TrialCompass: Visual Analytics for Enhancing the Eligibility Criteria Design of Clinical Trials

Rui Sheng, Xingbo Wang, Jiachen Wang, Xiaofu Jin, Zhonghua Sheng, Zhenxing Xu, Suraj Rajendran, Huamin Qu, and Fei Wang

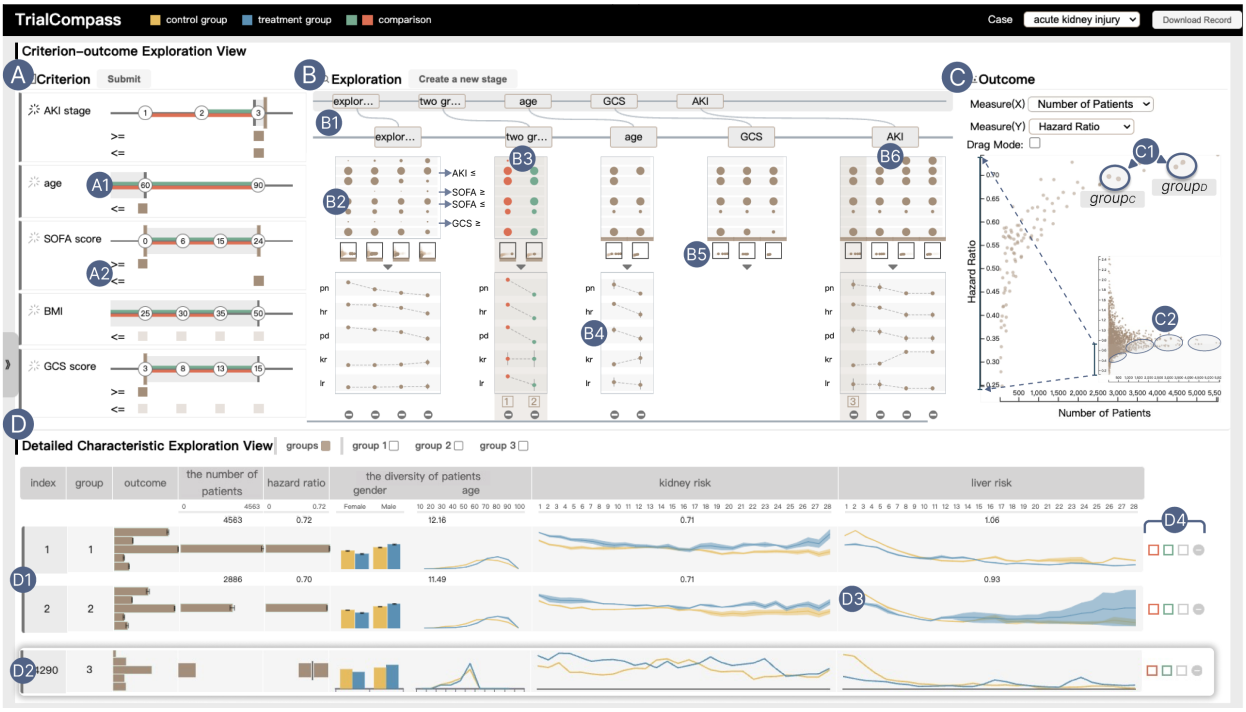

Fig. 1: *TrialCompass* enables clinicians to explore the vast space of eligibility criteria for clinical trials (approximately five thousand criterion candidates in this case). (A) Clinicians can adjust eligibility criteria based on their expertise or preferences in the Criterion View. (B) They can analyze the relationship between eligibility criteria and outcome metrics by creating stages throughout the exploration process, with history recorded for both. (C) Clinicians can explore the outcome metrics of all the criterion candidates in the Outcome View. (D) Clinicians can examine the detailed characteristics of the original EHR data for either individual candidates within a group or the summarization of a candidate group in the Detailed Characteristic Exploration View.

Abstract— Eligibility criteria play a critical role in clinical trials by determining the target patient population, which significantly influences the outcomes of medical interventions. However, current approaches for designing eligibility criteria have limitations to support interactive exploration of the large space of eligibility criteria. They also ignore incorporating detailed characteristics from the original electronic health record (EHR) data for criteria refinement. To address these limitations, we proposed *TrialCompass*, a visual analytics system integrating a novel workflow, which can empower clinicians to iteratively explore the vast space of eligibility criteria through knowledge-driven and outcome-driven approaches. *TrialCompass* supports history-tracking to help clinicians trace the evolution of their adjustments and decisions when exploring various forms of data (i.e., eligibility criteria, outcome metrics, and detailed characteristics of original EHR data) through these two approaches. This feature can help clinicians comprehend the impact of eligibility criteria on outcome metrics and patient characteristics, which facilitates systematic refinement of eligibility criteria. Using a real-world dataset, we demonstrated the effectiveness of *TrialCompass* in providing insights into designing eligibility criteria for septic shock and sepsis-associated acute kidney injury. We also discussed the research prospects of applying visual analytics to clinical trials.

Index Terms—Visual Analytics, Healthcare, Clinical Trials, Decision Making, Electronic Health Record (EHR)

1 Introduction

Clinical trials are studies on human subjects that assess the safety and effectiveness of new medical interventions (e.g., vaccines or drug usage), significantly advancing medicine. Completing a clinical trial is costly, often requiring around $2.87 billion [5]. Unfortunately, 86% of clinical trials fail in the initial step due to the inability to recruit suitable

- *R. Sheng, X. Jin, Z. Sheng, and H. Qu are with the Hong Kong University of Science and Technology. E-mail: {rshengac, xjinao, szh}@connect.ust.hk, huamin@cse.ust.hk.*
- *Z. Xu, S. Rajendran, and F. Wang is with Cornell University. E-mail: {zhx2005, sur4002, few2001}@med.cornell.edu.*
- *X. Wang is with Bosch. E-mail: wangxbzb@gmail.com.*
- *X. Wang and F. Wang are the co-corresponding authors.*

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

participants within the specified timeframe [25]. A major factor is the lack of well-designed **eligibility criteria**, including inclusion and exclusion criteria, to determine who is eligible to participate in the study [38]. Designing effective eligibility criteria is a challenging task. Stricter criteria may hinder enrollment, while more relaxed ones can increase risks of adverse outcomes (e.g., worsening of the illness or even death) for target interventions. Data-driven approaches [16, 29, 31, 38] have been proposed to help clinicians design more inclusive, safe, and effective criteria. For example, Trial Pathfinder [38] can simulate the impact of adjusting a criterion (e.g., changing the age requirement from under 60 to under 70) on potential trial outcomes. This information offers valuable insights for designing eligibility criteria.

Despite the availability of these tools, clinicians still encounter difficulties in designing suitable eligibility criteria. First, existing tools cannot help clinicians efficiently **explore the large space of potential criterion combinations**—that is, the extensive set of possibilities generated by adjusting various eligibility criteria. In practice, clinicians must assess how different combinations of criteria—such as age limits, medical history, and drug dosages—affect patient selection and outcomes. Each individual criterion often needs to be tested across multiple plausible settings based on clinical expertise. For example, age may be set to under 60, 65, or 70 years, while medical history may specify no surgery in the past one, three, or six months. This results in a combinatorial explosion of criterion candidates that are hard to explore. Second, existing tools fail to **provide contextual information on the temporal changes in patients' conditions** during the trial (e.g., whether liver risk consistently increases over time). They only help clinicians trade off different aggregated metrics of trial outcomes, such as the number of patients that can be recruited versus the final trial hazard ratio. However, clinicians still need to consider finer details. For instance, clinicians might weigh the final hazard ratio against changes in liver risk during the trial. Even with a low hazard ratio, they may still reject the eligibility criteria if they observe a persistent increase in liver risk. Third, the lack of support to **track the exploration history** in eligibility criteria and outcome metrics throughout the iterative exploration process poses a significant challenge. With numerous exploration operations, the iterative design process can quickly become highly complex. Therefore, this tracking is crucial for iterative design processes. Without clear records of how changes in criteria impact trial outcomes, clinicians cannot systematically refine them and confidently determine the final settings based on the explored candidates.

To address the above challenges, we developed a visual analytics system named *TrialCompass* to assist clinicians in designing eligibility criteria. To the best of our knowledge, we are the first to leverage visualization techniques to address such a significant problem in the healthcare domain. We first interviewed five experts to derive design requirements and developed a novel visual analytics workflow that enables clinicians to explore the expansive design space of eligibility criteria iteratively. Given that experts may employ their expertise to varying degrees during the exploration, this workflow provides two approaches: knowledge-driven and outcome-driven. The knowledge-driven approach enables experts to simulate outcomes by specifying different eligibility criteria based on their expertise. On the other hand, the outcome-driven approach supports experts in first examining a large number of criterion candidates to make an informed decision. These two approaches offer clinicians flexibility in iteratively refining the eligibility criteria. We have also integrated a history-tracking feature, allowing clinicians to trace their exploration process and understand the relationships between eligibility criteria, outcome metrics, and temporal details. In summary, we made the following contributions:

- We formulate the system design requirements for eligibility criteria design of clinical trials through collaboration with five experts in various specializations.
- We propose *TrialCompass*, a system integrating a novel workflow for iteratively exploring the large space of eligibility criteria through knowledge-driven and outcome-driven approaches.
- We utilize a real-world dataset to conduct expert interviews and case studies, discovering novel insights for two important diseases (i.e., septic shock and sepsis-associated acute kidney injury).

2 Related Work

2.1 Clinical trial design studies

Clinical trials are crucial for assessing the safety and efficacy of new medical treatments. Designing suitable eligibility criteria is vital for trial success, impacting participant recruitment and final results [11]. These criteria specify the conditions that participants should meet to be eligible for a clinical trial, often including factors such as age, gender, medical history, and current health status. However, designing criteria is a challenging task for clinicians since subtle adjustments in criteria may lead to big differences. For example, strict ones previously led to 80% of advanced non-small-cell lung cancer patients being excluded from trials, significantly contributing to an 86% deficit in meeting recruitment goals [17, 25]. Researchers have thus explored data-driven approaches to support designing eligibility criteria more inclusively and effectively [31, 38]. Specifically, these approaches utilize historical patient data to measure the potential outcomes of specified criteria, such as the number of eligible participants for recruitment and their efficacy. For example, Trial Pathfinder [38] measures the number of eligible patients and the hazard ratio of the setting criteria. Moreover, it reveals a criterion's impact on a trial's outcomes. However, current tools mainly rely on clinicians' expertise to iteratively generate hypotheses, without fully leveraging precomputed outcome variations under different eligibility criteria settings. In this trial-and-error process, the lack of effective visual support often causes clinicians to lose contextual understanding and makes it difficult to assess how individual criteria or their combinations impact outcomes, hindering rigorous and informed decision-making. Moreover, previous approaches fall short in supporting the comparison and balancing of multiple objectives (e.g., hazard ratios and patient counts). When conflicts arise, clinicians must manually make trade-offs, sometimes even consulting the original data, which is time-consuming. Adding to the complexity, desirable outcomes often vary depending on clinicians' goals. Their thinking may also evolve during the exploration process, making it difficult for automated algorithms to adapt effectively. These challenges highlight the need for human-in-the-loop tools that combine computational power with appropriate visualization techniques.

2.2 Visual analytics for clinical data

The deployment of visualizations in clinical research is becoming essential [12, 15, 24, 36, 46, 48, 53]. For example, Wang et al. [48] summarized visualization techniques for EHR data, which is one of the most significant and common data formats in the clinical domain. These visualization methods can help analyze the inherent complexity of clinical data and support critical decision-making in this high-stakes field. In this context, AI is increasingly integrated into clinical workflows, and visualization often serves as a bridge between AI and clinicians by addressing issues of uncertainty and interpretability [10, 26, 33]. For example, Cheng et al. [10] proposed using visualization to help clinicians link original data with AI-generated features for improved decision-making. In addition to AI, traditional statistical models remain widely used and are often more thoroughly validated in clinical practice. However, visualization is still essential for helping clinicians understand the large volumes of data these models produce [30, 32, 49]. For example, DPVis [32] employs Hidden Markov Models to calculate disease progression pathways, using visualization to provide clinicians with a more intuitive understanding of disease dynamics. Despite the evident benefits of visualization tools in enhancing various clinical workflows, their integration into clinical trial design remains limited. Although a recent study [34] takes an initial step, it just visualizes patients' temporal progression during the treatment. It cannot be applied to complex decision-making scenarios like eligibility criteria design. The application of visualizations has great potential to facilitate exploring the huge design space of eligibility criteria.

2.3 Visual analytics for multi-objective decision making

Our work aims to assist clinicians in refining eligibility criteria considering multiple objectives, which is a multi-objective decision-making problem. In the visualization community, considerable visual analytic

technologies have been explored to address multi-objective decision-making problems [8, 9, 21, 37, 40, 47, 50, 51, 54]. A key aspect of these tools is the ability to help users narrow down their choices from an immense selection of options. One of the strategies is to enable users to define their constraints and preferences. This approach allows users to see the effects of their constraints and filter the options accordingly [9,51]. However, it is challenging for experts to define constraints when designing eligibility criteria in a clinical trial context. Another strategy focuses on the discovery of user preferences through a process of heuristic exploration [21, 40, 50, 54]. This technique is particularly useful when users are unable to explicitly articulate their constraints or preferences. However, those works cannot support experts in systematically exploring the impact of various eligibility criteria on different outcome forms (i.e., multiple outcome metrics and original EHR data insights) and make informed decisions. Therefore, it is urgent to develop a new visual analytics workflow that helps clinicians navigate through the large space of eligibility criteria.

3 Design Study

We developed a system to help clinicians design eligibility criteria. Over the past six months, we have collaborated closely with five domain experts (E_a-E_e) with various clinical specialties. E_a is a urologist with over twenty years of clinical trial experience. E_b and E_c are nephrologists with approximately three years of experience. E_d is a professor who has dedicated five years to data-driven clinical trial design. Lastly, E_e, a doctoral student who specializes in ophthalmology with two years of expertise. We conducted one-hour semi-structured interviews with each expert to understand the eligibility criteria design challenges. From their requirements, we derived visual analysis tasks, then validated these through bi-weekly prototype feedback sessions. This study received IRB approval.

3.1 Factors Related to Eligibility Criteria Design

Data-driven design leverages historical patient EHR data to help define and refine eligibility criteria. For example, when testing a new drug, clinicians can explore past patient records to identify individuals who have taken similar compounds, using them as a reference for eligibility adjustments. Specifically, clinicians often first determine *eligibility criteria* to filter qualified patients based on their *original EHR data* and assign *medical interventions* to categorize them into *treatment and control groups*. Then, several *outcome metrics*, such as the hazard ratio and kidney risk ratio, can be calculated based on the original EHR data from the patients in the treatment and control groups. The *temporal details* extracted from EHR data aid in interpreting these outcome metrics and understanding the more nuanced results of the intervention. These factors are crucial to the design of eligibility criteria, ensuring that its results are reliable, valid, and applicable to the intended patient population. We have introduced those factors in detail as follows.

1) Eligibility criteria are the conditions designed by clinicians to recruit participants for clinical trials, categorized into inclusion and exclusion criteria [38]. Inclusion criteria define participant eligibility for a clinical trial, whereas exclusion criteria identify disqualifying traits. Eligibility criteria may restrict participant demographics (e.g., age being between 18 and 70), health status (e.g., specific disease diagnoses, or recent medication history), and other relevant variables. We define each unique combination of criteria with specific settings as a **criterion candidate** (e.g., age under 70, no heart surgery in the past three months, BMI under 30).

2) Treatment and control groups are differentiated by whether the enrolled participants receive the **medical intervention** being studied. The treatment group is given the medical intervention, while the control group receives a standard intervention (e.g., placebo) or no intervention [1]. Then clinicians will compare these two groups to assess the effectiveness and safety of the medical intervention.

3) Outcome metrics assess the potential results in a clinical trial from systematic analysis of the treatment and control groups [18]. Once clinicians establish their eligibility criteria, they can filter out qualified patients from historical patient datasets. Then they identify a medical intervention that is either identical to or a suitable proxy for the current medical intervention being studied. This intervention allows them to categorize the filtered patients into treatment and control groups. Next, they measure various outcome metrics based on the two groups. Through the literature survey, two outcome metrics (i.e., the number of patients and hazard ratio) are always seen as the primary focus of interest in clinical trial studies and calculated in data-driven approaches [31, 38]. We have introduced them in detail as follows.

- **The number of patients** refers to the total size of participants qualified for the clinical trial. Recruiting sufficient participants is critical to determine whether a clinical trial can proceed [13]. All five experts underscored the importance of the number of patients.
- **Hazard ratio** refers to the ratio of hazard rates between the treatment and control groups, where a value less than one indicates a positive effect for the treatment group [44]. This metric is often a direct indicator of the effectiveness of a medical intervention [4, 31, 38, 39]. Additionally, it is typically reported with a p-value to assess statistical significance. Its clinical acceptability varies by context—for instance, in the treatment of rare diseases, even a hazard ratio slightly below 1 may be considered meaningful.

We then conducted expert interviews to identify additional metrics of interest. First, several experts suggested examining the overall diversity of the recruited participants. Moreover, since most clinical trials focus on drug-based experiments, the experts were particularly concerned about the impact of drug usage on patients' organ function. They suggested evaluating kidney and liver function since the kidney is the primary organ responsible for the excretion of most drugs [3] and the liver also plays a crucial role in drug metabolism [35]. Specifically, they would like to understand the kidney and liver risks of patients over time. Additionally, the Charlson Comorbidity Index was mentioned to indicate patient mortality risk. However, due to a lack of relevant data on most patients, we decided not to incorporate this index. Below, we outline the additional outcome metrics used in our study.

- **The diversity of patients** refers to the demographic breadth represented by participants who fulfill the eligibility criteria, encompassing a variety of attributes such as age, gender, race, and others. Recruiting a patient population with diverse demographic characteristics can lead to more broadly applicable results [28]. A greater value of diversity indicates a larger variety within the studied patient population.
- **Kidney risk ratio** and **liver risk ratio** measure the incidence of adverse events related to the kidney and liver between the treatment and control groups, respectively. A value less than one suggests that the treatment group has a lower risk of experiencing adverse events. E_a, E_b, and E_c highlighted the necessity of evaluating the potential adverse reactions in survivors in the two groups. E_a mentioned that the hazard ratio usually reflects the survival rate difference between the treatment and control groups, which is a primary concern of a clinical trial. However, it is also important to comprehend the health condition of those surviving patients. The kidney and liver are the two most critical concerns.

Clinicians must balance the five outcome metrics when defining eligibility criteria for clinical trials, as these metrics can at times present conflicting priorities. For instance, relaxing the eligibility criteria may increase participant enrollment, yet also result in a higher hazard ratio. As such, clinicians need to carefully weigh the tradeoffs among the five outcome metrics during the eligibility criteria design process.

4) Temporal details are derived from the original EHR data. The five outcome metrics are aggregation values calculated through patient cohorts. Therefore, clinicians also need to examine the temporal detailed characteristics in those patients' conditions to gain a more comprehensive understanding to compare different criterion candidates. For example, E_c highlighted the importance of tracking changes in kidney and liver function over time, as the medical intervention may have varying onsets across participants.

3.2 Visual Analysis Tasks

Based on our interviews with five experts, we have abstracted this domain problem into a multi-objective decision-making problem. Then,

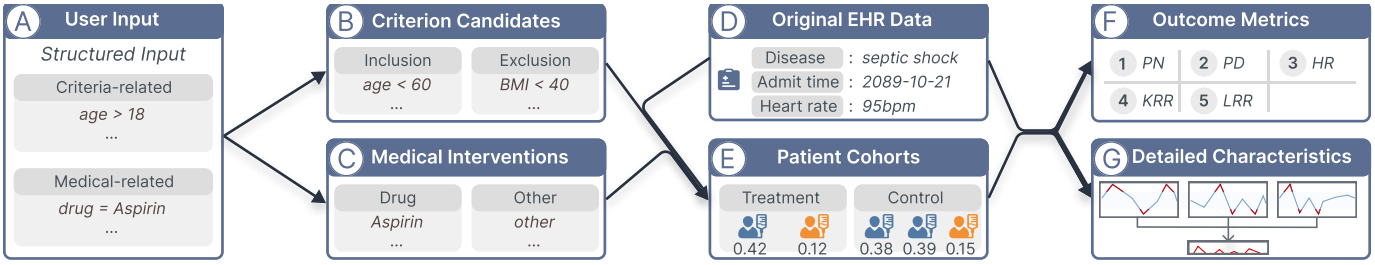

Fig. 2: (A) Clinicians input eligibility criteria and medical interventions. (B) Different criteria specifications can generate various criterion candidates. (C) The medical interventions will be used to divide patients into different groups. (D) The original EHR data of patients. (E) The system filters qualified patients and categorizes them into treatment and control groups based on eligibility criteria, medical interventions, and their original EHR data. (F) Measure the outcome metrics. (G) Organize the temporal detailed characteristics of the original EHR data.

two authors performed independent inductive coding of interview transcripts using the thematic analysis methodology [7]. Through iterative coding and regular discussions, we reconciled interpretations and collectively distilled five key analytical tasks through consensus.

T1 Support eligibility criterion specification for candidate generation. We need to enable experts to make initial settings for eligibility criteria. Moreover, experts should be able to specify which criteria are adjustable based on their requirements and the range of possible adjustments for each criterion. Following their settings, we can generate a series of criterion candidates.

T2 Present the outcomes of criterion candidates. Some experts (E_a, E_b, and E_e) emphasized the need to efficiently evaluate criteria configurations. As E_b noted, *"Manually testing different criteria combinations is prohibitively time-consuming."* E_a added, *"Seeing potential outcomes of various criteria candidates in advance can help me form testable hypotheses."* Therefore, our system needs to precompute and visually present all five outcome metrics for each candidate, enabling rapid comparative assessment.

T3 Support knowledge-driven and outcome-driven exploration. All five experts emphasized the need to integrate their expertise when evaluating criterion candidates. For instance, E_a explained, *"For several criteria, I have prior expectations from literature or other clinical trials. The system should let me quickly validate these hypotheses."* Experts also highlighted the benefits of using computed outcome metrics to address knowledge gaps. E_e mentioned that insufficient experience could be mitigated by leveraging big data to enhance exploration efficiency, a sentiment echoed by E_b. Consequently, our system should offer two exploration approaches: one that taps into clinicians' prior knowledge (knowledge-driven) and another that relies on measured outcomes (outcome-driven).

T4 Incorporate outcome metrics and temporal details for comparison. E_a, E_b, and E_c highlighted the need to examine details of the original EHR data. For instance, E_a mentioned, *"We also need to track organ function trajectories—not just snapshot values. If kidney function declines steadily post-treatment, this signals intolerable toxicity even if the aggregate risk ratio appears acceptable."* Therefore, it is crucial to facilitate the comparison of outcome metrics with temporal details.

T5 Facilitate the iterative navigation of criterion candidates. After exploring outcome metrics and temporal details of original EHR data, experts need to gain insights that could lead to further refinement of eligibility criteria. Given the extensive exploration history generated, experts E_a, E_c, and E_d emphasized the importance of systematically tracking and organizing this information.

4 Data Analysis

4.1 Data Description

In this work, we utilized the MIMIC-IV dataset [27] as a historical record of patient data to study eligibility criteria design. The MIMIC-IV dataset is a publicly available database that provides comprehensive clinical data from intensive care units (ICUs). It also includes de-identified electronic health records (EHR) of patients. Specifically, it contains detailed information on over a hundred thousand patients, such as specific diagnoses, observed values of physiological indicators, and medication history. This information allows us to ensure a sufficient sample size when studying different clinical trials.

4.2 Data Processing

We first use the eligibility criteria and specific medical intervention entered by experts to separate patients into treatment and control groups (Fig. 2-A-E). In this process, adjustments to the criteria will generate multiple criterion candidates and result in different patient compositions of the two groups. Then, our system can calculate the outcome metrics for each criterion candidate (Fig. 2-F). Finally, the temporal details of the treatment and control groups derived from the original EHR data will be systematically compiled and organized (Fig. 2-G).

4.2.1 Patient Cohort Construction

First, we can identify eligible patients based on structured eligibility criteria entered by clinicians and classify them into treatment and control groups (Fig. 2-E). When comparing the treatment and control groups, it is crucial to address the inherent differences in confounding factors, which are unrelated to the factors being studied. These confounding factors can introduce bias and affect the interpretation of the results. Therefore, we performed a matching process for the two groups under each criterion candidate. To achieve this, we leveraged the propensity score matching algorithm, like Trail Pathfinder [38], to reduce bias caused by confounding factors (e.g., race, gender, and birthplace). The propensity score is defined as the conditional probability of receiving the medical intervention given a set of observed confounding factors [43]. Specifically, this metric is used to identify matching patients within the control group to correspond with those in the treatment group, ensuring that their characteristics are comparable and allowing for a more accurate estimation of the treatment effect, which can be measured as follows:

$$e(F) = P(T = 1 \mid F),$$

where $e(F)$ represents the propensity score, T denotes the medical intervention assignment (1 for treatment, 0 for control), and F represents the confounding factors. Then, we iterated through each patient in the treatment group and identified the most similar patients in the control group based on their propensity scores (Fig. 2-E). We then compared the propensity score difference between the matched pairs with a specified caliper value. The caliper value, often set as the median absolute deviation of the propensity scores [2], serves as a threshold for acceptable similarity. If the difference in propensity scores is not greater than the caliper value, the pair is considered a match. Or we discard this particular sample. Finally, these pairs will be used for subsequent comparison between the treatment and control groups.

4.2.2 Outcome Metric Calculation

We calculated five outcome metrics to assess the potential results of each criterion candidate (Fig. 2-F). **The number of patients** is the count of qualified patients. **The diversity of patients** is calculated based on gender and age, which are commonly mentioned in data-driven eligibility criteria design [38]. We calculated the Shannon entropy based on the two attributes to represent the diversity of patients. We did not choose other diversity metrics, such as the Gini coefficient and

Simpson index. This is because these metrics have lower values when indicating higher diversity, which contradicts experts' intuition.

Then, we calculated the **hazard ratio** based on the treatment and control groups through training the Cox proportional-hazards model [45], which is a commonly used method for survival analysis. Specifically, the model is formulated as follows:

$$h(t|X) = h_0(t) \cdot e^{(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \beta_T T)},$$

where $h(t|X)$ represents the hazard function at time t given the covariates X, $h_0(t)$ represents the baseline hazard function, and $\beta_1, \beta_2, \ldots, \beta_p$ correspond to the regression coefficients associated with each covariate. Moreover, T indicates the medical intervention assignment. Finally, the hazard ratio can be represented by $HR = e^{(\beta_T)}$.

For the **kidney risk ratio and liver risk ratio**, we used serum creatinine (SCr) [6] and aspartate aminotransferase (AST) [42] as indicators, respectively. We extracted them from the original EHR dataset. Due to potential truncation, some patients may have missing or incomplete data. We addressed this by imputing values based on discharge status: for discharged patients, we replaced missing data with normal values, assuming recovery. For patients with missing data before discharge or death, we discarded samples with significant gaps and applied interpolation for the rest [14]. We then calculated daily kidney and liver risk ratios, averaging them to provide overall assessments for each.

4.2.3 Temporal Detail Organization

Our experts (E_a-E_e) pointed out that it is challenging to understand the demographic characteristics of the recruited patients only through the diversity score. Therefore, we compiled the distribution of gender and age for both the treatment and control groups. In addition, the hazard ratio needs to be accompanied by a confidence interval to allow clinicians to understand the significance of the results. Furthermore, considering that the kidney and liver risk ratios represent an average of the risk over a period, clinicians need to examine how these risks evolve over time in the two groups. Therefore, we summarized the risk degree for the organs over time (Fig. 2-G). For patients who were alive at a specific time, we calculated the average degree of abnormality in their indicators—such as SCr for kidney risk and AST for liver risk—for both the treatment and control groups, based on reference ranges.

5 Visual Design

TrialCompass provides three views to support clinicians in designing eligibility criteria: the Criterion Specification View (Fig. 3), the Criterion-outcome Exploration View (Fig. 1-A, B, C), and the Detailed Characteristic Exploration View (Fig. 1-D).

5.1 Criterion Specification View

This view facilitates clinicians in entering eligibility criteria for enrollment and defining the particular medical intervention that differentiates between treatment and control groups (**T1**). Clinicians can choose to create criteria for medical interventions, inclusion criteria, and exclusion criteria (Fig. 3-A, B). The system supports the use of "AND" and "OR" within a single criterion, as well as several aggregation functions like the minimum and maximum (Fig. 3-C). It also supports the combination of multiple criteria, such as "at least two eligibility criteria must be met" (Fig. 3-D). In addition, our system provides clinicians with user-friendly prompts, like displaying detailed explanations when hovering over an entity in a drop-down list. Finally, our system allows clinicians to customize uncertain criteria that they would like to adjust. They can click the corresponding button to set multiple adjustment values (Fig. 3- C1). By combining these adjustments for each criterion, our system can generate criterion candidates, evaluate their outcome metrics, and organize the temporal details of the original EHR data.

5.2 Criterion-outcome Exploration View

This view allows clinicians to systematically navigate through a comprehensive array of criterion candidates along with their associated outcome metrics. There are three sub-views: the Criterion View, the Outcome View, and the Exploration View.

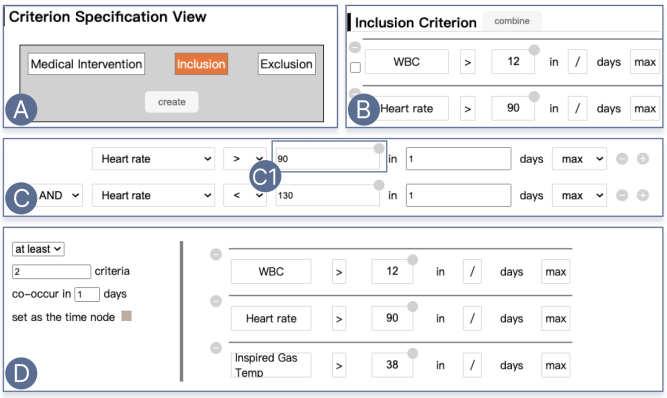

Fig. 3: (A-D) The Criterion Specification View enables clinicians to specify inclusion criteria, exclusion criteria, and medical interventions.

5.2.1 Criterion View

This view (Fig. 1-A) enables experts to adjust each eligibility criterion through sliders based on their prior knowledge or insights during the exploration process (**T3**).

Description: This view is organized sequentially, listing each eligibility criterion that experts need to adjust. A slider is provided for each criterion to allow clinicians to make adjustments. If clinicians initially specify a maximum and minimum constraint for a criterion, two sliders will appear to represent these limits. Additionally, to aid experts in understanding the range, the region of the effective range is highlighted. Beneath each tick mark on the slider, color-coded visualizations indicate how many criterion candidates satisfy the corresponding constraint (Fig. 1-A2). They can interact with the scatter plot in the Outcome View to show the distribution of the selected candidates. Considering that clinicians often need to compare two candidates, two colors are used above and below the slider to display the distinctions between the criteria of two candidates selected in the other views (Fig. 1-A1).

Justification: Originally, we utilized a set of polylines along parallel sliders to display the combinations of various eligibility criteria (Fig. 4-A). However, this approach introduced visual confusion. Interleaved polylines made it difficult for experts to discern differences between two candidates. Furthermore, in this situation, the upper and lower limit requirements of a criterion will be separately displayed, which does not align with the customary practices of experts.

5.2.2 Outcome View

This view (Fig. 1-C) reveals the values of outcome metrics associated with all potential criterion candidates using a scatter plot (**T2**). Experts can select any two of the five metrics to represent on the axes. We avoided using glyphs to prevent visual clutter and potential interpretation difficulties. Instead, experts can set two axes to represent each criterion candidate as a point in the scatter plot, allowing for zooming during exploration. They can also lasso interesting candidates, which will be shown in the Exploration View for further analysis (**T3**).

5.2.3 Exploration View

This view (Fig. 1-B) enables clinicians to conduct systematic exploration. It allows experts to navigate through the vast space of eligibility criteria based on stages. In each stage, the exploration process is recorded through a snapshot, assisting clinicians in tracking and analyzing their exploration path (**T5**). This feature helps them to understand and narrow down the complex criterion space effectively.

Description: This view helps experts understand the relationship between criteria and outcomes through their exploration history. Given the complexity of adjusting criteria in the Criteria View and selecting candidates in the Outcome View, systematically recording this operation history is essential for clarity. To achieve this, we introduced stages (Fig. 1-B1). Experts can create a new stage whenever they consider an exploration action to be independent of previous ones. Their exploration will be recorded through a snapshot, where they can assign an importance level, keywords, and detailed descriptions to this stage.

This enables precise documentation and easy review of previous operations. Furthermore, we provide a condensed stage visualization that displays all important stages at the top of the Exploration View.

In each stage snapshot, our system visualizes the exploration history from two main operations: adjusting eligibility criteria in the Criterion View and selecting criterion candidates based on outcome metrics in the Outcome View. First, a matrix displays changes in eligibility criteria (Fig. 1-B2). Typically, each row in a matrix represents a criterion. When a criterion has a minimum and maximum threshold set, two rows will represent them respectively. Columns correspond to exploratory records, and circles indicate values—larger circles represent higher values, with exact figures displayed on hover. Second, we present changes from selecting criterion candidates in the Outcome View using two methods. Thumbnails of the scatter plot provide an intuitive overview of operations (Fig. 1-B5). Considering that the scatter plot in the Outcome View might involve changing the horizontal and vertical axes, we display these axes in the thumbnail if they have been changed. Additionally, line charts illustrate the average values of five outcome metrics during exploration (Fig. 1-B4), helping experts comprehend how different outcome metrics fluctuate throughout their operations.

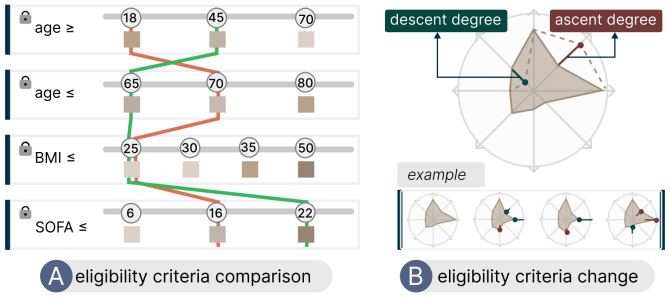

Fig. 4: (A) The alternative design is to compare the eligibility criteria of two individual criterion candidates. (B) The alternative design is to track the changes in multiple eligibility criteria.

Justification: First, we initially leveraged line charts to display the change in eligibility criteria. However, as the number of eligibility criteria increases, each line chart occupies a smaller portion of the space. This diminishes the visual amplitude of variations, making it challenging to discern changes. Then, we used radar charts to represent changes in criteria (Fig. 4-B). However, we discovered that experts encountered difficulty in comparing radar charts to understand the changes in a specific criterion over time. Second, we just recorded how eligibility criteria changed and the five outcome metrics evolved over time with each operation conducted by clinicians at first. However, when faced with a large amount of historical data, experts often struggle to keep track of their previous operations and relevant records during the current exploration. Therefore, we introduced the concept of stages.

5.3 Detailed Characteristic Exploration View

This view (Fig. 1-D) presents temporal detailed characteristics of the original EHR data in two modes: group and individual (**T4**). For the group mode, the visualization presents the average and standard deviation of temporal detailed characteristics for all the criterion candidates in a group (Fig. 1-D1). The individual mode displays each criterion candidate in a group individually (Fig. 1-D2). Specifically, it first displays the index and five outcome metrics. Additionally, histograms indicate the number of patients and hazard ratios. Different colored line charts represent the treatment and control groups for the distributions of gender, age, kidney function over time, and liver risk over time. Line charts are used instead of histograms to reduce clutter and enhance trend analysis. Experts can select two groups in group mode or two candidates in individual mode for comparison (Fig. 1-D4). The criteria for the two selected groups are showcased in the Exploration View (Fig. 1-B3), while the criteria for two individual candidates are displayed in the Criterion View (Fig. 1-A1).

6 Case Study

We have conducted case studies based on the MIMIC-IV dataset [27] for two different diseases (i.e., septic shock and sepsis-associated acute kidney injury). For the first case, we invited a new clinician E_f, who has over five years of experience in clinical trials in sepsis. For the second case, we invited our previous expert E_b, who specializes in kidney disease research and has three years of experience in clinical trials. We thoroughly documented the exploration process of both experts to showcase how our system assists in eligibility criteria design.

6.1 Case I: Septic Shock

E_f is an expert in sepsis research and is very interested in a historical clinical trial[1] investigating the efficacy of ***hydrocortisone*** (a kind of drug) in patients with septic shock. She desired to assess whether the eligibility criteria in this clinical trial could be refined. In addition, she would like to check whether and how to add two more criteria that are commonly seen in other related clinical trials.

Specifying the eligibility criteria (T1). First, E_f established the eligibility criteria and medical intervention based on this clinical trial. Among these, she focused on adjusting two specific criteria. She thought one was slightly relaxed, while another was too restrictive to potentially exclude some patients who could benefit. The first is **age** (the first row in Fig. 5-A), as this clinical trial requires patients above 18 years old. However, she desired to understand how age might impact the efficacy of hydrocortisone. For instance, she was considering whether hydrocortisone might be less effective in older patients. This consideration could lead her to establish an upper age limit for the trial. Secondly, she noticed that it specified the requirement for patients to be on **mechanical ventilation** (the second row in Fig. 5-A), which usually indicates a severe condition. She wondered if hydrocortisone was also suitable for patients with less severe conditions.

Additionally, E_f introduced two new criteria for adjustment. These criteria were not considered in the original trial but are frequently included in other related clinical trials. E_f deemed them important based on her expertise. The first was that patients should not have undergone **cardiac surgery** (the third row in Fig. 5-A) within the past six months. From her expertise, patients who had previously undergone cardiac surgery and subsequently developed sepsis were at a higher risk of developing septic cardiomyopathy, which carried a higher mortality rate. Therefore, she hypothesized that hydrocortisone might not be as effective for these patients. Finally, she desired to assess whether to recruit patients with obesity, as **obesity** (the fourth row in Fig. 5-A) can exacerbate organ damage caused by septic shock. She set the range of those criteria within the Criterion Specification View. Initially, she set the criteria to match the original clinical trial, finding a hazard ratio of 1.225 with a statistically significant confidence interval. The kidney and risk ratios were around 2, indicating high stakes. Despite the large sample size, these metrics led her to refine the eligibility criteria.

Knowledge-driven exploration (T3). Initially, E_f desired to examine whether patients who did not require mechanical ventilation could still benefit from hydrocortisone. Therefore, she explored the impact of the mechanical ventilation requirement by creating a stage in the Exploration View and specifying the slider in the Criterion View. She found that the group without receiving mechanical ventilation (the first column in Fig. 5-B1) showed a deterioration in various outcome metrics. This was indicated by a higher hazard ratio, as well as higher kidney and liver risk ratios (the second, fourth, and fifth line charts in Fig. 5-B1). Upon hovering over the hazard ratio line chart, she noticed that this group particularly had a hazard ratio greater than 1 (1.11). Conversely, the group that received mechanical ventilation (the second column in Fig. 5-B1) had a hazard ratio below 1 (0.97). *"A hazard ratio below 1 indicates a positive effect of the treatment. Therefore, it suggests that hydrocortisone may not have a positive effect on patients without mechanical ventilation."* Then, E_f would like to examine the potential outcomes from interactions between the mechanical ventilation criterion and other criteria, as she was concerned this might lead to changes in the effects. Therefore, E_f decided to analyze all the

[1] https://clinicaltrials.gov/study/NCT01448109

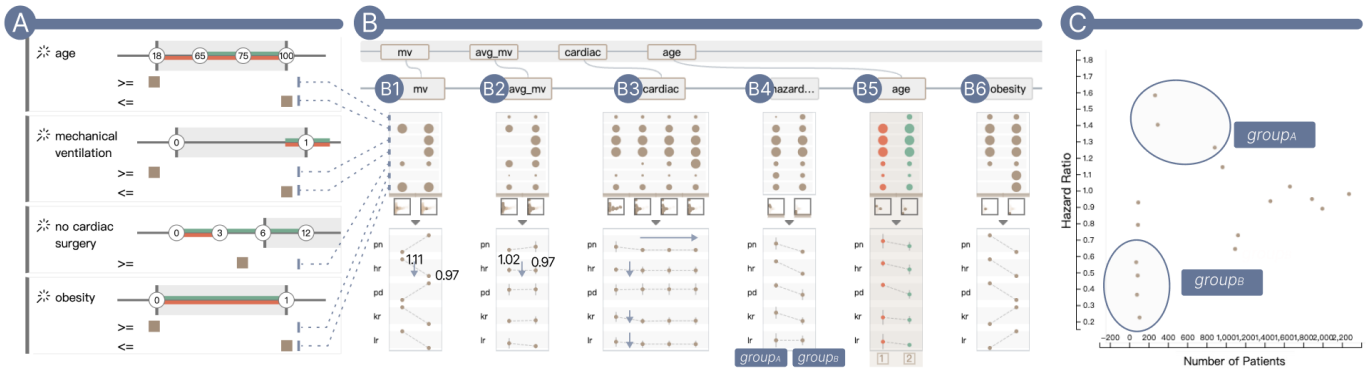

Fig. 5: (A) Specify four eligibility criteria. (B) The exploration process. (B1-B2) Investigate the impact of mechanical ventilation. (B3) Study the impact of cardiac surgery. (B4) Identify the factors (i.e., age and obesity) that can impact the hazard ratio through the outcome-driven approach. (B5) Validate the impact of age. (B6) Validate the impact of obesity. (C) Explore two groups of candidates with hazard ratios greater than or less than 1.

criterion candidates requiring mechanical ventilation alongside those not requiring mechanical ventilation. She found that the average hazard ratio of the former was less than 1 (the second column in Fig. 5-B2, the second line chart). This reinforced her belief in following the historical clinical trial and targeting patients undergoing mechanical ventilation. *"Mechanically ventilated patients often have more severe systemic inflammation. In this context, hydrocortisone seems to be effective at modulating the inflammation and improving their symptoms."*

E_f proceeded to evaluate the requirement of duration after cardiac surgery. She hoped to determine whether the implementation of this criterion could mitigate the hazard ratio and other risk ratios. Therefore, she created a new stage to analyze the criterion candidates with varying time requirements after cardiac surgery. Her comparisons of all five outcomes across different time requirements (indicated in the line charts in Fig. 5-B3) revealed that candidates without a time requirement (the first column in Fig. 5-B3) had a higher average hazard ratio (greater than 1), kidney risk ratio, and liver risk ratio. This aligned with her prior knowledge that patients with a history of cardiac surgery who later developed sepsis faced an elevated risk of septic cardiomyopathy, which in turn carried a higher mortality rate. Furthermore, she noticed that there was almost no difference in the outcome metrics between the time requirements of 3, 6, or 12 months (the last three columns in Fig. 5-B3, the line charts). Therefore, E_f decided to set the inclusion criterion requiring a duration after cardiac surgery greater than 6 months. This decision was predicated on the medical understanding that a 3-month threshold typically denotes potential safety, whereas a 6-month threshold signifies a more fundamental level of safety. By implementing the 6-month benchmark, E_f could ensure a baseline level of safety while still avoiding the exclusion of patients who could potentially benefit.

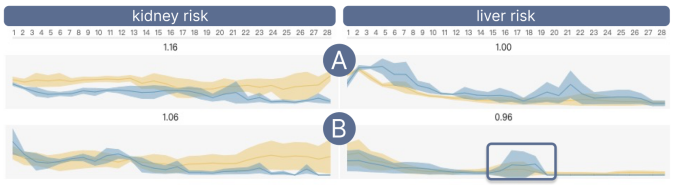

Fig. 6: The temporal details of the two groups. (A) The group with minimum age limits set at 18. (B) The group with minimum age limits set at 65. E_f found that older patients exhibit lower kidney and liver risks, but caution is advised regarding the mid-term increase in liver risk.

Outcome-driven exploration (T3). After finalizing the two criteria, E_f used the Outcome View to examine the factors influencing the hazard ratio. To understand how age and obesity jointly influence the outcome, she performed a lasso selection on the scatter plot based on y-axis values (**T2**) to compare two distinct groups: $group_A$ (hazard ratio > 1) and $group_B$ (hazard ratio < 1) (Fig. 5-C1). She discovered that the criterion candidates in $group_B$ had a higher minimum age requirement and a greater number of patients with obesity (as indicated by the circle size in the first row and the last row of the matrix in Fig. 5-B4). She also checked it through the heatmap under each slider in the Criterion View. *"This indicates that recruiting patients with older age may lead to better treatment outcomes. Furthermore, patients with obesity can also benefit from the treatment."*

Employing a similar approach as before, E_f discovered that patients over the age of 65 (the second column in Fig. 5-B5, the line charts) had better treatment efficacy, indicated by lower hazard ratios (**T5**). She also observed that increasing age had a minimal impact on kidney and liver risk ratios (Fig. 5-B5, the fourth and fifth line charts). *"Although there was minimal change in the kidney or liver risk ratios, it is still important to identify when these patients might experience abnormalities."* Therefore, E_f delved into the Detailed Characteristic Exploration View (**T4**). She discovered that older patients in the treatment group were more likely to experience mid-term liver issues (as indicated by the steep increase in the blue line during the mid-term period in Fig. 6-B). *"This may be due to liver-related side effects that arise after administering a certain amount of hydrocortisone. I find it acceptable since these patients are expected to recover in the later stages."* Finally, E_f made an intriguing observation during her analysis of obesity: patients with obesity (the second column in Fig. 5-B6) indeed displayed a favorable response to the treatment since their hazard ratio was lower. This led her to consider the presence of the obesity paradox. *"Despite the potential adverse effects of obesity on organ function, several studies have indicated that individuals with obesity exhibit lower mortality rates. This phenomenon has been observed in some diagnosis scenarios. It seemed to be also present in this case."* However, she also emphasized the importance of monitoring the proportion of abnormalities in liver and kidney function, as indicated by the last two line charts in Fig. 5-B6. Despite this, she ultimately decided to include patients with obesity, as the hazard ratio was remarkably low.

The final decision. Finally, E_f reviewed all of her explorations again. Using the stage-based visualization, she systematically rechecked the reason behind each decision and then summarized the insights. She determined the following key criteria based on her findings: recruiting patients undergoing mechanical ventilation, setting a minimum time frame of 6 months after cardiac surgery, including patients with obesity, and enrolling as many elderly patients as possible. *"This system has helped me improve potential outcomes compared to the historical eligibility criteria. My key concern—the hazard ratio—has shifted from greater than 1 to less than 1, which is very promising."* She also noted that the kidney and liver risk ratios are below 1, boosting her confidence in her decision-making *"Although patient recruitment has decreased, this likely filters out those who do not respond well to treatment."*

6.2 Case II: Sepsis-associated Acute Kidney Injury

E_b, an experienced nephrologist, was interested in studying the effects of aspirin on sepsis-associated acute kidney injury. He hoped to leverage our system to optimize five eligibility criteria.

Specifying the eligibility criteria (T1). He set the inclusion criteria as patients with sepsis-associated acute kidney injury. He then used aspirin to divide the treatment and control groups. Next, he selected five criteria which were always considered in kidney-related diseases and

he was interested in (Fig. 1-A). The first was the **AKI stage**, categorized into three levels, indicating the severity of kidney dysfunction. The second was **age**. In some clinical trials, older patients are often excluded due to potential organ decline and reduced treatment compliance. The third was the **SOFA score**, reflecting the degree of organ failure and providing insights into the patient's current health condition. The fourth was **BMI**, which is used to evaluate whether an individual is within the healthy weight range. Excessively overweight patients are sometimes excluded from clinical trials for safety reasons. The fifth was the **GCS score**, which is widely used in emergency medicine to assess a patient's level of consciousness. Then, E_b inputted these criteria and manually corrected them. Initially, the AKI stage was required to be over 1, and the age was limited to below 60. The SOFA score had to be less than 15, and no specific requirements were set for the GCS score. Additionally, patients whose BMI was larger than 35 were excluded from the study. These criteria identified approximately 1,000 eligible patients. The analysis yielded a statistically significant hazard ratio of 0.59. Additionally, the kidney risk ratio was below 1, while the liver risk ratio exceeded 1. E_b considered the hazard ratio favorable but deemed the patient sample size insufficient. He hoped to greatly increase the patient enrollment, while still maintaining the low hazard ratio and other favorable outcome metrics.

Outcome-driven exploration (T3). Given the numerous criteria and their interactions, E_b found it difficult to adjust the criteria and examine the outcomes. However, with so many potential candidates, selecting the ones for further examination was also challenging. Therefore, he decided to explore the relationship between the criteria and outcome metrics first to see how to reduce the exploration space. Therefore, E_b initially identified four regions (Fig. 1-C2) on the edges of the scatter plot where the hazard ratio and the number of patients were balanced (**T2**). *"These regions represent a trade-off, where increasing the number of patients can lead to a higher hazard ratio."* From the records in the Exploration View, he observed that the size of the circles remained relatively unchanged in the second, fourth, fifth, and seventh rows of the matrix in Fig. 1-B2, respectively. This indicated that the upper limit for the AKI stage, the lower and upper limit for the SOFA score, and the lower limit for the GCS score were stable in these four regions. Consequently, he established these criteria (i.e., *the AKI stage* ≤ 3, $0 \leq$ *the SOFA score* ≤ 24, and *the GCS score* ≥ 3) since he believed that within this range, he was more likely to find the criterion candidate he was satisfied with. E_b mentioned, *"Patients with higher AKI stages indicate more severe kidney injury, while lower GCS scores suggest more significant brain impairment. This suggests that severely affected patients may benefit from aspirin. Furthermore, the overall impact of the SOFA score seems minimal."*

Next, E_b selected two smaller groups (i.e., $group_C$ and $group_D$) for in-depth analysis (Fig. 1-C1). Therefore, he examined their details in the Detailed Characteristic Exploration View (Fig. 1-D) (**T4**). While $group_C$ exhibited a slightly lower hazard ratio (0.70) compared to $group_D$ (0.72), he noticed that the number of patients decreased by almost half (Fig. 1-D1). Additionally, he noted that the treatment group in $group_C$ exhibited a continuous deterioration of liver function by the end of the experiment (Fig. 1-D3). Therefore, he was inclined to choose $group_D$ for further exploration. Then, E_b examined the two individual criterion candidates in $group_D$. Based on the comparison in the Criterion View, he found that both selected candidates had an upper age limit of 90 and an upper GCS score limit of 15 (Fig. 1-A). To validate this, E_b adjusted the corresponding slider and discovered that there were not sufficient patients aged under 60 (as indicated by the first line chart in Fig. 1-B4) (**T5**). He further confirmed that setting the upper age limit to 90 was a preferable choice. Using the same approach, E_b determined the upper limit of the GCS score.

Knowledge-driven exploration (T3). E_b started to determine the remaining two eligibility criteria: the lower limit of the AKI stage and whether to enroll patients with high BMI. E_b believed that patients with an AKI stage of 1 (indicating a less severe condition) might be more prone to experiencing side effects rather than benefits. As expected, based on the second line chart in Fig. 1-B6, he found that as the lower limit of the AKI stage increased, the hazard ratio slightly decreased. However, he said, *"Although including patients with an AKI stage of 1 increases the hazard ratio, it remains at an acceptable level. Considering the obvious increase in the number of patients and the overall lower kidney risk in this group, it seems reasonable to set the lower limit of the AKI stage as 1."* He applied the same approach to another criterion and concluded that the study should also include patients with a higher BMI.

The final decision. E_b discovered that the clinical trial would include over 5,000 patients after the exploration. Compared to the initial 1,000 patients, this represented a 5-fold increase in the number of patients. In addition, E_b found that the other outcome metrics remained acceptable. There was a slight increase in the hazard ratio from the initial value of 0.59 to 0.72. Although the two values are distant from 1, they are not excessively low, indicating a similar level and suggesting that the drug is effective in the selected population. Furthermore, he found that the kidney risk ratio remained below 1, and there was a notable decrease in the liver risk ratio compared to the initial settings. Overall, E_b expressed excitement about these insights, as they indicate a substantial number of patients without a significant increase in risk.

7 Expert Interview

To further evaluate the effectiveness of our system, we conducted one-on-one interviews via Zoom with five clinicians (P_a-P_e) who had not participated in the design process of *TrialCompass* and had never used our system. These participants were recommended by experts from our earlier formative study, based on our inclusion criteria—namely, having at least three years of experience in clinical trial design and having completed at least one full trial cycle. To ensure objectivity, we did not allow the experts to contact the participants directly; instead, we reached out via email or phone. Additionally, participants came from different hospitals, enhancing the independence and generalizability of their feedback. Among them, three were male and two were female clinicians, with an average age of 41 (ranging from 26 to 44). Their average experience in clinical trials was 9.4 years (ranging from 3 to 21 years). All of them specialized in kidney-related fields, as we intended to use the sepsis-associated acute kidney injury (introduced in Case II in Sec. 6.2), a common and severe illness, as their exploration scenario. As eligibility criteria, outcomes, and detailed characteristics represent fundamental components of eligibility criteria designs, our interview specifically investigated how the Criterion-Outcome Exploration View and Detailed Characteristic Exploration View can facilitate the exploration and decision-making processes. We began the session by providing a 10-minute introduction to the research background. Following that, we demonstrated the system functionality and usage through Case I for 20 minutes. Next, the experts could familiarize themselves with the system for 15 minutes. Then, the experts utilized our system to adjust the eligibility criteria in Case II for 30 minutes. During the process, we asked the experts to think aloud, allowing us to record their audio during the exploration process and facilitate our subsequent analysis. Afterward, we conducted a 12-minute interview to gather the experts' feedback on using the system. Lastly, we invited the experts to complete a questionnaire to rate our system's usability, which took approximately three minutes.

System Workflow. We summarized how our two approaches (i.e., the knowledge-driven approach and the outcome-driven approach) facilitate their decision-making process, respectively.

◇ ***The knowledge-driven approach.*** All the experts have emphasized a key advantage. The knowledge-driven approach allows them to obtain a precise understanding of the outcomes related to eligibility criteria that they are familiar with at first. P_e expressed, *"Clinicians might have a general sense of eligibility criteria, but it may not be specific enough. Therefore, we desire to know its exact outcomes through the knowledge-driven approach, thus guiding further exploration."* Additionally, several experts also mentioned that they could avoid examining too many options in the Outcome View. P_c stated that he could significantly reduce the number of options in the Outcome View once he determined several criteria through his knowledge upfront. Lastly, P_b emphasized the coherence provided by the knowledge-driven approach. By adjusting the criteria with his expertise, he expressed greater confi-

dence in maintaining consistency throughout the optimization process.

◇ ***The outcome-driven approach.*** The most prominent advantage of adopting the outcome-driven approach is quickly exploring and determining multiple criteria. Additionally, clinicians can fill their knowledge gaps regarding complex combinations among the eligibility criteria. Furthermore, the outcome-driven approach can help uncover more optimal candidates. Initially, P_a established two types of eligibility criteria: one for what he perceived as the worst-performing group and the other for the best-performing group. P_a observed a hazard ratio of 0.79 with a sample size of 3541 in the first group, while the second group had a smaller sample size of 512 and a lower hazard ratio of 0.59. This discrepancy led him to consider the possibility of better candidates lying between these two extremes. However, managing multiple criteria proved challenging, even for experienced clinicians. Consequently, he adopted the outcome-driven approach to explore additional points along the spectrum. Finally, we were surprised to discover that P_d started the exploration process through the outcome-driven approach. She mentioned that she hoped to prevent her knowledge bias. *"Clinicians might possess similar knowledge bases. If I aim to discover more promising designs for eligibility criteria, starting the exploration through the outcome-driven approach could be effective."*

Visual Designs and Interactions. All the experts praised the clear and user-friendly visual designs and interactions of our system. First, they highlighted the smooth combination of the three sub-views in the Criterion-outcome Exploration View, which allowed for easy criterion setting, candidate outcome examination, and exploration history tracking. P_a found the interaction between the Criterion View and Outcome View to be highly suited to his needs. He expressed, *"This exploration approach can also be extended to various stratified research in the medical field."* P_a and P_e both expressed high appreciation for the ability to create stages, as it allowed them to effectively organize and recall different explorations. We evaluated their perception of the system using the NASA Task Load Index [23], a 7-point scale. First, the overall system design is not complex, as indicated by the average scores for mental demand (3.4), physical demand (2.4), effort (2.8), and frustration (1.8). However, the temporal demand score is 4.6. This could be attributed to the iterative nature of the exploration process, where experts engage in repeated iterations. Furthermore, the average performance score is 1.7 (with scores closer to 1 indicating a higher level of perceived performance perfection), indicating that the experts are highly confident of the final result obtained from our system.

Suggestions. Experts offered several suggestions for further enhancements. First, they recommended highlighting regions in the Outcome View to make it easier to identify interesting candidates. Second, P_e suggested providing greater flexibility in organizing exploration, such as enabling a hierarchical tree format. Lastly, several experts proposed automatically recommending potentially important criteria to ensure they are not overlooked due to knowledge limitations.

8 Discussion

Design Implications. We have identified two important aspects during the system design process. First, our findings highlight the importance of supporting clinicians in systematically tracking their iterative exploration process. In various clinical scenarios, the decision-making process is always non-linear and exploratory, often requiring backtracking, hypothesis refinement, and contextual sense-making [20, 41]. While prior work in visual analytics [52] has emphasized the value of provenance tracking and cognitive support, our work reinforces these findings. We observed that allowing clinicians to define their own reasoning stages and annotate their thought process—through note-taking and snapshotting—helps externalize reasoning, making it easier to revisit and replicate successful strategies while avoiding redundant exploration. Another key implication is the need to support the adaptive integration of clinical expertise within exploratory decision workflows. We observed that clinicians often shift between knowledge-driven and outcome-driven strategies, depending on their evolving goals and domain understanding. This echoes findings from VBridge [10], where clinicians alternated between forward and backward analyses when collaborating with AI. Our results suggest that such hybrid reasoning patterns may also apply in broader, non-AI contexts. Therefore, future systems should therefore accommodate these fluid transitions and support flexible, mixed-strategy reasoning.

Generalizability. In this work, we focus on the design of eligibility criteria in clinical trials, a crucial step before participant enrollment. However, conducting a clinical trial involves multiple steps [19]. For instance, clinicians are tasked with patient screening during participant enrollment to confirm that participants adhere to the trial's criteria. Furthermore, after enrollment, they need to conduct experimental procedures and monitor patient conditions. Finally, they will assess adverse events and interpret the final results. Our system can also be utilized in several steps. For example, the Criterion-outcome Exploration View can help clinicians organize and analyze the final treatment effectiveness of different subgroups. Beyond clinical trials, our system can also inspire other decision-making scenarios. For example, our system integrates outcome metrics with temporal detailed characteristics. This can be used for long-term investment allocation. Currently, various models have been developed to predict how much profit and risk investment will bring [22]. They allow investors to simulate various investment allocation strategies and assess the potential outcomes. Investors also require detailed information (e.g., market context or current investment trend) with these outcome metrics during their decision-making process. Additionally, the combination of the knowledge-driven and outcome-driven approaches can allow them to better leverage their prior domain knowledge as well as data-driven techniques.

Scalability. First, scalability issues can arise in the Outcome View. When the number of criteria or possible adjustments for each criterion increases, the number of criterion candidates will grow rapidly. However, since experts' initial exploration focuses on understanding the overall distribution in the scatter plot, even with a high number of criterion candidates, they can explore based on distribution without needing to examine individual candidates. Thus, the candidate count does not hinder their exploration. Additionally, our system allows clinicians to take an iterative approach. They can start by avoiding overly granular adjustments for each eligibility criterion. After gaining insight into the appropriate adjustment range, they can then fine-tune the criteria more precisely. In the future, we plan to introduce a sampling approach. When data volumes are large, uniform sampling initially will not compromise users' judgment of the overall distribution. As users refine their selections, we can gradually reintroduce previously hidden candidates. This method can ensure smooth exploration while maintaining the system's rendering capabilities. Scalability issues may also arise in the Exploration View with numerous exploration records, making it difficult for experts to grasp the overall context. Currently, we allow clinicians to specify important stages. In the future, we could add features like hierarchical organization of exploration history.

Limitations. Firstly, our system incorporated five outcome metrics based on our literature survey and expert interviews. While we expanded the metrics compared to previous tools, some suggested metrics were not included due to the dataset limitation. Incorporating more relevant metrics in the system can be a future work. For example, detailed heart-related risk metrics could provide important insights into cardiovascular complications and help better assess patient safety during trials. Secondly, our system is currently limited to examining kidney and liver details from EHR data. Clinicians may need to explore more detailed information and define a broader spectrum of risk events, which presents an opportunity for future enhancements.

9 Conclusion

In this work, we proposed *TrialCompass*, a visual analytics system to assist clinicians in designing eligibility criteria for clinical trials. We developed a novel workflow that enables exploration of eligibility criteria through both knowledge-driven and outcome-driven approaches. Additionally, we integrated a history-tracking feature to support clinicians in their iterative design process. Using the MIMIC IV dataset, we conducted expert interviews and case studies, uncovering new insights for eligibility criteria in two major diseases. Finally, we highlighted several research opportunities that arise from applying visualization techniques to enhance clinical trial workflows.

Acknowledgments

We sincerely thank all our collaborators and reviewers. We also appreciate experts who participated in the development and validation of our system. Lastly, we express our gratitude to all the partners who have helped us with our project.

References

[1] R. Aggarwal and P. Ranganathan. Study Designs: Part 4 – Interventional Studies. *Perspectives in Clinical Research*, 10(3):137–139, 2019. doi: 10.4103/picr.PICR_91_19 3

[2] P. C. Austin. Optimal Caliper Widths for Propensity-score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies. *Pharmaceutical Statistics*, 10(2):150–161, 2011. doi: 10.1002/pst.433 4

[3] E. Bartoli. Adverse Effects of Drugs on the Kidney. *European Journal of Internal Medicine*, 28:1–8, 2016. doi: 10.1016/j.ejim.2015.12.001 3

[4] M. Berk, R. L. Woods, M. R. Nelson, R. C. Shah, C. M. Reid, E. Storey, S. Fitzgerald, J. E. Lockery, R. Wolfe, M. Mohebbi, S. Dodd, A. M. Murray, N. Stocks, P. B. Fitzgerald, C. Mazza, B. Agustini, and J. J. McNeil. Effect of Aspirin vs Placebo on the Prevention of Depression in Older People: A Randomized Clinical Trial. *JAMA Psychiatry*, 77(10):1012–1020, 2020. doi: 10.1001/jamapsychiatry.2020.1214 3

[5] B. E. Blass. *Basic Principles of Drug Discovery and Development*. Elsevier, 2015. doi: 10.1016/C2017-0-02030-X 1

[6] A. G. Bostom, F. Kronenberg, and E. Ritz. Predictive Performance of Renal Function Equations for Patients with Chronic Kidney Disease and Normal Serum Creatinine Levels. *American Society of Nephrology*, 13(8):2140–2144, 2002. doi: 10.1097/01.asn.0000022011.35035.f3 5

[7] V. Braun and V. Clarke. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa 4

[8] G. Carenini and J. Loyd. ValueCharts: Analyzing Linear Models Expressing Preferences and Evaluations. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 150–157, 2004. doi: 10.1145/989863.989885 3

[9] L. Chen, H. Wang, Y. Ouyang, Y. Zhou, N. Wang, and Q. Li. FSLens: A Visual Analytics Approach to Evaluating and Optimizing the Spatial Layout of Fire Stations. *IEEE Transactions on Visualization and Computer Graphics*, 30:847–857, 2024. doi: 10.1109/TVCG.2023.3327077 3

[10] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):378–388, 2022. doi: 10.1109/TVCG.2021.3114836 2, 9

[11] Y.-E. Claessens, P. Aegerter, H. Boubaker, B. Guidet, and A. Cariou. *Critical Care*, 17(3), article no. R89, 9 pages, 2013. doi: 10.1186/cc12734 2

[12] F. Dabek, E. Jimenez, and J. J. Caban. A Timeline-based Framework for Aggregating and Summarizing Electronic Health Records. In *2017 IEEE Workshop on Visual Analytics in Healthcare*, pp. 55–61, 2017. doi: 10.1109/VAHC.2017.8387501 2

[13] M. Desai. Recruitment and Retention of Participants in Clinical Studies: Critical Issues and Challenges. *Perspectives in Clinical Research*, 11(2):51–53, 2020. doi: 10.4103/picr.picr_6_20 3

[14] J. D. Dziura, L. A. Post, Q. Zhao, Z. Fu, and P. Peduzzi. Strategies for Dealing with Missing Data in Clinical Trials: From Design to Analysis. *Yale Journal of Biology and Medicine*, 86(3):343–358, 2013. doi: 10.4103/cmi.cmi_8_24 5

[15] A. Faiola and C. Newlon. Advancing Critical Care in the ICU: A Human-centered Biomedical Data Visualization Systems. In *Proceedings of the International Conference on Ergonomics and Health Aspects of Work with Computers*, pp. 119–128, 2011. doi: 10.1007/978-3-642-21716-6_13 2

[16] Y. Fang, H. Liu, B. Idnay, C. Ta, K. Marder, and C. Weng. A Data-driven Approach to Optimizing Clinical Study Eligibility Criteria. *Journal of Biomedical Informatics*, 142:104375, 2023. doi: 10.1016/j.jbi.2023.104375 2

[17] L. Fehrenbacher, L. Ackerson, and C. Somkin. Randomized Clinical Trial Eligibility Rates for Chemotherapy (CT) and Antiangiogenic Therapy (AAT) in a Population-based Cohort of Newly Diagnosed Non-small Cell Lung Cancer (NSCLC) Patients. *Journal of Clinical Oncology*, 27(15_suppl):6538–6538, 2009. doi: 10.1200/jco.2009.27.15_suppl.6538 2

[18] J. C. Ferreira and C. M. Patino. Types of Outcomes in Clinical Research. *Jornal Brasileiro de Pneumologia*, 43(1):5, 2017. doi: 10.1590/S1806-37562017000000021 3

[19] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of Clinical Trials*. Springer, 2015. doi: 10.5213/inj.2013.17.2.96 9

[20] N. Gladstone. Comparative Theories in Clinical Decision Making and Their Application to Practice: A Reflective Case Study. *British Journal of Anaesthetic & Recovery Nursing*, 13(3-4):65–71, 2012. doi: 10.1017/S1742645612000435 9

[21] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013. doi: 10.1109/TVCG.2013.173 3

[22] C. Gu. Research on Prediction of Investment Fund's Performance before and after Investment Based on Improved Neural Network Algorithm. *Wireless Communications and Mobile Computing*, 2021(1):5519213, 2021. doi: 10.1155/2021/5519213 9

[23] S. G. Hart. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909 9

[24] J. S. Hirsch, J. S. Tanenbaum, S. Lipsky Gorman, C. Liu, E. Schmitz, D. Hashorva, A. Ervits, D. Vawdrey, M. Sturm, and N. Elhadad. HARVEST: A longitudinal Patient Record Summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014. doi: 10.1136/amiajnl-2014-002945 2

[25] G. D. Huang, J. Bull, K. J. McKee, E. Mahon, B. Harper, J. N. Roberts, C. R. P. Team, et al. Clinical Trials Recruitment Planning: A Proposed Framework From the Clinical Trials Transformation Initiative. *Contemporary Clinical Trials*, 66:74–79, 2018. doi: 10.1016/j.cct.2018.01.003 2

[26] Z. Jiang, H. Chen, R. Zhou, J. Deng, X. Zhang, R. Zhao, C. Xie, Y. Wang, and E. C. Ngai. HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1205–1215, 2024. doi: 10.1109/TVCG.2023.3326943 2

[27] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. MIMIC-IV, A Freely Accessible Electronic Health Record Dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x 4, 6

[28] M. D. Kelsey, B. Patrick-Lake, R. Abdulai, U. C. Broedl, A. Brown, E. Cohn, L. H. Curtis, C. Komelasky, M. Mbagwu, G. A. Mensah, R. J. Mentz, A. Nyaku, S. O. Omokaro, J. Sewards, K. Whitlock, X. Zhang, and G. S. Bloomfield. Inclusion and Diversity in Clinical Trials: Actionable Steps to Drive Lasting Change. *Contemporary Clinical Trials*, 116:106740, 2022. doi: 10.1016/j.cct.2022.106740 3

[29] J. Kim, C. Ta, C. Liu, C. Sung, A. Butler, L. Stewart, L. Ena, J. Rogers, J. Lee, A. Ostropolets, P. Ryan, H. Liu, S. Lee, M. Elkind, and C. Weng. Towards Clinical Data-driven Eligibility Criteria Optimization for Interventional COVID-19 Clinical Trials. *Journal of the American Medical Informatics Association*, 28(1):14–22, 2021. doi: 10.1093/jamia/ocaa276 2

[30] Y.-H. Kuo, B. Martínez-López, and K.-L. Ma. Investigating Animal Infectious Diseases with Visual Analytics. In *IEEE Pacific Visualization Symposium*, pp. 71–81, 2023. doi: 10.1109/PacificVis56936.2023.00015 2

[31] S. A. Kwee, L. L. Wong, C. Ludema, C. K. Deng, D. Taira, T. Seto, and D. Landsittel. Target Trial Emulation: A Design Tool for Cancer Clinical Trials. *JCO Clinical Cancer Informatics*, (7):e2200140, 2023. doi: 10.1200/CCI.22.00140 2, 3

[32] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng. DPVis: Visual Analytics With Hidden Markov Models for Disease Progression Pathways. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3685–3700, 2021. doi: 10.1109/TVCG.2020.2985689 2

[33] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027 2

[34] Z. Li, X. Liu, Z. Cheng, Y. Chen, W. Tu, and J. Su. Trialview: An ai-powered visual analytics system for temporal event data in clinical trials. In *Proceedings of the Hawaii International Conference on System*

Sciences, p. 1169—1178, 2024. doi: 10.48550/arXiv.2310.04586 2

[35] A. Licata. Adverse Drug Reactions and Organ Damage: The Liver. *European Journal of Internal Medicine*, 28:9–16, 2016. doi: 10.1016/j.ejim.2015.12.017 3

[36] C. G. Linhares, D. M. Lima, J. R. Ponciano, M. M. Olivatto, M. A. Gutierrez, J. Poco, C. Traina, and A. M. Traina. ClinicalPath: A Visualization Tool to Improve the Evaluation of Electronic Health Records in Clinical Decision-Making. *IEEE Transactions on Visualization and Computer Graphics*, 29(10):4031–4046, 2023. doi: 10.1109/TVCG.2022.3175626 2

[37] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, and Y. Wu. SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):1–10, 2017. doi: 10.1109/TVCG.2016.2598432 3

[38] R. Liu, S. Rizzo, and S. e. a. Whipple. Evaluating Eligibility Criteria of Oncology Trials using Real-world Data and AI. *Nature*, 592(7855):629–633, 2021. doi: 10.1038/s41586-021-03430-5 2, 3, 4

[39] S. M. Opal, P.-F. Laterre, B. Francois, S. P. LaRosa, D. C. Angus, J.-P. Mira, X. Wittebole, T. Dugernier, D. Perrotin, M. Tidswell, L. Jauregui, K. Krell, J. Pachl, T. Takahashi, C. Peckelsen, E. Cordasco, C.-S. Chang, S. Oeyen, N. Aikawa, T. Maruyama, R. Schein, A. C. Kalil, M. Van Nuffelen, M. Lynn, D. P. Rossignol, J. Gogate, M. B. Roberts, J. L. Wheeler, J.-L. Vincent, and f. t. ACCESS Study Group. Effect of Eritoran, an Antagonist of MD2-TLR4, on Mortality in Patients With Severe Sepsis: The ACCESS Randomized Trial. *Journal of the American Medical Informatics Association*, 309(11):1154–1162, 03 2013. doi: 10.1001/jama.2013.2194 3

[40] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer. WeightLifter: Visual Weight Space Exploration for Multi-Criteria Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):611–620, 2017. doi: 10.1109/TVCG.2016.2598589 3

[41] T. Pelaccia, J. Tardif, E. Triby, C. Ammirati, C. Bertrand, V. Dory, and B. Charlin. How and When Do Expert Emergency Physicians Generate and Evaluate Diagnostic Hypotheses? A Qualitative Study Using Head-Mounted Video Cued-Recall Interviews. *Annals of Emergency Medicine*, 64(6):575–585, 2014. doi: 10.1016/j.annemergmed.2014.05.003 9

[42] D. Pratt and M. Kaplan. Evaluation of Abnormal Liver-Enzyme Results in Asymptomatic Patients. *The New England Journal of Medicine*, 342(17):1266—1271, 2000. doi: 10.1056/nejm200004273421707 5

[43] P. R. Rosenbaum and D. B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41 4

[44] P. Royston and M. K. B. Parmar. Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-event Outcome. *BMC Medical Research Methodology*, 13:152, 2013. doi: 10.1186/1471-2288-13-152 3

[45] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore. Hazard Ratio in Clinical Trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792, 2004. doi: 10.1128/aac.48.8.2787-2792.2004 5

[46] N. Sultanum, F. Naeem, M. Brudno, and F. Chevalier. ChartWalk: Navigating Large Collections of Text Notes in Electronic Health Records for Clinical Chart Review. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1244–1254, 2023. doi: 10.1109/TVCG.2022.3209444 2

[47] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert. Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):288–297, 2018. doi: 10.1109/TVCG.2017.2745078 3

[48] Q. Wang and R. S. Laramee. EHR STAR: The State-Of-the-Art in Interactive EHR Visualization. *Computer Graphics Forum*, 41:69–105, 2022. doi: 10.1111/cgf.14424 2

[49] Q. Wang, T. Mazor, T. Harbig, E. Cerami, and N. Gehlenborg. ThreadStates: State-based Visual Analysis of Disease Progression. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):238–247, 2022. doi: 10.1109/TVCG.2021.3114840 2

[50] D. Weng, R. Chen, Z. Deng, F. Wu, J. Chen, and Y. Wu. SRVis: Towards Better Spatial Integration in Ranking Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):459–469, 2019. doi: 10.1109/TVCG.2018.2865126 3

[51] D. Weng, H. Zhu, J. Bao, Y. Zheng, and Y. Wu. HomeFinder Revisited: Finding Ideal Homes with Reachability-Centric Multi-Criteria Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 12 pages, 2018. doi: 10.1145/3173574.3173821 3

[52] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the Analysis of User Interactions and Visualization Provenance. In *Computer Graphics Forum*, vol. 39, pp. 757–783, 2020. doi: 10.1111/cgf.14035 9

[53] Y. Zhang, K. Chanana, and C. Dunne. IDMVis: Temporal Event Sequence Visualization for Type 1 Diabetes Treatment Decision Support. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):512–522, 2019. doi: 10.1109/TVCG.2018.2865076 2

[54] X. Zhao, Y. Wu, W. Cui, X. Du, Y. Chen, Y. Wang, D. L. Lee, and H. Qu. SkyLens: Visual Analysis of Skyline on Multi-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):246–255, 2018. doi: 10.1109/TVCG.2017.2744738 3