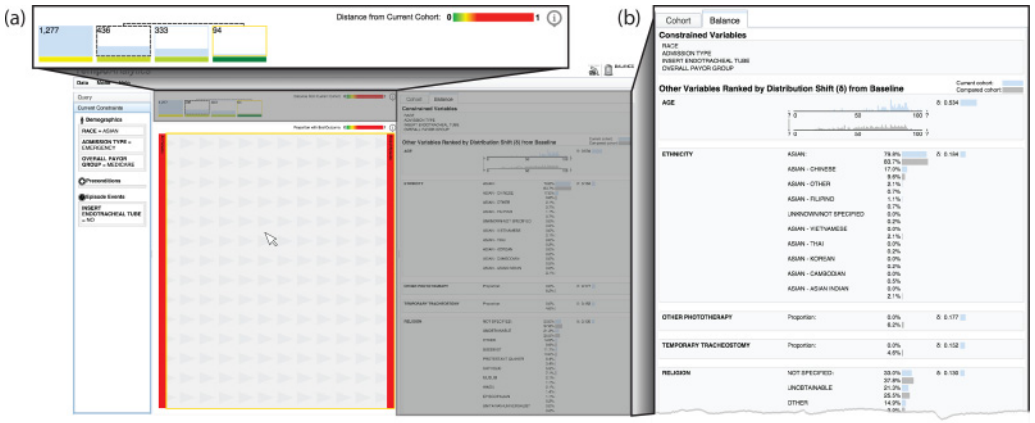

Fig. 3. The user interface for our AC-based visualization system extends the baseline interface of Figure 1 in two key ways. First, (a) the breadcrumb panel showing the user's history of datasets includes a visualization of the δ metric and supports new interactions to support dataset comparisons. Second, (b) a new Balance Panel has been added to visualize detailed ranked lists of univariate differences between datasets.

breadcrumb design and supports a new set of user interaction capabilities. Second, highlighted in Figure 3(b), a new *Balance Panel* supports detailed univariate comparisons for selected pairs of datasets. A high-level representation of the two main interaction paths supported by these panels is shown in Figure 4.

4.3.1 The Contextualized Breadcrumb. The contextualized breadcrumb panel is designed to help users understand (1) the filters that have been applied at each step of the data selection process and (2) how the dataset d_{active} compares to those previously visualized datasets in terms of underlying variable distributions. To achieve this goal, a glyph-based design has been developed as illustrated in Figure 5. In this design, each dataset in the provenance model, starting with d_0 and continuing to d_{active}, is represented by its own glyph. The glyphs are positioned from left to right, with d_0 appearing first. As new datasets are visited, the chain of glyphs is extended to the right as shown in Figure 6.

Each glyph shows the size of the dataset with both a number and a blue indicator whose height is proportional to the size of the dataset. This provides a simple bar-chart view of the changing dataset size as filters are applied. In addition, a color-coded rectangle, which we call a δ *bar*, is positioned at the bottom of each dataset's glyph. For each dataset d_i, the color of the bar is determined by the value of $\delta(d_i, d_{active})$. This value is then mapped to a green-to-yellow-to-red color scale. A red δ bar represents a dataset that has major differences in variable distributions compared to the active dataset. A green bar, meanwhile, represents a dataset that is very similar to the active dataset. This explains, for example, why the active dataset (highlighted with the gold border) has a pure green δ: It shows no bias when compared to itself (i.e., $\delta(d_{active}, d_{active}) = 0$).

Users can access additional information via interaction. Mousing over the portion of the glyph containing the blue size indicator shows the specific filter applied to arrive at the corresponding dataset. Meanwhile, mousing over the δ bar provides a high-level summary of the bias detected between the dataset and $d_{actcive}$. In particular, as shown in Figure 6(f), users can see the actual δ score as well as a list of the three variables with the largest difference in distributions as measured by Equation (1).

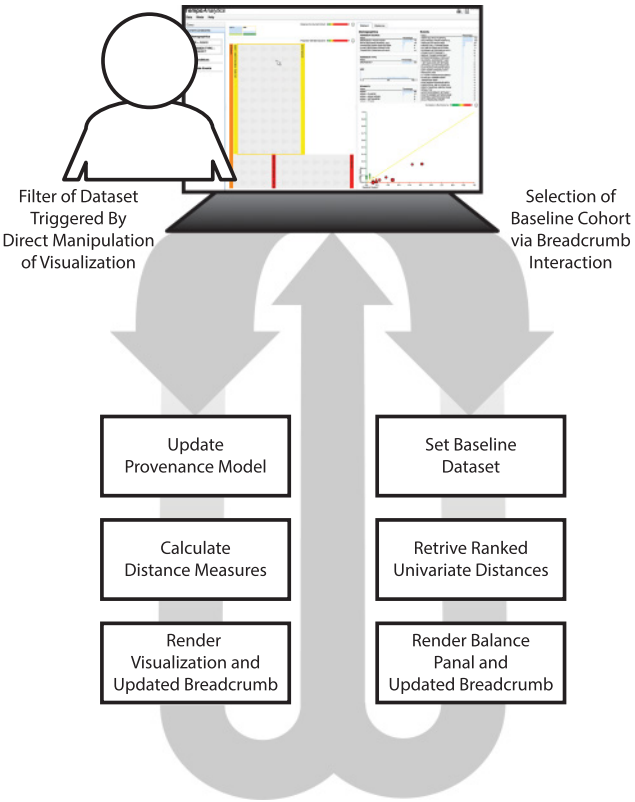

Fig. 4. Users interact with the system via one of two main interaction paths. (a) Normal data selection is performed via the visualization panel. Users navigate the active dataset and apply filters via direct manipulation of the visualization's graphical elements. In response, the system updates the provenance model, calculates new distance measures, and renders the updated visualization and breadcrumb views. (b) The user interacts with the breadcrumb to initiate a comparison between the active dataset and a user-selected baseline.

Finally, users can click on any glyph in the breadcrumb to select the corresponding dataset. If the selected dataset is *not* d_{active}, then the breadcrumb view is updated with a dotted black selection line (see Figure 5) that connects the selected dataset with d_{active}. Coordinated with this selection, the user interface displays the Balance Panel described later to support detailed investigation of the differences between the datasets. Clicking on the active dataset returns the visualization to its normal exploratory data selection model. Right-clicking on a dataset brings up a context menu that allows users to "go back" to an earlier dataset to explore alternative selections.

4.3.2 The Balance Panel. When a dataset d_i (other than d_{active}) is selected in the contextualized breadcrumb view, the Balance Panel (Figure 3(b)) is presented to the user with the aim of supporting detailed comparisons between the selected dataset and the user's currently active dataset. This panel, in essence, provides users with a prioritized visualization of the individual univariate H scores (Equation (1)) that contribute to the overall δ value.

The top of the balance panel provides a list of constrained dimensions. While a detailed list of active constraints are also included in the leftmost sidebar of the user interface (where the query is performed), this section of the panel provides users with a reminder that the listed variables have

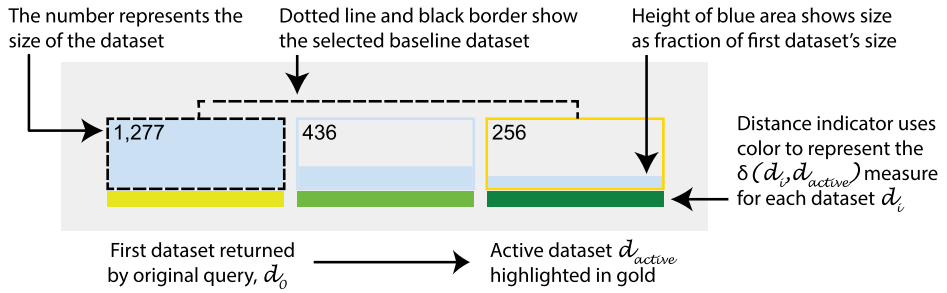

Fig. 5. The contextualized breadcrumb view uses color-coded bars at the bottom of each glyph to encode the δ measures. Interaction capabilities revert to prior datasets or select a prior dataset for detailed comparison in the Balance Panel.

been explicitly constrained and are therefore omitted from the data that is displayed below it. In particular, these are the dimensions for which w_i is set to zero in Equation (2).

The remainder of the balance panel provides a ranked list of univariate visualizations. The visualizations are sorted by H score, placing at the top of the list the variables whose distributions have been changed the most given the filters that created d_{active} from d_i. For each variable, two distributions of values are visualized: (a) the values from the active dataset and (b) the values from the baseline dataset. By visualizing the two distributions in juxtaposition with each other, the panel allows users to quickly "drill down" into a given H score and see what specific differences there are in values between the two datasets.

An example of the univariate visualizations included in the balance panel is shown in Figure 3(b). This example shows that the Age variable exhibits the largest change in distribution given its location at the top of the panel. The histograms illustrate the reason for the high ranking: The active cohort (in blue) has far fewer young patients. This produces a distribution with a much higher average and lower variance. The second largest shift takes place within the ethnicity variable. The detailed view of the distribution of values, however, shows an interesting pattern. While there are fewer patients identified as simply "Asian," there are more patients identified as "Asian–Chinese." This manifests itself as a large change in distribution, but in practice the difference may not be semantically meaningful. This example shows a key benefit of providing these tools to domain expert users. What appears statistically meaningful may in fact be semantically insignificant, and including visualization tools within the process can help users make more informed analytical decisions.

The balance panel is critical for users, because it allows them to understand *why* a large δ has developed during the data selection process. An example workflow from our motivating scenario is illustrated in Figure 7. First, a user is shown iteratively applying multiple filters (Figures 7(a)–(c)) before clicking on the contextual breadcrumb to learn which variables have been most biased (Figure 7(d)). The balance view tells the user that the age variable is most biased variable. Using the balance view to look at other datasets in the breadcrumb, the user learns that this bias introduced by a filter to a specific type of hospital admission.

5 USER STUDY FOR TASK PERFORMANCE EVALUATION

To evaluate the benefits of AC to users during high-dimensional data selection, we conducted a user study comparing our AC-based visual data selection prototype to a baseline system in which AC features were disabled. The study required users to perform tasks related to the selection of a patient cohort from a database [30, 33] containing electronic medical record data for approximately

Fig. 6. The contextualized breadcrumb view grows the right as users apply new filters. ((a)–(d)) The color-coded δ bars are updated in response to these new datasets, providing up-to-date representation of the measured bias. (e) Mousing over the main glyph area shows the filters applied at that step, while (f) hovering over the δ bar.

30,000 patients with over 3,000 variables.[1] We note that while the data was medical in nature, the study tasks were designed to require no medical knowledge or background.

5.1 Participants and Testing Environment

A total of 18 users (12 female, 6 male) were recruited to participate in the study, with ages within the range of 21–50. All participants were college educated with degrees in science, technology, engineering, or math (STEM) fields. As a result, all were familiar with the basic statistical concepts relevant to the study. Moreover, all participants had either completed or were currently enrolled in a graduate degree program. Backgrounds included public health, statistics, and information science. However, none of the participants were epidemiologists or clinicians at ICISS. The study tasks (see the Electronic Appendix for this article) were designed to be completed by users without detailed domain knowledge.

The 18 participants were randomly assigned into two groups of nine: a *Baseline* group and an *AC* group. None of the participants had any prior experience with any of the software evaluated in the

[1]Variables include demographics, thousands of distinct medical procedures, and other medical events (e.g., admission and discharge).

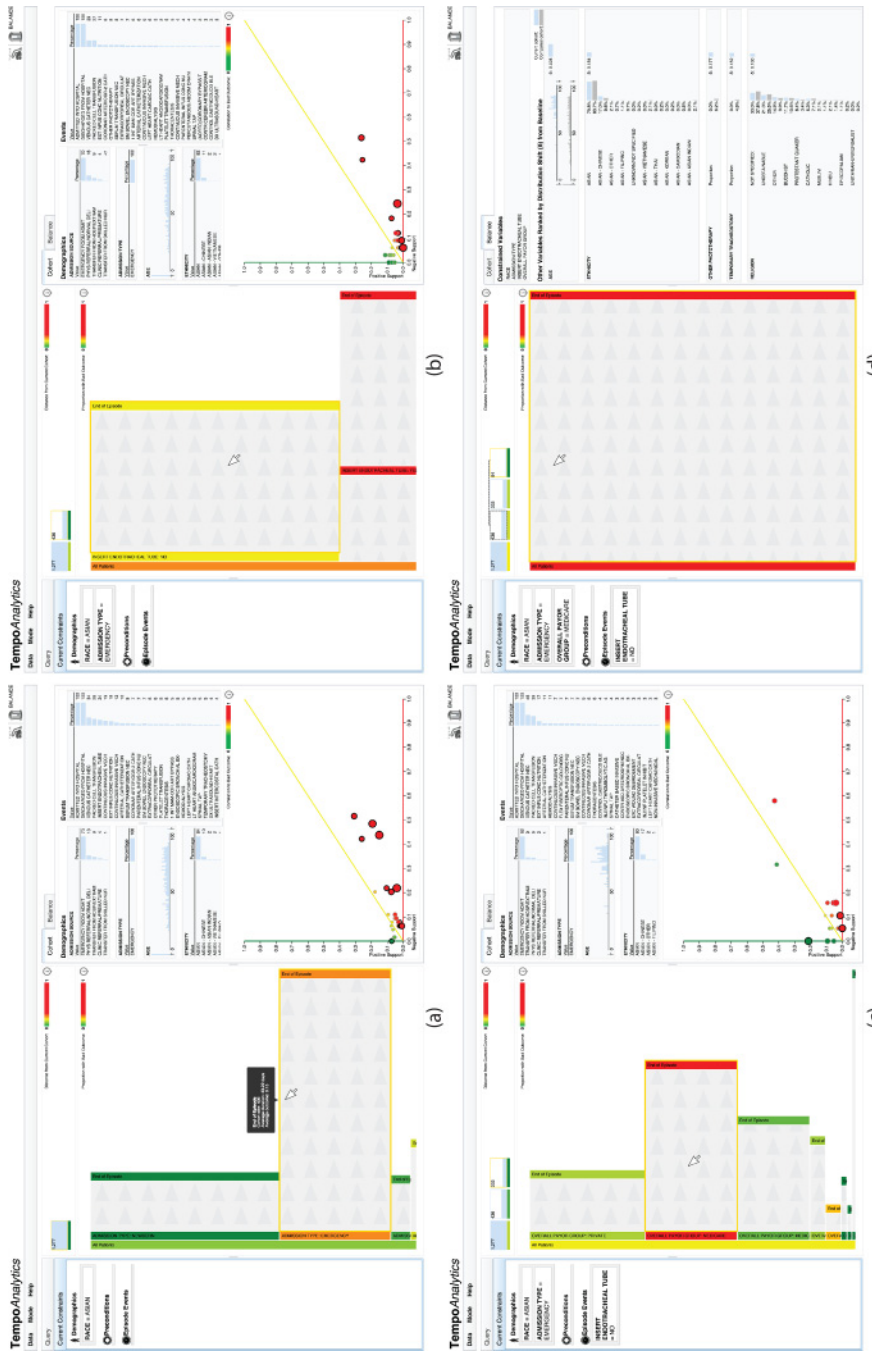

Fig. 7. A sample interaction sequence showing a user applying multiple filters to narrow down to a narrow cohort of patients. The breadcrumb at the top shows the growing provenance model as it expands from (a) one step to (d) four steps. In (d), the user has switched to the Balance view to inspect the differences between two datasets as represented by the dotted black line in the breadcrumb view.

study. The study tasks were performed on a 13.3-inch laptop computer with a display resolution of 1,280 × 800 pixels and a 60Hz refresh rate.

5.2 Procedure

Each user took part in single study session lasting roughly one hour. Each session began with a brief introduction to the data that would be used during the study, followed by a tutorial during which the moderator described the various features of the software that would be used to complete the study tasks. Users in the *AC* group were given access to the advanced AC features described in this article. Users in the *Baseline* group were given access to a version of the same software but with the AC-related features disabled (see Figure 1). Regardless of group, participants were then instructed to practice with their assigned software tool and asked to perform six specific practice tasks. The moderator provided additional help and instruction when needed during these practice tasks.

Once a user was comfortable using the software assigned to her/his group, the experimental portion of the study began. The participant was asked to perform six formal study tasks that were very similar to the six practice tasks but focused on different subsets of the study dataset. Both groups of users performed identical tasks with identical data, but with different versions of the software.

While the formal study tasks were performed, the moderator did not provide any assistance or instruction in how to use the system. The moderator recorded both accuracy and time-to-completion for Tasks T1–T4. Tasks T5 and T6 focused on the more subjective tasks of bias assessment and confidence, respectively. Once finished with the study tasks, users completed a post-session questionnaire with nine 5-point Likert scale questions (Q1–Q9) and two free-response questions that asked users to comment regarding the most and least helpful aspects of the software they were given to perform the tasks.

Finally, after the questionnaire, participants in both groups were debriefed by the moderator to gather additional qualitative feedback. Moreover, users in the *Baseline* group were given a demonstration of the full AC prototype and asked to comment regarding the additional features. More details, including the full text for all tasks and questions used in the study are provided in the Electronic Appendix for this article.

5.3 Results and Discussion

The results from our study show that AC can help contextualize the visual data selection process, and help users more effectively detect and characterize emerging selection bias. This section presents the results obtained from our study and discusses the implications of those results for AC-based visualization systems.

5.3.1 Study Tasks. Each participant in the study completed six study tasks T1–T6 (see the Electronic Appendix for this article). The first four tasks (T1–T4) were designed to reflect the iterative "assess, revise criteria, assess, revise criteria" workflow that is typical of the cohort selection process. Overall, users from both the *AC* and *Baseline* groups completed these four timed tasks (T1–T4) with high accuracy. However, there were statistically significant[2] differences observed in task completion time for two of the four timed tasks as illustrated in Figure 8.

T1: Users were asked to compare the mortality rate for across different patient subgroups within the same dataset. This is a task for which AC was not expected to provide any meaningful benefit,

[2]Statistical significance for tasks T1–T6 was determined using a standard t-test comparing results between *AC* and *Baseline*. A total of 18 observations were used for each t-test (one for each participant) with nine observations in each group.

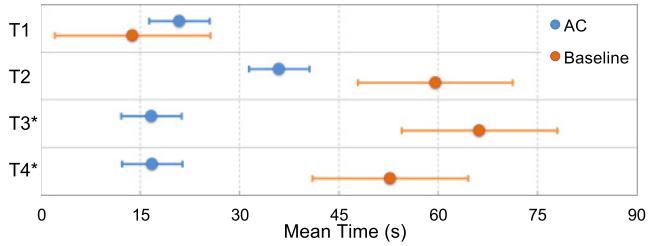

Fig. 8. Mean time-to-completion (in seconds) for the four timed tasks in our study (T1–T4). Error bars show the standard error. Tasks T3 and T4 were performed significant faster by the AC group ($p < 0.01$).

and therefore similar performance was expected across groups. Indeed, both groups answered accurately well with only one incorrect answer. A user in the *AC* group arrived at the wrong answer after confusing the variable "Admission Type" with "Admission Source" and therefore selected the wrong variable for comparison. However, the user did perform the task correctly even if the final response did not provide the correct answer.

T2: Users were asked to apply a filter to create a new dataset, then compare the values observed for a single variable across the "before" and "after" datasets. Users in all groups arrived at the correct answer, while *AC* user performed the task more quickly. The difference in timing, however, was not statistically significant given our sample size. We hypothesize that the Balance Panel proved useful for those that completed this task in the least amount of time, but more study is required to reach a stronger conclusion.

T3: Users were next asked to compare more broadly the two datasets produced after T2, comparing all variables rather than a single specific dimension. All users were again able to provide accurate answers. However, the *AC* group arrived at their answers in significantly less time ($p < 0.01$). This result suggests that the Balance Panel was highly beneficial in helping users characterize differences between datasets.

T4: Users were asked to apply an additional filter, resulting in three datasets in the breadcrumb panel. Users were then asked to identify which of the first two datasets was most similar to the new active dataset that was just created. AC proved most useful in this case. Participants in the *AC* group performed the task significantly faster and provided accurate answers. Meanwhile, participants in the *Baseline* group faced far more difficulty resulting in significantly slower times ($p < 0.01$). Moreover, one user in the *Baseline* group simply abandoned the task claiming it was not possible to answer. The task completion time for this user were therefore omitted from the results shown in Figure 8 and our statistical analysis. The user did remain to complete the study session, but including his results for T4 would have produced an even stronger effect. Moreover, we believe that this user's behavior is emblematic of the much more difficult cognitive work required by participants in the *Baseline* group to complete this task.

T5 and T6: Unlike the previous tasks, T5 and T6 asked participants for subjective assessments. Users asked (T5) to state how representative the final dataset was with respect to the original query result; and (T6) state how confident they were in their assessment. Interestingly, both responded similarly that the final dataset was quite biased (*AC*: mean of 1.89; *Baseline*: mean of 1.69; on a 5-point scale with 1 = "very biased" and 5 = "highly representative") and had similar levels of confidence in their assessments (*AC*: mean of 3.44; *Baseline*: mean of 3.88; on a 5-point scale with 1 = "very unsure" and 5 = "highly confident"). The differences were not statistically significant, making it impossible to draw any firm conclusions. However, we do note that the *AC* user were less extreme, reporting in answers for both questions which, on average, were closer to the midpoint

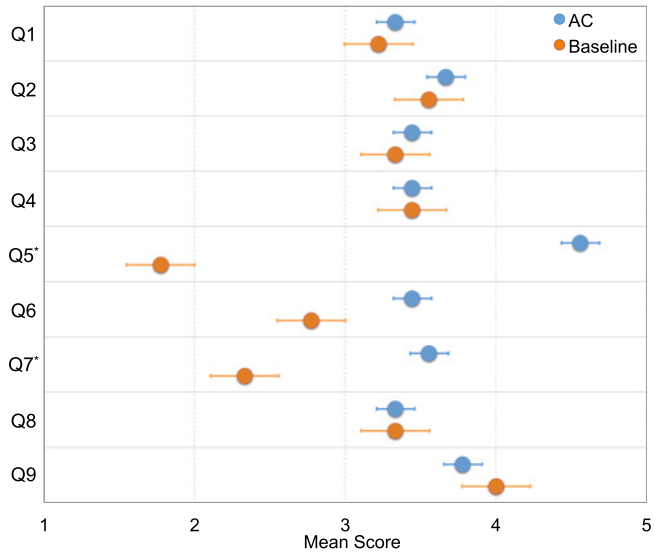

Fig. 9. Average responses for each of the nine 5-point Likert scale questions (Q1–Q9). Error bars show the standard error. Questions Q5 ($p < 0.05$) and Q7 ($p < 0.1$) show statistically significant differences between groups.

of the scale. The subjective feedback reported in the next section provides a more nuanced view of the participants' opinions.

5.3.2 Questionnaire and Moderator Debriefing. Users' feedback gathered via the questionnaire (see Figure 9) suggests that users clearly recognized certain benefits of using an AC-enabled visualization system for data selection. Moreover, these benefits appear to come without any penalty in terms of ease-of-use. Because the questionnaire collected user responses via a 5-point Likert scale, a non-parametric Mann-Whitney analysis was performed to analyze the results.[3]

Q1–Q4: Q1 asked users how easy or hard it was to "learn how to interpret" the visualization, while Q2 as how easy or hard it was to "learn how to interact with" the visualization tool. Q3 asked users to score how easy the system was to use *after* the initial learning curve. The responses were similar for both the *AC* and *Baseline* groups for all three questions, falling in the middle of the range. This suggests the inclusion of AC capabilities does not have any meaningful impact on either the learning-curve of ease-of-use for the system.

From the moderator debriefings, it was observed that users with prior experience with other data visualization tools found the interface very easy and convenient to use, while others took more time to become comfortable. However, even the participants with a steeper learning curve were able to complete the study session within the expected time.

Q5–Q7: The next three questions all suggested that the *AC* group felt more empowered for key selection bias assessment tasks. Q5 asked if it was easy to compare a dataset from one step to a dataset from another step in the selection process. Users in the *AC* group were far more likely to agree, and the difference was statistically significant ($p < 0.05$). Similarly, participants from the *AC*

[3]Statistical significance for Q1–Q9 was determined using a Mann-Whitney U test comparing results between the *AC* and *Baseline* groups. A total of 18 responses were used for each of these U tests (one for each participant) with nine responses in each group.

group were in stronger agreement with Q6, which asked users if the system made it easy to detect when a filter produce datasets that exhibited large amounts of selection bias. The difference for Q6, however, was not statistically significant given our relatively small sample size. Finally, the *AC* group was significantly ($p < 0.10$) more likely to agree with Q7, which asked if the visualization made it easy to learn which dimensions were most biased from the original dataset after applying multiple filters.

Q8: Participants responded similarly across groups to Q8, which asked about comparing variable distributions across any arbitrary pair of datasets rather limiting the comparisons to d_{active}. In both the Baseline and AC software prototypes, users would need to revert to a prior dataset using the breadcrumb view, which made this task harder to perform than other comparisons.

Q9: Perhaps surprisingly, users from the *Baseline* group agreed more strongly that the system they used in their study session provided sufficient information to properly assess the validity of the dataset they selected. The differences were not statistically different. However, we believe that the users with AC were provided with a more nuanced view that highlighted areas of bias. This in turn fosters deeper suspicion—and rightly so—within the user population about the quality of their data selection.

These results show that users in the *Baseline* group were equally or more confident in their final data selection. However, this confidence is misplaced as *Baseline* users also found it harder to compare datasets, harder to assess differences in individual dimensions across steps of a data selection process, and performed tasks more slowly and with more cognitive effort. We believe that this highlights the critical benefits AC, without which users may dangerously proceed with an analysis of a dataset both confident of its quality and ignorant of any underlying bias which may threaten then validity of subsequent analysis results.

These conclusions were further supported by the feedback in the open-ended questions at the end of the questionnaire. Users in the *AC* group said the color-mapped indicators for the δ measure of each dataset "useful"/"very helpful" for dataset comparison. Meanwhile, five of the eight users in the *Baseline* group expressed a wish for more straightforward ways to compare the datasets rather than manually going back and manually comparing variable by variable.

Finally, at the very end of the study session for members of the *Baseline* group, the moderator revealed the AC features that were made available to users in the *AC* group. Every *Baseline* user said the study tasks would have been much easier with these new features. Comments from these users after the reveal included "nice!," "that's cool," and "a significant improvement if you want to compare between cohorts." Meanwhile, users in the *AC* group felt the tasks would have been impossible without the added features, or at least would have been much harder and taken much longer.

Interestingly, one user in the *AC* group brought up in the discussion that he would have had less confidence in his assessments (e.g., T6, Q9) if he had been in the *Baseline* group without access to the AC-based features. However, as already discussed, the study results don't support his assertion. Instead, we believe that users without AC—rather than wishing for features such as those that AC provide—simply proceed with false confidence, because the bias being introduced during selection is often hidden within the many variables that have been omitted from the visualization.

6 QUALITATIVE EVALUATION VIA PRACTITIONER INTERVIEWS

To complement the results from our formal user study, we collaborated with a team of population health researchers from the UNC Lineberger Comprehensive Cancer Center to conduct a real-world evaluation of our methods. We applied our AC-enabled visualization software (as used by the *AC* group in Section 5) within the ICISS system first described in the motivating scenario of Section 3, and conducted semi-structured interviews with the researchers to better understand

how AC techniques align with their day-to-day needs and how such tools could be extended in the future.

6.1 Procedure and Participants

As described in Section 3, ICISS is a "Big Data" resource that is built and maintained by the UNC Lineberger Comprehensive Cancer Center to support a range of retrospective analysis studies. While the full system contains data for over six million patients, the typical workflow for ICISS analysts begins with a "rough cut" of the overall population, which is then further narrowed down with a series of filters to produce the final cohort for a specific study. Therefore, the first step in evaluating the *AC* prototype with ICISS analysts was to import a rough cut of data into the visual analytics system.

We obtained an existing rough cut dataset of 18,945 breast cancer patients that was being actively used within a research study supported by ICISS. In the conduct of that study ICISS analysts further refined the rough cut cohort, eventually producing a formal study cohort of 2,640 patients. That smaller subset (about 14% of the original rough cut) was obtained by iteratively applying further inclusion or exclusion criteria based on exploratory analyses of the remaining data.

To help analysts to more directly compare and contrast our *AC*-based visualization tools with their existing set of tools, we exported data for the full set of 18,945 patients from a SAS-based data repository and transformed the data into the format required by our prototype. We then ingested those data files into our software prototype.

Once the prototype software was ready to use with the ICISS dataset, we conducted a series of semi-structured interviews with seven ICISS analysts over a period of four days. All seven subjects were full-time employees of the UNC Lineberger Comprehensive Cancer Center and had at least Masters-level education in a quantitative discipline. Reflecting the interdisciplinary needs of a data-driven population-based surveillance system, employees were trained in fields such as epidemiology, biostatistics, computer science and medicine. All seven employees had real-world experience working with ICISS data within a variety of population health activities.

During the interviews, analysts were first briefed on the software prototype's design and capabilities. This briefing included a demo of basic functionality as well as some hands on exercises. The analysts were then asked to perform free-form data exploration and selection using the software. While the analysts used the software to explore their dataset, a moderator took notes about their behavior with the system and recorded users' feedback.

6.2 Results and Discussion

The interviews captured a wide range of feedback from ICISS analysts. Given the analysts focus on their daily workflow (rather than the comparatively narrow contributions of the novel methods outlined in this article), many comments were focused on specific user interface design suggestions or missing features that would make the software more suitable as a day-to-day tool. This focus was expected given the analysts' perspective that were showing them a new tool, rather than evaluating a single feature. This tendency was exacerbated by the fact that we asked the users to examine their own data with the new tool.

High level feedback suggested that the prototype was "easy to understand" and that it was "easy to interact with" the software. The analysts also felt that the cohort comparison features were valuable. The results here in general reflected the findings of the user study described in Section 5.

However, the users also suggested certain features which our prototype did not support. First, multiple users requested the ability to compare included versus excluded patients at each step. For example, suppose a user starts with one dataset which they then filter to include only men. The current prototype allows users to compare the overall cohort with the cohort of men. However,

there is no way in the current prototype to compare the cohort of men with the cohort of patient excluded by the gender filter: the corresponding cohort of women. Given this feedback, this would be a very valuable feature to explore in future work. Such a feature would also result in a stronger alignment between the AC method's breadcrumb view and the CONSORT flow diagram used widely for the reporting of clinical trials [2, 5, 21].

Related to this, users expressed the desire for a richer provenance model that captured more than a linear chain of cohorts. In particular, while the ability to revert to prior cohorts to explore new alternatives was valued, users expressed dissatisfaction with the fact that pruned datasets were lost. Such a change would increase the power of the tool, but come at the expense of a more complex user interface. For power users such as those at ICISS, a more complex interface may be acceptable and the idea will be explored in future work.

Users also recognized the value of seeing a ranked list of shifted variables. However, some also expressed the opinion that treating all variables as equal was not appropriate. The methods described in this article do indeed treat all variables as equal contributors to the distance measure. However, the distance measure defined in Equation (2) can be easily extended to support a per-variable weighting factor. As described in this article, the δ measure restricts values of the weight w_i to either zero or one. This is used to ensure that shifts to explicitly filtered variables are not reflected in the measure's value.

In practice, however, the same weighting algorithm could be used to allow users to explicitly ignore variables even if they have not been used in a filter operation. In this way, users could manually mark variables as unimportant to have them excluded from the distance calculations. Moreover, the algorithms for computing δ allow, in theory, for weights of any value in the range $[0, 1]$. This could allow relative weighting between variables. The real challenge to overcome in address this concern, however, is not algorithmic. Instead, it is a user interface challenge. We plan in future work to experiment with alternative ways of allowing users to define per-variable weights in an intuitive and efficient manner. The manual configuration of a large number of weights is a difficult user interface challenge, so the use of intelligent interface methods for this purpose has significant potential.

7 CONCLUSION

Large-scale datasets are being gathered in many domains with the goal of supporting data-driven decision making. While interactive data visualization can play a critical role in this process, there is a critical mismatch between (1) the very large number of dimensions in many complex real-world datasets and (2) the much smaller number of dimensions that can be concurrently visualized using modern techniques. This gap in dimensionality can place an analysis at high risk of hidden selection bias during exploratory data selection tasks. This article described AC, a novel approach to interactive visual data selection that is specifically designed to combat this challenge. The AC approach (1) captures a model of users' visual data selection activity, (2) computes metrics over that model to quantify the amount of selection bias after each step, (3) visualizes the metric results, and (4) provides interactive tools that help users detect and assess sources of bias as they emerge.

The results from our formal user study provide evidence for the benefits of our approach. However, there remain many avenues for future work. In particular, we plan to examine intelligent ways to help users minimize the impact of bias through the use of intelligent data transformation operations. In addition, we plan to explore other data quality measures which can be computed in similar ways to address challenges beyond selection bias. We also plan to expand our metric to account for differences in variable importance as requested by analysts in our expert user interviews. Finally, we plan to conduct more thorough, longitudinal evaluations of our approach through a series of multi-dimensional long-term case studies [36].

APPENDIX

A USER STUDY TASKS

The user study described in Section 5 included 18 participants split into two groups. As described in Section 5.2, each participant was asked to perform six practice tasks and six study tasks. The tasks were designed to be similar to each other, but focused on different subsets of the study dataset. This section of the appendix includes the actual tasks used during the study.

A.1 Practice Tasks

The six practice tasks were as follows:

Practice Task 1:

- **Moderator:** Query for a cohort of patients where "Race = Hispanic"
- **Question:** Which admission type has the best outcome (fewest deaths)?

Practice Task 2:

- **Moderator:** Add a filter for "Emergency Admissions"
- **Question:** How does age compare between the two cohorts (just "Emergency Admissions" versus all admission types)?

Practice Task 3:

- **Question:** What other differences exist between these two cohorts (just "Emergency Admissions" versus all admission types)?

Practice Task 4:

- **Moderator:** Starting with the emergency admissions cohort, apply a new filter for "VENOUS CATHETER NEC = Yes"
- **Question:** This new cohort is most similar to (a) the original query cohort, (b) the emergency admissions cohort, or (c) not sure/I'd have to guess.

Practice Task 5:

- **Question:** How representative is this cohort of the overall population? [Asked with a scale of 1 through 5, where 1 = very biased and 5 = highly representative]

Practice Task 6:

- **Question:** How confident are you in your assessment (in Practice Task 5)? [Asked with a scale of 1 through 5, where 1 = very unsure and 5 = highly confident]

A.2 Study Tasks

The six study tasks (referred to as **T1** through **T6** in the article) were as follows:

Study Task 1 (**T1**):

- **Moderator:** Query for a cohort of patients where "Race = Asian"
- **Question:** Which admission type has the best outcome (fewest deaths)?

Study Task 2 (**T2**):

- **Moderator:** Add a filter for "Emergency Admissions"
- **Question:** How does age compare between the two cohorts (just "Emergency Admissions" versus all admission types)?

Study Task 3 (**T3**):

- **Question:** What other differences exist between these two cohorts (just "Emergency Admissions" versus all admission types)?

Study Task 4 (**T4**):

- **Moderator:** Starting with the emergency admissions cohort, apply a new filter for "VENOUS CATHETER NEC = Yes"
- **Question:** This new cohort is most similar to (a) the original query cohort, (b) the emergency admissions cohort, or (c) not sure/I'd have to guess.

Study Task 5 (**T5**):

- **Question:** How representative is this cohort of the overall population? [Asked with a scale of 1 through 5, where 1 = very biased and 5 = highly representative]

Study Task 6 (**T6**):

- **Question:** How confident are you in your assessment (in Practice Task 5)? [Asked with a scale of 1 through 5, where 1 = very unsure and 5 = highly confident]

B POST-SESSION QUESTIONNAIRE

As described in Section 5.2, every participant in the user study completed a post-session questionnaire. The first nine questions (referred to as **Q1** through **Q9** in the article) all asked the user to answer using the same five-point scale: Strongly Disagree, Somewhat Disagree, Neither Agree nor Disagree, Somewhat Agree, and Strongly Agree.

- **Q1**: It was easy to learn how to interpret the visualization's graphic design.
- **Q2**: It was easy to learn how to interact with the visualization system.
- **Q3**: Once I was finished learning how the visualization system worked, I found it easy to use.
- **Q4**: It was easy to remember the sequence of cohorts that were visualized as new filters were applied.
- **Q5**: It was easy to compare cohorts from one step of the analysis to another.
- **Q6**: It was easy to detect when a filter produced cohorts that exhibited large amounts of selection bias.
- **Q7**: The visualization made it easy to learn which dimensions were most biased with respect to the original dataset after applying multiple filters.
- **Q8**: The visualization made it easy to learn which dimensions were most biased when comparing any arbitrary pair of cohorts.
- **Q9**: The visualization system provided me with valuable information that could help me improve the validity of my analysis result.

The post-session questionnaire concluded with two free-response questions.

- **Q10**: What did you like best about the visualization system?
- **Q11**: What did you dislike most about the visualization system?

REFERENCES

[1] Amy P. Abernethy, Lynn M. Etheredge, Patricia A. Ganz, Paul Wallace, Robert R. German, Chalapathy Neti, Peter B. Bach, and Sharon B. Murphy. 2010. Rapid-learning system for cancer care. *J. Clin. Oncol.* 28, 27 (Sep. 2010), 4268–4274. DOI: http://dx.doi.org/10.1200/JCO.2010.28.5478

[2] D. G. Altman, K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Gøtzsche, T. Lang, and Consort Group (Consolidated Standards of Reporting Trials). 2001. The revised consort statement for reporting randomized trials: Explanation and elaboration. *Ann. Intern. Med.* 134, 8 (April 2001), 663–694.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. 1998. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. 52–60.

[4] L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. 2005. VisTrails: Enabling interactive multiple-view visualizations. In *IEEE Visualization*. 135–142. DOI: http://dx.doi.org/10.1109/VISUAL.2005.1532788

[5] C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, and D. F. Stroup. 1996. Improving the quality of reporting of randomized controlled trials. The consort statement. *J. Am. Med. Assoc.* 276, 8 (Aug. 1996), 637–639.

[6] E. Bertini, A. Tatu, and D. Keim. 2011. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. Vis. Comput. Graph.* 17, 12 (Dec. 2011), 2203–2212. DOI: http://dx.doi.org/10.1109/TVCG.2011.229

[7] Andreas Buja, Dianne Cook, and Deborah F. Swayne. 1996. Interactive high-dimensional data visualization. *J. Comput. Graph. Stat.* 5, 1 (Mar. 1996), 78–99. DOI: http://dx.doi.org/10.2307/1390754

[8] Nan Cao, David Gotz, Jimeng Sun, and Huamin Qu. 2011. DICON: Interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comput. Graph.* 17, 12 (2011), 2581–2590. DOI: http://dx.doi.org/10.1109/TVCG.2011.188

[9] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. Scatterplot matrix techniques for large N. *J. Am. Stat. Assoc.* 82, 398 (1987), 424–436.

[10] Keke Chen and Ling Liu. 2004. VISTA: Validating and refining clusters via visualization. *Inf. Vis.* 3, 4 (Dec. 2004), 257–270. DOI: http://dx.doi.org/10.1057/palgrave.ivs.9500076

[11] N. Elmqvist and J. Fekete. 2010. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Trans. Vis. Comput. Graph.* 16, 3 (2010), 439–454. DOI: http://dx.doi.org/10.1109/TVCG.2009.84

[12] Lynn M. Etheredge. 2007. A rapid-learning health system. *Health Affairs* 26, 2 (Mar. 2007), 107–118. DOI: http://dx.doi.org/10.1377/hlthaff.26.2.w107

[13] Charles P. Friedman, Adam K. Wong, and David Blumenthal. 2010. Achieving a nationwide learning health system. *Sci. Transl. Med.* 2, 57 (Nov. 2010). DOI: http://dx.doi.org/10.1126/scitranslmed.3001456

[14] D. Gotz and D. Borland. 2016. Data-driven healthcare: Challenges and opportunities for interactive visualization. *IEEE Comput. Graph. Appl.* 36, 3 (May 2016), 90–96. DOI: http://dx.doi.org/10.1109/MCG.2016.59

[15] D. Gotz and H. Stavropoulos. 2014. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Trans. Vis. Comput. Graph.* 20, 12 (2014), 1783–1792. DOI: http://dx.doi.org/10.1109/TVCG.2014.2346682

[16] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI'16)*. ACM, New York, NY, 85–95. DOI: http://dx.doi.org/10.1145/2856767.2856779

[17] David Gotz and Zhen Wen. 2009. Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 315–324. DOI: http://dx.doi.org/10.1145/1502650.1502695

[18] D. Gotz and M. X. Zhou. 2008. Characterizing users' visual analytic activity for insight provenance. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2008 (VAST'08)*. 123–130. DOI: http://dx.doi.org/10.1109/VAST.2008.4677365

[19] Julian Heinrich and Daniel Weiskopf. 2012. State of the art of parallel coordinates. In *Eurographics 2013—State of the Art Reports*, M. Sbertand L. Szirmay-Kalos (Eds.). The Eurographics Association. DOI: http://dx.doi.org/10.2312/conf/EG2013/stars/095-116

[20] Stacie Hibino and Elke A. Rundensteiner. 1997. User interface evaluation of a direct manipulation temporal visual query language. In *Proceedings of the 5th ACM International Conference on Multimedia*. ACM, New York, NY, 99–107. DOI: http://dx.doi.org/10.1145/266180.266342

[21] Sally Hopewell, Allison Hirst, Gary S. Collins, Sue Mallett, Ly-Mee Yu, and Douglas G. Altman. 2011. Reporting of participant flow diagrams in published reports of randomized trials. *Trials* 12 (Dec. 2011), 253. DOI: http://dx.doi.org/10.1186/1745-6215-12-253

[22] Alfred Inselberg and Bernard Dimsdale. 1990. *Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry*. IEEE Computer Society Press, Los Alamitos, CA, 361–378.

[23] Institute of Medicine. 2012. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Technical Report. Retrieved from http://iom.nationalacademies.org/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx.

[24] T. J. Jankun-Kelly, Kwan Liu Ma, and Michael Gertz. 2002. A model for the visualization exploration process. In *Proceedings of the IEEE Conference on Visualization*. Washington, DC, 323–330. http://dl.acm.org/citation.cfm?id=602099.602149

[25] M. Kreuseler, T. Nocke, and H. Schumann. 2004. A history mechanism for visual data mining. In *Proceedings of the IEEE Symposium on Information Visualization*. 49–56. DOI: http://dx.doi.org/10.1109/INFVIS.2004.2

[26] Lineberger. 2014. UNC Lineberger Comprehensive Cancer Center. Integrated Cancer Information and Surveillance System. Retrieved from http://iciss.unc.edu/.

[27] Jie Lu, Zhen Wen, Shimei Pan, and Jennifer Lai. 2011. Analytic trails: Supporting provenance, collaboration, and reuse for visual data analysis by business users. In *Proceedings of the Human-Computer Interaction (INTERACT'11)*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Volume 6949 in Lecture Notes in Computer Science. Springer, Berlin, 256–273. http://link.springer.com/chapter/10.1007/978-3-642-23768-3_22

[28] Travis B. Murdoch and Allan S. Detsky. 2013. The inevitable application of big data to health care. *J. Am. Med. Assoc.* 309, 13 (Apr. 2013), 1351–1352. DOI: http://dx.doi.org/10.1001/jama.2013.393

[29] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. 2011. Analytic provenance: Process+interaction+insight. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 33–36. DOI: http://dx.doi.org/10.1145/1979742.1979570

[30] PhysioNet. 2016. MIMIC II: Clinical Database Overview. Retrieved from http://physionet.org/mimic2/mimic2_clinical_overview.shtml.

[31] David Pollard. 2002. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

[32] Alexander Rind. 2013. Interactive information visualization to explore and query electronic health records. *Found. Trends Hum.-Comput. Interact.* 5, 3 (2013), 207–298. DOI: http://dx.doi.org/10.1561/1100000039

[33] Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Crit. Care Med.* 39, 5 (May 2011), 952–960.

[34] Jinwook Seo and Ben Shneiderman. 2002. Interactively exploring hierarchical clustering results. *IEEE Comput.* 35 (2002), 80–86.

[35] Ben Shneiderman. 1994. Dynamic queries for visual information seeking. *IEEE Softw.* 11, 6 (Nov. 1994), 70–77. DOI: http://dx.doi.org/10.1109/52.329404

[36] Ben Shneiderman and Catherine Plaisant. 2006. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*. ACM, New York, NY, 1–7. DOI: http://dx.doi.org/10.1145/1168149.1168158

[37] Y. B. Shrinivasan, D. Gotz, and Jie Lu. 2009. Connecting the dots in visual analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2009 (VAST'09)*. 123–130. DOI: http://dx.doi.org/10.1109/VAST.2009.5333023

[38] Yedendra Babu Shrinivasan and David Gotz. 2009. Connecting the dots with related notes. In *CHI '09 Proceedings of the Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 3649–3654. DOI: http://dx.doi.org/10.1145/1520340.1520549

[39] Yedendra Babu Shrinivasan and Jarke J. van Wijk. 2008. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1237–1246. DOI: http://dx.doi.org/10.1145/1357054.1357247

[40] Douglas G. Simpson. 1987. Minimum hellinger distance estimation for the analysis of count data. *J. Am. Stat. Assoc.* 82, 399 (Sept. 1987), 802–807. DOI: http://dx.doi.org/10.2307/2288789

[41] James Thomas and Kristin Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.

[42] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. 2000. Guidelines for using multiple views in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM, New York, NY, 110–119. DOI: http://dx.doi.org/10.1145/345513.345271

[43] Matthew Ward, Georges Grinstein, and Daniel Keim. 2010. *Interactive Data Visualization: Foundations, Techniques, and Applications* (1 ed.). A K Peters/CRC Press, Natick, MA.

[44] Zhiyuan Zhang, David Gotz, and Adam Perer. 2015. Iterative cohort analysis and exploration. *Information Visualization* 14, 4 (2015). DOI: http://dx.doi.org/10.1177/1473871614526077

Received June 2016; revised February 2017; accepted March 2017