DOI: 10.1111/cgf.14305
Eurographics Conference on Visualization (EuroVis) 2021
R. Borgo, G. E. Marai, and T. von Landesberger
(Guest Editors)

Volume 40 (2021), Number 3

SumRe: Design and Evaluation of a Gist-based Summary Visualization for Incident Reports Triage

T. Kakar[1], X. Qin[1], T. La [2], S. K. Sahoo[2], S. De[2], E. A. Rundensteiner[1], and L. Harrison[1]

[1] Computer Science Department, Worcester Polytechnic Institute, Worcester, USA
[2] Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Maryland, USA

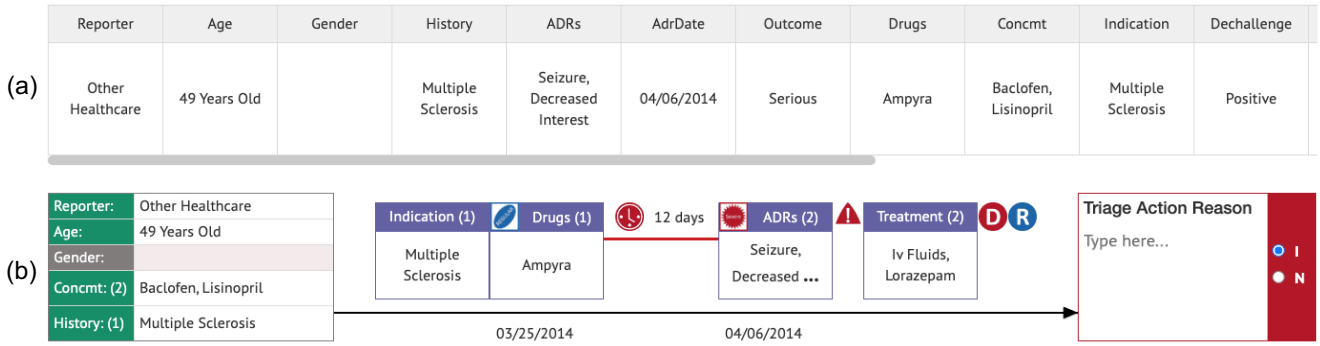

Figure 1: *(a) Triage information displayed using Tabular layout, modeled after current practices at US Food and Drug Administration (FDA). (b) Following interviews with FDA analysts, our gist-based summary, SumRe, supports their triage workflows by displaying the same information using visual encodings to facilitate analysis of critical incident details at a glance along with recording analyst's decision making.*

Abstract

Incident report triage is a common endeavor in many industry sectors, often coupled with serious public safety implications. For example, at the US Food and Drug Administration (FDA), analysts triage an influx of incident reports to identify previously undiscovered drug safety problems. However, these analysts currently conduct this critical yet error-prone incident report triage using a generic table-based interface, with no formal support. Visualization design, task-characterization methodologies, and evaluation models offer several possibilities for better supporting triage workflows, including those dealing with drug safety and beyond. In this work, we aim to elevate the work of triage through a task-abstraction activity with FDA analysts. Second, we design an alternative gist-based summary of text documents used in triage (SumRe). Third, we conduct a crowdsourced evaluation of SumRe with medical experts. Results of the crowdsourced study with medical experts ($n = 20$) suggest that SumRe better supports accuracy in understanding the gist of a given report, and in identifying important reports for followup activities. We discuss implications of these results, including design considerations for triage workflows beyond the drug domain, as well as methodologies for comparing visualization-enabled text summaries.

1. Introduction

Incident reports detail issues related to products or services are submitted to regulatory agencies. For instance, the Federal Aviation Agency (FAA) analyzes service difficulty reports about aircraft's maintenance issues [MR12]. The US Consumer Financial Protection Bureau [Bur11] is responsible for collecting and analyzing consumer complaints about unfair or deceptive financial practices. The U.S. Food and Drug Administration reviews drug safety reports to identify previously undiscovered adverse drug reactions [HVG08], one of the leading causes of death worldwide [LPC98].

What is common about these incident report systems is that they all include semi-structured text data. An incident report might include information about the who or what organization submitted it, structured information about the incident type or details, quantitative measures pertaining to the incident, (*e.g.* money lost, medical measurements), *etcetera*. While there have been several research efforts targeting text data, for example the Yelp review analysis system [FPB16] from Felix *et al.*, and JigSaw [GLK*12] from Görg *et al.*, these typically focus on overview and exploratory analysis tasks. In contrast, triage workflows demand individual attention on individual reports, suggesting alternate system visualization designs and workflows are needed.

In this work, we thus address challenges in triage workflow from

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

a visualization standpoint by focusing on drug safety *incident reports triage*. At the US FDA, analysts encounter hundreds of reports in their caseloads where they review information within each report to determine if the report needs further investigation. For example, differences in patients' demographics, medical histories, and other drugs concurrently taken can make an outsize impact on the criticality of a report, even among reports that focus on the same drugs and adverse reactions. Analysts must dedicate substantial time and attention to minimize the risk that life-threatening safety issues are missed amidst the influx of reports received on a daily basis.

Through interviews with analysts at the US FDA and task-characterization, we establish the promise of visualization-enabled designs for improving reports triage. Currently, analysts at the FDA use a traditional SQL-style table layout (hereafter referred to as Tabular layout) to summarize key information in the reports narratives (see Figure 2). A visual examination of the Tabular layout suggests room for improvement (*e.g.* the unused white space, need to scroll, etc.). To improve on this design, we observed FDA analysts using the tool as part of their daily work. We also conducted a series of interviews and discussions. We report this task characterization in this manuscript, followed by describing how we utilized this to inform our alternative design and its evaluation.

Informed by the task characterization, we design SumRe, a compact gist-based visual representation of an incident report, which aims to align with and facilitate analysts' triage workflows. The design of SumRe follows a multi-stage model which decomposes subtasks in the triage process, as reported by analysts. For instance, we characterize the most-used data elements in report triage as *triage cues*, and use them to inform visual encoding choices including space, color, and icons (following Maguire [Mag14]). To evaluate SumRe, we conduct a crowdsourced controlled study with twenty medical experts, comparing SumRe to the Tabular baseline. The quantitative results of this study suggest that SumRe aids analysts in accurately assessing high-level qualities of reports (*i.e.* the "gist") as well as the detail-oriented assessment for determining whether reports need followup. Moreover, our results suggest that SumRe leads to better short-term recall, and qualitatively different insights compared to the Tabular form. Participants also reported that analyzing reports with SumRe was a more enjoyable experience.

We summarize our main contributions as follows:

- We contribute a task characterization of incident report triage by observing 6 FDA domain experts at work using think-aloud protocols and a series of followup discussions. These activities form the basis of a set of design requirements for a compact yet expressive design to support efficient *reports triage*.
- We design a *gist-based visual summary*, SumRe, drawing on principles from information visualization and glyph-based visual design. SumRe facilitates tasks from domain-specific workflows using visual encodings such as spatial alignment, color, and word-scale icons with the goal to reduce effort by analysts by supporting the "gist"-based reasoning given a report.
- We develop an empirical study of triage-centered tasks to compare SumRe and the Tabular baseline. Results from 20 medical experts suggest SumRe facilitates high-level analysis as well as accurate assessment of details needed to identify the most important reports, with comparable performance across other key metrics.

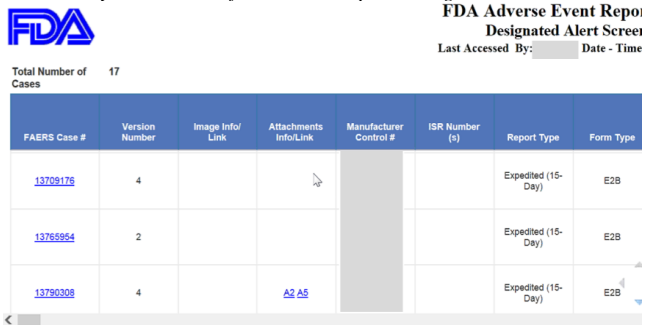

Figure 2: *Currently FDA uses this standard Tabular layout for incident triage, where scrolling is needed to view a report. We aim to design and evaluate an alternative visual summary to better communicate critical safety report information.*

Taken together, these efforts underscore the need for techniques and evaluation methodologies that target the role of visualization in the difficult yet critical work of incident reports triage.

2. Background and Related Work

SumRe relates to previous work in three ways, including visualization for text documents; visualization for triage in domains such as cyber security; and efforts in improving drug-safety issue management. We briefly cover prior work in these areas, focusing on prior work which we draw from in the design and evaluation of SumRe.

2.1. Visual Analysis of Text Documents

The vast majority of text visualization techniques designed for the exploration of a document corpus display metadata about the corpus, such as results of document clustering [CWDH09], topics [LZP*12], and name-entities [GLK*12]. Feature Lens [DZG*07] allows the visual exploration of frequent text patterns in text collections. TextTile [FPB16] allows users to explore a set of documents by providing interaction operations and views. These approaches provide an aggregated summary of a set of documents using multiple linked views for exploratory analysis, our focus instead is to provide a glancable summary of each individual report for triage.

Visualization techniques for a single document also exist. Wordclouds, also known as tag-clouds, have been widely used as an exploratory tool to provide a high-level overview of a single or multiple text documents [VW08, KHGW07]. Docuburst [CCP09] uses a space-filling approach to visualize document content by depicting relevant terms along with their semantic relationships. Word Trees [WV08], a graphical version of 'keyword-in-context', visualizes sentences sharing the same beginning in the form of a tree. Phrase Nets [VHWV09] use a graph-based visualization to display relationships among words. Similarly, work on visual summaries of individual text documents also exists. For instance, Liu et al [LSL03] visually summarize the emotions in a text document. In the literary domain, long texts such as books and novels have been visually summarized at multiple levels of abstraction to facilitate navigation within the document [KJW*14]. For example, Document Cards [SOR*09] provide a compact summary of a publication consisting tf-idf-based keywords and images from the publication to support exploration.

2.2. Triage in Other Domains

Existing research in document triage has focused on designing thumbnails [AKG*10] or image-based [CAS13] web-page pre-

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

views to help users select relevant webpages. Other efforts have studied user behavior during the triage of research articles [BBM*06] such as skimming through the headings and titles. Trist [JWS*05], an information triage tool, provides an overview of large document corpora to help in document comparison and trend analysis. Approaches on designing visualizations to triage emails also exist [NBS05, HHT14]. CueT [ALK*11], a network alarm triage tool, uses machine learning to prioritize network alarms to help operators quickly identify and fix them. While we share CueT's idea of helping analysts in identifying important reports, CueT is neither designed nor evaluated for incident report triage.

2.3. Incident Report Analytics

Much of the existing work in Pharmacovigilance has focused on applying computational techniques such as natural language processing (NLP) on incident reports to extract name-entites from unstructured text [WQK*17, KWMJ*15, STKO13]. Publicly available online tools such as OpenFDA [KHXM*15] and OpenFDAVigil [BvHH*16] help the general public explore incident reports and learn basic statistics about a certain drug or reaction. Other approaches have visualized relationships between drugs and their reported reactions [YBH14, JXYXJ*15, KQR*19]. These tools help in exploring the reports at an aggregate level; but they are not designed to visually summarize an individual report.

3. A Task-Characterization of Incident Report Analysis

The FDA regularly receives incident reports about medication errors and adverse reactions through their post-marketing drug surveillance program called FAERS [FA15]. Each report has structured information such as patients demographics, therapy and event related information, as well as a text narrative describing the details of the reaction suspected to be caused by the drug. FDA analysts review hundreds of incident reports on a daily basis. Their goal is to identify reports indicative of potential safety problems such as a previously undiscovered reaction that needs regulatory action. Regulatory action includes adding a warning to the drug label or in worst cases removing the drug from the market.

Incident reports triage is a process where drug analysts review each individual report related to a certain drug or reaction from a larger batch of reports. The goal of individual reports triage is to assess the association between the drug and reactions, and to determine whether a report requires further investigation. More specifically, analysts review certain structured information that we call **triage cues** in each report and formulate a hypothesis as to whether the report is indicative of a potential issue and thus should be investigated. Investigation beyond this stage means that she would read the text narrative to sift through the details of the incident.

After assessing the report's contents, if an analyst considers a report indicative of a potential safety issue, further evidence is then sought by collecting similar reports and considering the domain-related plausibility of the incident. When both of these conditions are met, the evidence is compiled along with recommendations for a regulatory action to mitigate the safety issue. While steps beyond this triage process are important, they are sufficiently complex to require design and interventions beyond the scope of this work. The complexity of the post-triage process, combined with the fact that FDA analysts spend a substantial portion of their time on triage itself, leads us to constrain the scope of this work to focus on triage with its own distinct set of challenges.

3.1. Triage Requirements Elicitation and Characterization

To better understand what requirements were necessary to aid analysts in the identification of reports indicative of investigation, we conducted one-hour in-person semi-structured interviews with six (6) drug safety analysts at the FDA. In these sessions, we observed the analysts performing their regular triage tasks, and asked follow-up questions to characterize their thought processes. In these efforts, our goal was to understand what parts of the report summary are attended to during triage, and how these parts are analyzed to render an assessment. We recorded and transcribed these sessions to facilitate the synthesis of requirements. As an intermediary goal, following these sessions we had additional remote discussions with these analysts to clarify and then crystallize a model of their process and key requirements. Details of the minor requirements are reflected in the design process as discussed in Section 4.

3.2. Triage Cues: The Critical Information Used in Triage

One major requirement that emerged from interviews with FDA analysts is the relative weighting of report summary information, which we characterize as *triage cues*. In particular, analysts seek information about the patient, such as their demographics and history, the incident, such as the drug and reaction, and the events after the incident, such as, interventions taken to mitigate the incident. Figure 3 depicts common triage cues (data attributes) in the sequence they appear in a report (arrows and spatial position) as well as their priority in being assessed during triage within a report (color). The majority of these triage cues are high cardinality, that is, they have many categories, such as different drugs or reactions. For instance, Aspirin and Tylenol are two categories of the drug cue, while nausea and headache are categories of the reaction cue.

3.2.1. Primary Triage Cues for Hypothesis Formation

Here we describe how these triage cues are assessed during triage. Some triage cues help analysts to determine whether the report should be investigated or not. We call these cues **Primary Triage Cues** as analysts assess them first to know if the incident is serious and/or plausible, depending upon the quantity and quality of useful information present in the report.

Primary Triage Cues for Seriousness. To assess seriousness, analysts look at the outcome of the report, where "serious" typically means a negative outcome, such as the patient dying or being hospitalized (Fig. 3). Analysts also prioritize a report if the drug is a new molecular entity, i.e., a new drug. This is because analysts closely monitor new drugs as they have not been in the market for a long time and chances of them causing undiscovered adverse reactions (ADRs) are high compared to those in long-term use. Analysts also look for the severity of the reported adverse reaction. For instance, a renal failure or seizure is a severe reaction and thus worth investigating as compared to a headache or nausea.

Primary Triage Cues for Plausibility. Analysts also form a hypothesis about the importance of a report by evaluating the plausibility of the incident by reviewing the 'Onset', 'Dechallenge', and 'Rechallenge' triage cues (Fig. 3). Onset is the duration between the date when the drug was taken and the date when the reaction was observed. Onset helps analysts assess the possibility of the reaction

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

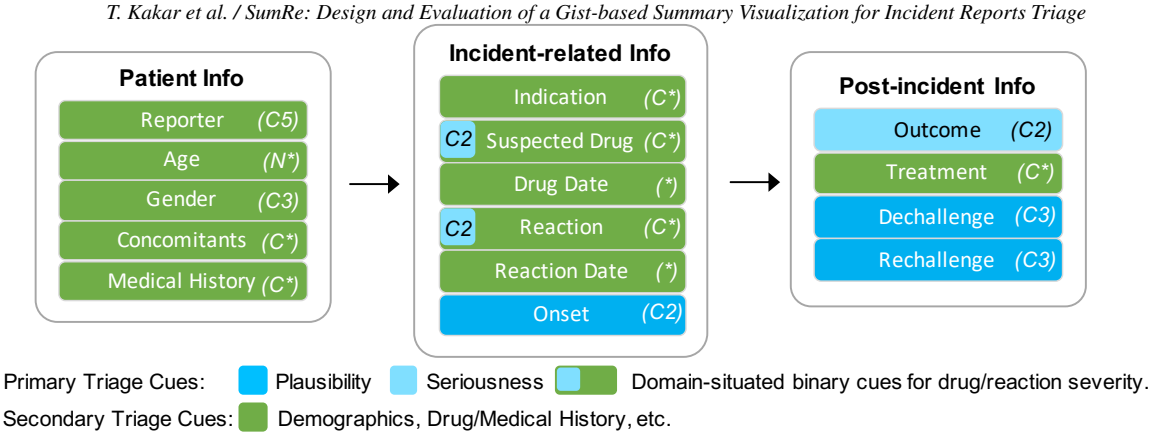

Figure 3: *Several design activities with FDA analysts led to the depicted multi-stage model of drug incident reports triage. This model drives the spatial layout and other design elements in SumRe by reflecting the sequence of cues drug analysts examine related to patient characteristics.* Triage cues *are a key part of incident report triage. Analysts first seek primary cues to form an assessment of plausibility, severity, and likelihood of further investigation. Analysts then review secondary cues that may support or challenge their hypothesis. (Xn) represents the attribute type (where X=C means Categorical and X=N Numeric) and its cardinality. E.g. (C*) means a high-cardinality (*) categorical attribute (C) with hundreds of categories.*

being associated with the drug. Dechallenge and Rechallenge, on the other hand, are clinical actions taken by medical professionals to assess if the reaction is associated with the drug. Dechallenge implies that the reaction disappeared after stopping the drug, and rechallenge that the reaction recurred after restarting the drug.

3.2.2. Secondary Triage Cues to Support or Reject Hypothesis

After forming a hypothesis about whether the association between the drug and reaction is plausible or if the report is serious and hence important by reviewing the Primary Triage Cues, analysts seek additional information to further support their initial hypothesis. We call these cues **Secondary Triage Cues**. For instance, the analyst may want to know if the drug is known to be causing the reaction, or patient's medical history may be the reason for the incident (reaction). Similarly, reporter type is important for triage because if a severe and rare reaction is reported by a medical professional then it has more value than when reported by a patient due to the former's medical expertise. Hence analysts will prioritize investigating it further. These secondary cues support or reject the analyst's hypothesis about whether the report needs investigation or not.

Missing Information. Information including key triage cues can be missing in incident reports due to poor reporting – a world-wide problem plaguing Pharmacovigilance [BNL14]. Missing information may complicate the triage process because based on the type of missing triage cues, the assessment of a report can become challenging. However, if a report has many triage cues missing and there is less information to assess an incident, then analysts are likely to quickly form a hypothesis that the report does not need investigation.

3.3. Design Goals for a Compact Visual Summary – SumRe

Based on our interviews with the FDA analysts (Sec. 3.1), we identified the following design goals for SumRe.

DG1: Provide a compact view to facilitate comprehensive analysis. Analysts review multiple triage cues (Fig. 3) collectively to decide if a report needs further investigation. Our goal is to provide a compact view of this expansive and informative data to make glancing at a report feasible.

DG2: Differentiate among diverse triage cues. Analysts prioritize primary cues for forming a hypothesis about the report, and thereafter tend to focus on secondary cues to seek supporting evidence for their hypothesis. The design thus needs to support an ease in differentiating between these classes of triage cues.

DG3: Facilitate capturing triage artifacts. Our goal is to allow analysts capture their triage related comments and actions; with the aim to support them in keeping track of their analyses as well as facilitate information recall at a later stage.

4. Designing SumRe: a Visual Summary for Reports Triage

Following our collaboration with FDA analysts and initial task characterization activities, we developed an alternative summary method, SumRe. SumRe seeks to address the identified challenges in incident report triage, while drawing on visualization principles to effectively align with the identified primary/secondary cue workflow we observed from analysts. The final design as depicted in Fig. 4 is a result of multiple iterations and discussions with the domain experts. After initial prototype designs, we refined the aforementioned requirements by obtaining further details on how analysts process triage cues via discussions with the FDA experts. One outcome of this activity was the use of icons to summarize information. Further iterations explored alternatives related to the order and visual encodings of the triage cues.

4.1. Overall SumRe Layout

The information in our design is structured from left to the right following the order of events happened to the patient within a report. We used a primary visual channel— spatial position— to represent the sequence of the cues to follow this natural flow of the events. For instance, SumRe can be read starting from patient and reporter, to the details of the incident and the events afterwards.

One example outcome of the design process comes from the placement of the indication component, which is the disease for which drug is taken or prescribed. During our discussions with the FDA analysts, they suggested to place indication next to the drug as it would help them know right away why the person took the drug.

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

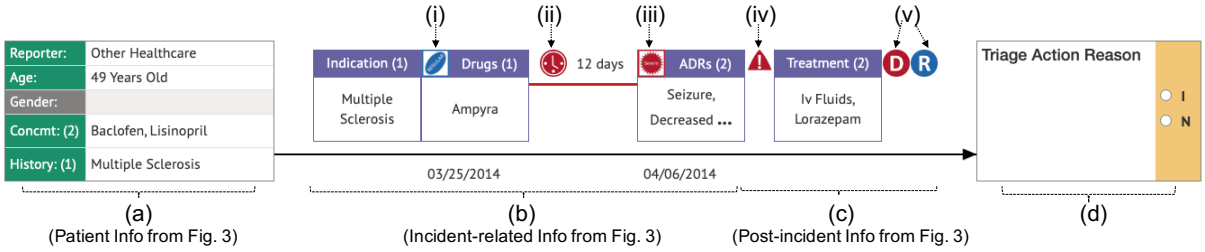

Figure 4: *SumRe layout to triage incident reports. (a) Profile Panel presenting information about the patient and the reporter. (b) Incident Panel representing all the information associated with the reaction (incident). (c) Analysis Panel allows analysts to add comments and triage action (Investigate or Do Not Investigate). Icons: (i) Regular drug, (ii) Onset, (iii) Severe Reaction, (iv) Serious outcome, (v) Dechallenge and Rechallenge respectively. Missing information is represented with grey color. Arrows are added for illustration and are not part of the design.*

According to one expert, "Indication is important for our analysis, as most of the times one drug can be used to treat multiple symptoms. Such as, Propranolol is used to lower blood pressure but it is also prescribed to prevent migraines. Knowing the drug's association with the reaction for a certain indication is helpful". Below we discuss further components of the design.

Profile Panel. The Profile Panel (Fig. 4a) displays the triage cues related to the patient. Basically it answers the questions "who did the incident happen to, and who reported it". These cues represent supplementary information and are not associated directly with the incident. Following expert feedback, we added the counts for the high-dimensional cues to allow analysts to assess the information on the fly. We were told that having many underlying conditions (history) and taking multiple drugs at a time hinders the analysis. According to one expert, "We are looking for confounders. e.g., If we have a report where patient has taken 10 drugs and has a reaction, that report is not gonna help us in assessing the incident as compared to a report where a patient has taken only one or maybe two drugs".

Incident Panel. The Incident Panel (Fig. 4b & c) summarizes the information associated with the reaction from the time the drug was taken until the patient recovered from the reaction. The arrow on the timeline depicts the sequence of the events. The information on timeline reads like a story. That is, for a certain disease, the patient took the drug on a date and after [onset] days the patient experienced the reaction with a serious outcome [hospitalization] and was then treated with [Treatment]. In some cases, the drug was dechallenged and rechallenged to assess its association with the reaction.

Analysis Panel. The Analysis Panel (Fig. 4d) allows analysts to add their comments as well as triage action to a report summary to help them in distinguishing between triaged and untriaged reports as well as help them in reviewing their assessment about the report at a later time **DG3**). The triage actions at this stage are if they will investigate (I) a report further or not (N). The Analysis Panel is placed at the end of the summary following the workflow of analysts, that is, they read a report from left-to-right and end by adding notes to the end of the report when writing a review.

4.2. Design Considerations for Triage Cues Depictions

To design a compact visual summary (**DG1**), some triage cues needed to be made glanceable. Existing research shows that icons and symbols are effective for getting the gist [Mat06]. As only primary cues in our data have a few categories, we design icons using Borgo et al.'s [BKC*13] guidelines to represent them and used text, space, and color to represent rest of the triage cues (**DG2**). Redundant encodings are used to help analysts easily differentiate among various triage cues [BBK*15] and to facilitate multiple analysis strategies. For instance, patient-related cues have the most left spatial position and a green color (Fig. 4).

Overall Visual Encoding for Icons. The primary triage cues contain 2-3 categories, at most. For instance, the primary cue *outcome* has two categories, *serious* and *non-serious*. Therefore, the visual channel with the strongest 'popout' effect to differentiate between these categories is color after position, as position has been used for the overall layout of SumRe. We use color hues to differentiate between categories of a primary triage cue. The colors are chosen based on the semantic meaning of the category. For instance, red represents information that is serious or plausible and needs attention, grey encodes missing information, and blue represents non-serious or implausible. For consistency, similar color encoding is used across all icons. Similarly, the shape channel [Mag14] is selected to differentiate among different primary triage cues. The details on selection of each shape is discussed below.

Designing Onset's Icon. As onset is the duration between the date when the reaction was observed and the date on which the drug was taken, we use a time symbol to represent it. Moreover, onset represents a connection between the drug and reaction so we represent the connection using a semantic encoding of link [Mag14]. Due to poor quality of reporting, sometimes the onset is missing in the report or it may even be implausible, such as depicting that the reaction happened before the drug. Although only domain experts can verify the actual plausibility of the onset as it varies from drug to drug, we consider an onset only to be plausible if it is at least a positive value depicting that the reaction happened after taking the drug. According to one domain expert "For some drugs a certain reaction may appear the same day, while for others it may take months, so the onset really depends on the drug and reaction". Therefore, we only represent if onset is reported or is missing and leave further assessment to the human experts.

Designing Dechallenge & Rechallenge's Icons. Dechallenge means the reaction's disappearance after stopping the drug (a positive fact), while rechallenge means the reaction's recurrence after restarting the drug (a negative fact). Our first designs for these elements included variations of metaphoric icons to represent 'stop' and 'restart' signs to represent dechallenge and rechallenge, respectively. Other design alternatives included using the alphabets 'D' and 'R' with and without background. All these designs were shown to the experts and multiple rounds of discussions led to the final design as depicted in Fig. 5 (b & c). The circles were added to the alphabet to improve their visibility in the presence of other triage cues in SumRe (Fig. 4)

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

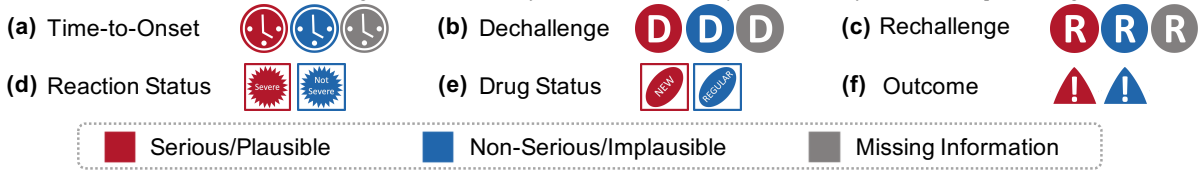

Figure 5: *Icons for the Primary Triage Cues. Red color depicts a serious or plausible cue, blue color encodes non-serious or implausible cues, and grey represents missing information.*

Designing Icons for Drug and Reaction Status. Two of the primary triage cues correspond to domain knowledge about the reaction and drug – both critical for triage. For drug, the status represents if the drug is new or old (regular), and for reaction, status means if it is severe or non-severe (Fig. 5d,e). At the FDA, a drug is considered new for three years after its approval and being in the market. We considered various design options for the drug-status cue including pictures of medicine and different shapes of pills. For the reaction-status, metaphoric representations, such as, variation of 'star shapes' to depict issue or negative effect were preferred by experts as compared to the alphabetic representation (A for ADR). After many design iterations, we selected the final design as depicted in Fig. 5(d,e) with the surrounding rectangle used to enhance their popout effect when displayed along-side the secondary cues (Fig. 4).

Designing Outcome's Icon. The outcome of a reaction has two values: serious and non-serious. We designed multiple variations of metaphoric symbols for 'danger' and 'alarming' signs using a variety of shapes (circular, triangular) as well as using alphabetical representations (S and NS). Based on the feedback by the experts, the triangle symbol (Fig. 5f) is selected as compared to a circle to avoid confusion with other circular icons, i.e., icons for onset, dechallenge and rechallenge.

Representing Missing Information. As missing information is key in report quality assessment, we use a grey color to represent missing information in SumRe for both primary and secondary cues. In informal pilot studies, we observed that participants had difficulty perceiving the amount of grey presented and were unconsciously biased towards the missing secondary cues as compared to primary cues, due to the small size of icons. We conducted followup pilots using different shades of grey for both primary and secondary cues. The final shade that was best perceived is presented in Fig. 4.

5. Crowdsourced Evaluation of SumRe with Medical Experts

SumRe is designed to support effective triage of incident reports by providing a visual summary of each individual report. To evaluate the design of SumRe, we have conducted a controlled user study to assess its usefulness in reports triage in comparison with the current techniques used for triage of these reports at the FDA.

5.1. Overall Study Design

5.1.1. Participants

Newly developed crowdsourcing platforms such as Prolific [PS18] provide means for participants to self-report occupation and expertise on the platform itself. This presented an opportunity to evaluate SumRe in a crowdsourced setting, meeting the dual goal of maintaining social constraints at the time of the study (COVID'19) and evaluating whether platforms such as Prolific provide relevant expertise for domain-focused visualization evaluation. (We note that FDA analysts were not available during the study period, which was conducted during the pandemic.)

We recruited 20 medical experts (5 male, 10 female, 5 unspecified) using Prolific filter criteria to target participants who identify as Doctors, Nurses, Pharmacists, and Emergency medical workers. Recruited participants included medical students, and retired medical professionals demonstrated that they were familiar with drugs and their related adverse reactions, particularly judging from their free-response answers. Based on completion times (around 55 minutes) in pilot experiments, each participant was paid over $10.00 exceeding US Minimum Wage. All participants viewed an IRB-approved consent form.

5.1.2. Study Design

The experiment followed a within-subjects design where each participant performed the tasks using both conditions (layouts) to minimize the effects of participant's expertise and domain knowledge. Both conditions and datasets were balanced across participants.

5.1.3. Baseline Condition

Analysts at the FDA currently use a Tabular layout (Fig. 2) to skim the structured information of a report to make a decision on whether to dive into reading the actual report and more deeply investigate the narrative. We use a Tabular Layout similar to (Fig. 2) as a baseline to reflect this current triage workflow. The order of the data elements in the Tabular layout follows the same order used in current tool at the FDA for triage (Fig. 2). We added two columns appended to the end of the Table, one for adding comments and other for adding the proposed triage action. However, in practice, the current Tabular layout at the FDA (Fig. 2) does not capture any annotations and it is read only. For consistency and sake of a fair comparison, we present similar information in both the Tabular layout and SumRe.

5.1.4. Interaction Design

SumRe is designed to communicate a gist of all triage cues using visual encodings, while addressing limitations inherent in the analysts' current Tabular layout. For example, the current Tabular layout requires analysts to scroll to view all data related to a particular report. In this way, eliminating interaction is also part of the SumRe design. One interaction pattern that was important to support is accessing the underlying report text, which analysts sometimes need to fully investigate the relevance of a particular report. With the Tabular view, accessing the underlying text is accomplished by a link to the report in one of the table entries. In SumRe, we provided similar functionality by including ellipses (...) in various places that brought up relevant portions of the underlying report, and enabled analysts to view the entire report if needed. Other interactions such as sorting and grouping may be useful in triage practice. However, we did not include such interactions in this study design to facilitate controlled comparison between SumRe and the Tabular layout.

5.1.5. Datasets

For our datasets, we used FDA Adverse Event Reports from 2014-2019 [FA15]. As the study was within-subjects, we curated two data sets, each containing a total of eight reports. This small set is a representative of the FDA workflow in that, during our requirement gathering sessions, we observed that an analyst would have reports within a filtered set varying from as few as 3 to 10 or more in a

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

No	Tasks	Description
T1	Triage	Select triage action by identifying reports that are or are not indicative of investigation.
T2	Triage	Identify reports that are the most worthy of investigation.
T3	Triage	Identify reports with the most complete information.
T4	Getting the gist	What happened to the patient after the serious outcome in the report with [X] search criteria?
T5	Getting the gist	What common pattern do you observe about [X] cues in reports with [Y] search criteria?
T6	Getting the gist	Contrast report [X] and report [Y].
T7	Exploration	Explore the reports freely and report on your findings. Is there anything surprising or interesting?

Table 1: *Task Questions and Description*

single "unit" of work. Another consideration is fatigue, in that too many reports may drive up study completion time. Based on pilot studies with analysts, we found that a report takes ~2 minutes to triage, and used this finding to inform the overall study length.

For both datasets, six reports were about one drug, 'Ampyra' and 'Harvoni', respectively, while two reports in each set were about other drugs. This reflects the domain workflow, where analysts may come across important reports not related to the drugs they are responsible for, yet their identification is crucial. Thus, we added two reports related to other drugs to evaluate the analysts' capability of "serendipitous" discovery following a similar task used in network security analysis [TRYB07].

To reflect the ratio of reports with incomplete to complete information received by the FDA, half of the reports were chosen with missing information, while the other half had complete information. We consider a report complete if it contains more than 80% of triage cues following [BNL14] guidelines, and others as incomplete. We selected one report as indicative of investigation in each of these complete and incomplete subsets, verified by domain experts. In practice, the ratio of complete to incomplete and important to non-important is small, but for study purposes we kept it balanced.

For both SumRe and the Tabular baseline, we used structured information from the reports such as drugs and reactions as well as manually extracted missing data points from the text such as dechallenge and rechallenge. The current workflow at the FDA follows a similar manual practice, while ongoing machine learning research efforts underway to automatically extract this information [WQK*17, WQK*20] can be utilized in future generation systems.

5.1.6. Procedure of the Study

After completing the consent form, participants were presented with two video tutorials; one for demonstrating the study task and other for the layouts, followed by a guided tour of both layouts. Participants were allowed time to practice with each of the layouts. They were given two mock tasks with multiple-choice questions to ensure their understanding of the layouts and tasks. After accurate completion of the mock task with as many attempts as possible, participants started the study with the Triage task followed by a set of overview and exploration questions. After completing the Triage task twice each with one of the two layouts, participants were asked to fill out a demographic questionnaire and qualitative survey to provide feedback on both layouts as well as their preferred layout. Thereafter, participants performed the Recall task. A help reference for the concepts used in the study such as factors that make a report "indicative" or "not" of investigation was provided in each layout.

5.1.7. Study Tasks

We designed the study tasks to be reflective of the triage tasks performed by the analysts at the FDA. Although, SumRe is the visual representation of a single report, but practically set of reports filtered for a particular drug or reaction are examined one by one to assess if it requires further investigation. This is synonymous to the webpage previews [AKG*10] such as Google search results displayed as a set, while each preview represents a summary of an individual webpage for the users to take a triage action of whether they will open the page.

Triage Tasks The goal of the triage task was to assess the participants' performance in discerning the gist of a given set of reports to not only identify concerns but also grasping the content of the reports with both layouts. Participants were asked to put themselves into the role of a drug safety analyst who needs to analyze each report and decide whether the report demands further investigation by taking the respective triage action. Once the triage action was captured for all reports (T1), participants were asked to complete the tasks in Table. 1. The tasks T1-T3 were to assess their performance in accurately taking a triage action, identifying safety issues, and assessing reports quality, all crucial for incident reports triage.

As SumRe provides an overview of an incident report, we also include low-level tasks (Table 1, T4-T6) involved in getting the gist of a report to assess participant's performance in searching, understanding, and comparing reports information [HH11,BBB*18]. We ask participants to explore reports (T7) following low-level tasks for overview visualization [HH11] to identify interesting patterns.

Recall Task. Recall is important in reports triage as analysts review reports regularly for a screened issue and come across safety issues that they may have seen earlier, which could vary from minutes, to days, to weeks. This could be helpful in escalating an issue for investigation if they were to encounter it a second time. We include a short-term retention task to assess the effectiveness of SumRe and follow the tasks designed by Bateman et al. [BMG*10]. The study of longer retention could be a topic for future work.

Participants were not told about this task during the instructions in the beginning of the study to prevent intentional learning. After completing the tasks (Table. 1) for both layouts, participants filled the demographic and qualitative survey to clear their visual and linguistic memory before starting the recall task. For each layout, we presented equally blurred summaries of two reports to the participant, including some that were indicative and others they were non-indicative of investigation from the triage task. The order of the layouts and datasets followed the same order from the triage task to ensure similar duration between recall and the triage task. For each layout, participants were asked to recall and report as much information about the two reports as possible.

5.1.8. Measurements

We collect both *q*ualitative and quantitative measures throughout the study. For each trial, we capture the *s*tart and *e*nd time, this

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

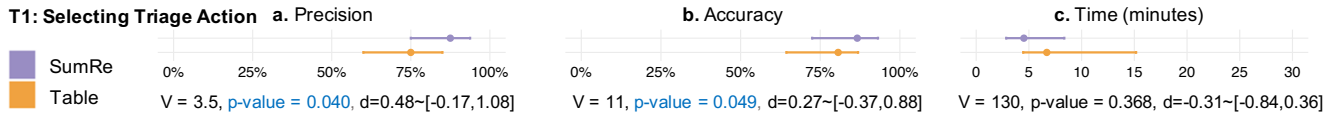

Figure 6: *Triage task to select an action (Investigate or not) on each individual report. SumRe has significantly high precision, i.e., correct selection of reports indicative of investigation.*

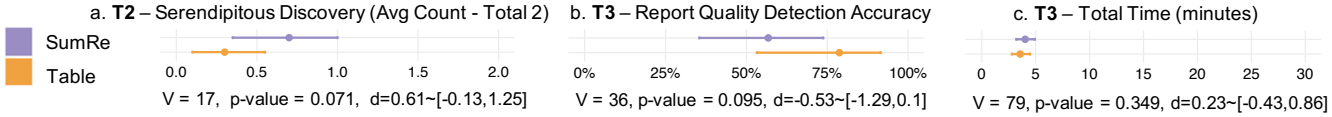

Figure 7: *(a). Average count of identified unrelated issues. (b,c). Overall accuracy and time for identifying if reports are complete. Although insignificant, but the Tabular layout outperforms SumRe.*

allows us to evaluate the average time spent on tasks for each condition. During the task, we measure time spent on triaging the reports and answering the questions, submitting the answers, participants' confidence in the submitted answer, and the perceived difficulty of the task, all on a 7-point Likert scale. Through free-response questions, we also collect qualitative answers and feedback for each task which help us assess participants expressiveness and reasoning about the report information. We also collect demographics and free-response feedback on the overall summary and tasks design. Following [SES16], we also asked participants to select their preferred layout for future tasks and provide a reason for their preference. When calculating correctness, we use non-binary rules that map to a 0–1 scale, corresponding to key "parts" of an answer. For example, we give 0.5 points for an answer that contains only the severity of the reaction, if the task asked for identifying severity and count. Additional details on the scoring method are in the supplement.

5.1.9. Pilots, Analysis, and Experiment Planning

We conducted several pilot studies to evaluate system usability, data collection, tasks, measures, and clarity of our procedure. Due to the limitations of null hypothesis significance testing, we base our analysis on best practices for fair statistical communication in HCI [Dra16] by reporting confidence intervals and effect sizes following APA guidelines. We compute 95% bootstrapped confidence intervals [Cum13] and effect sizes using Cohen's d to indicate a standardized difference between two means. For each task, we display in our figures the accuracy and time results in the form of CI, along with p-values from a paired Wilcox Signed-Rank Test.

5.1.10. Hypotheses

We developed a set of hypotheses to assess how the two summary layouts would compare for different types of tasks. We present the hypotheses below and later use them to discuss our results.

H1: Efficient Reports Triage. Analysts should be able to triage reports quickly and accurately with SumRe due to the compact layout and the visual encodings designed to highlight and differentiate between different triage cues.

H2: Serendipitous Discovery. Analysts may identify unrelated safety issues with SumRe due to the visual encodings to highlight primary triage cues.

H3: Report Quality Assessment. Analysts will identify reports with more missing information more accurately with SumRe due to an explicit encoding of missing data.

H4: Accurate Gist Detection. Analysts will perform better in getting the gist of the reports with SumRe due to the spatial alignment of patient and incident related information in SumRe.

H5: Insight Generation. When freely exploring the reports, analysts will gain different types of insights due the different structures and emphasis of both summary designs.

H6: Better Information Recall. Analysts would be able to recall more items with SumRe due to the memorable nature of glyphs and visual cues.

H7: Triage User Experience. Overall analysts will report a positive experience when completing the tasks with SumRe.

5.2. Study Results

We report on the results of the study conducted with 20 participants (in Prolific, an additional 14 started but returned the study before completion). We group the results based on our hypotheses.

5.2.1. Reports Triage

Shown in Figure 6.b, participants had a relatively higher accuracy in selecting the triage action (T1) with SumRe ($M = 0.87\ [0.72, 0.93], d = 0.27 [0.37, 0.88]$) as compared to the Tabular layout ($M = 0.81\ [0.64, 0.87]$). Although, the practical effect size is small when comparing groups overall, we note that the experiment was within-subjects, so the differences found were based on comparing within participants. In addition, qualitative analysis of the triage action indicates that participants were able to correctly identify the reports indicative of investigation, as verified with the precision measure (Fig. 6.a), SumRe ($M = 0.88\ [0.74, 0.94]$), the Tabular layout ($M = 0.75\ [0.59, 0.86], d = 0.48\ [-0.17, 1.08]$). For the reports not indicative of investigation, the spread in mistakes was similar with both layouts.

There was little difference in participants' time spent on taking the triage action (Fig. 6.c) with SumRe ($M = 4.5\ minutes\ [2.8, 8.3], d = -0.31\ [-0.84, 0.36]$) compared to the baseline ($M = 6.7\ minutes\ [4.5, 15.1]$). Results did show that the help page was accessed more frequently under table ($M = 1.9\ [0.9, 2.9]$) as compared to SumRe ($M = 0.8\ [0.3, 1.3], d = -0.6\ [-1.18, 0.03]$). The help page contained information on triage criteria and general guidelines for assessing the report's importance (primary triage cues, quality of information, etc.). This partially supports (H1). While triage with visual layout is not more efficient in terms of time, it is comparatively accurate, particularly in identifying reports that are indicative of investigation.

5.2.2. Serendipitous Discovery

Shown in the Figure 7a, participants identify a similar number of safety issues with SumRe ($M = 0.7\ [0.35, 0.95], d =$

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

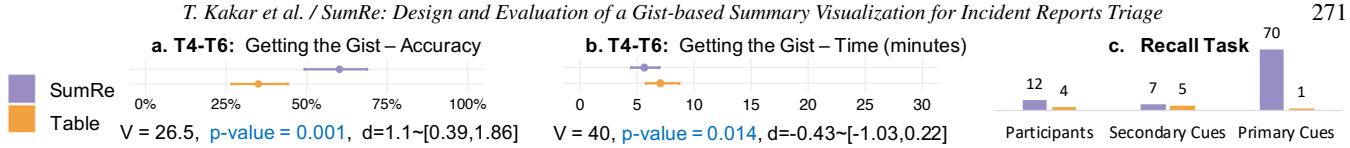

Figure 8: *(a,b) Overall accuracy and time for getting-the-gist tasks (T4-T6). SumRe significantly outperforms the Tabular layout in both accuracy and speed. (c) Count of participants able to recall information along-with type of information recalled.*

0.61 $[-0.13, 1.25]$) as compared to the Tabular layout ($M = 0.3$ $[0.05, 0.57]$). Serendipitous discovery is important in reports triage to ensure safety issues are not missed, even if it is unrelated to the drugs analysts are monitoring currently. Overall, with the Tabular layout, 5 participants were able to identify unexpected issues as opposed to SumRe with 11 participants. Qualitative analysis indicates that participants were able to easily identify the primary triage cues despite of missing information in these reports. On the other hand, with the Tabular layout primary cues are not prominent and hence participants mostly focused on the missing information in these reports considering them as non-indicative of investigation. Ultimately, the non-significant result means H2 is inconclusive.

5.2.3. Identification of Report Quality

For this task (T3), participants spent relatively more time on questions with SumRe ($M = 4.02$ $[3.2, 4.9]$) as compared to the Tabular layout ($M = 3.56$ $[2.79, 4.48]$). Participants also had comparatively better accuracy in identifying the report quality with the Tabular layout ($M = 0.79$ $[0.53, 0.91]$) than SumRe ($M = 0.57$ $[0.35, 0.74]$). Both of these differences are insignificant (Fig. 7b,c). Qualitative analysis of selecting the wrong reports as complete indicates that in SumRe participants were not able to accurately detect the missing information for the patient related triage cues and the dechallenge and rechallenge icons. This might be due to the variable shapes and sizes of these cues making it difficult to assess which information is present, as compared to the Tabular layout, where all triage cues look similar and missing information is represented as an empty cell. These results do not support the hypothesis (H3).

5.2.4. Getting the Gist of Reports

The overall accuracy and time for all three tasks (T4-T6) are shown in (Fig. 8a,b). There was a significant difference in overall accuracy between SumRe ($M = 0.58$ $[0.48, 0.68], d = 1.1$ $[0.39, 1.86]$) and the Tabular layout ($M = 0.34$ $[0.25, 0.43]$). Participants also took less time with SumRe ($M = 5.6$ $[4.5, 6.95], d = -0.43$ $[-1.03, 0.22]$) as compared to the Tabular layout ($M = 7.05$ $[5.8, 8.7]$). This supports the hypothesis (H4) that visual cues help in quickly locating and extracting information from reports. For the comparison task (T6), there was no significant difference between the two layouts, that is, SumRe (0.53 $[0.44, 0.68], d = 0.29$ $[-0.41, 0.86]$), and baseline ($M = 0.47$ $[0.38, 0.55]$).

5.2.5. Insight Generation

T7 instructed participants to freely explore the data and report on any insights they derived from their exploration. To analyze these responses, we performed qualitative coding of the responses following the guidelines provided in [SND04]. We consider one observation about the data as an insight and do not count general comments toward insights, such as, 'hard to read information in this layout'.

For SumRe we received a total of 24 insights of which 21 were distinct, while for the Tabular layout, 23 insights of which 15 were distinct. We consider insights duplicate if two insights are discussing the same facts, for instance, "I was shocked at the lack of useful information in some", and "There is a lot of missing information". We categorized insights into a set of codes that were derived by an initial open coding of the data.

Two types of insights were markedly more common in the SumRe layout: **Report-level**, and **Unexpected**. We categorize an insight as **Unexpected** if it has not been part of the answers in the previous tasks and is an observation about the data, such as, *"Two of the patients had mental problems and they may have taken the drug incorrectly"*, and *"I'm sure that there are other things to mention but the one that immediately popped up was in R5, Cardiac Arrest is described as non severe ADR? That's odd! "*. In the first insight, the participant is reasoning about patient's capability of administering the drug correctly, while in the second, the reaction status of cardiac arrest is being questioned. FDA does not consider cardiac arrest a severe reaction due to its high background rate making the assessment of its association with the drug difficult.

Report-level insights were observations about multiple triage cues in a single or multiple reports. For instance, *"The reports submitted by pharmacists are surprisingly undetailed, they lack further information about the patient and nature of the ADRs which would be helpful in the triage process. Being a pharmacist myself I thought that the reports submitted by pharmacists would at least contain more information about the patient's concomitant medications"*, and *"It was interesting that so many of the very long adverse reactions reported had an implausible onset"*. These insights are considering multiple attributes, such as patient information and reporter in the first example, and reactions and onsets in the second.

For the Tabular layout, the most common insights were categorized as **Attribute-Level** and **Guided**. Attribute-level insights are those that are focused on a single triage cue or attribute. For instance, *"5 reporters were other healthcare"*, and *"There appears to be more men than women taking part"*. The first example is focused on the frequency of a certain reporter type (Other Healthcare), and the second is about the gender.

On the other hand, **Guided** insights were those that have been observed during completing the previous tasks. Such as, *"Also the information in some reports was not much and it did not help me..."*, and *"The most plausible adverse reactions to Harvoni were seizure, hypotension, and acute renal failure so pretty serious"*. The first insight is pointing towards the quality of information, one of the questions from the Overview task. The second insight indicates the severity of the reactions which was part of the triage criteria. A possible explanation for having more attribute-level insights with the Tabular layout could be that column data can be viewed and compared easily in the Tabular layout due to visual proximity. On the other hand, SumRe is designed to make all the report data accessible, resulting in more Report-level insights with SumRe. Consequently, we consider hypothesis H5 that the interfaces would lead to different insights to be supported.

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

5.2.6. Information Retention (Recall)

For the recall task, 4 out of 20 participants were able to recall information with the Tabular layout, resulting in a total of 6 triage cues, including 1 primary and 5 secondary cues. With SumRe, 12 participants were able to recall a total of 77 triage cues consisting of 70 primary and 7 secondary cues (Fig. 8c). Responses were scored based on specific answers. Because of the blurred images, the report text was not readable. However, with SumRe, due to the redundant encodings, that is, spatial alignment and use of other visual channels, the icons for primary cues could be interpreted, resulting in a high count of information retention (H6). The high recall due to redundant encodings aligns with the findings from [BBK*15].

5.2.7. Preference and Qualitative Feedback

For participant's preference we followed the task from [SSK16] and after the completion of Triage task, asked participants to select their preferred layout if they were to perform another triage task. 90% of the participants preferred to use SumRe layout for Triage tasks, considering it faster, easier, appealing, engaging, and enjoyable (H7). This is also verified by significant differences in participant's overall perceived ease and confidence ratings for all the tasks (see supplement). Some of comments by participants in the favor of Visual layout included *"The visual layout was more appealing to use and I enjoyed the work more, it felt less like monotonous work and more like a pattern recognition game almost. It was more user friendly", "The visual layout made it significantly easier and faster to perform the tasks. Moreover, it was less straining for my eyes compared to the table"*. For the Tabular layout, participants' remarks included *"For me Table was easy to use and see the ADRs"*.

6. Discussion

Taken together, results from these design and evaluation efforts suggest that SumRe outperforms the Tabular layout in "gist"-based reasoning as well as identifying reports that need investigation (based on detail assessment). Both of these tasks are crucial for identifying critical issues in drug safety. These findings may hold implications more broadly for visualization efforts targeting triage workflows and semi-structured text.

6.1. Triage Design Shapes Analysis, Exploration, and Insights

In our study, participants were more engaged and expressive in their findings and feedback with the visual layout as compared to the Tabular layout. For instance, participants left a total of 23 voluntary comments (avg length 37.5 characters) with the Tabular layout, as compared to 36 with visual (avg length 47.2 characters). We also observed that participants are able to correct their triage actions later in the study due to the visual feedback, that is, color for Investigate and Do not Investigate actions (Fig. 1), while we did not see this behavior with the Tabular layout. This was an expected outcome of the design of our Analyses panel (DG3 - Section. 4), which we did not evaluate exclusively. This could be useful in the real-world scenario to re-assess one's analyses and correct mistakes if needed. We also noticed that the insights provided by participants while using SumRe reflected reasoning about the data beyond the provided facts. This behavior is reflected in prior studies on visualization systems involving insight-based reasoning, *e.g.* Chang *et al.* [CZGR09].

6.2. Design Implications for Triage Workflows

In this work, we decomposed incident triage through task characterization, *e.g.* the development of triage cues and layout models that drive the design of SumRe. By focusing on both the characteristics of the data involved in triage, as well as the triage workflow itself, these efforts could provide a useful baseline for investigating how visualization might improve triage workflows in other domains. For example, the design of symbols and icons to meet the dual goal of compactness and attention management might be leveraged in future designs.

One possibility is drug incident response beyond the US FDA. We designed SumRe based on the Pharmacovigilance workflows at the FDA, however, Pharmacovigilance systems are present in many countries [JA10] who collect and analyze drug incident reports with similar data elements and goals [BNL14]. Similarly, tables and spreadsheets are generally used to display and analyze structured data, particularly, those with many textual attributes such as the FAA analyzes service difficulty reports about aircraft's maintenance issues [MR12]. Our proposed design is an initial step forward towards introducing gist-based visualization for this common type of rich data. Moreover, characterization of the triage task and design of a controlled experiment can be adapted for triage workflows in other domains dealing with incident reports.

Other possibilities focus on drug incident response itself. Suggestions to improve the visual design involve the incident-panel with the timeline which currently displays cumulative information about the drugs and onset, to keep it as similar to the Tabular layout as possible. However, more sophisticated designs to display multiple onsets for multiple drugs while keeping the design compact are possible. For instance, one approach could be adding interactions and overlays with further quantitative information about drugs and adverse events on top of SumRe.

6.3. A Need for Visualization Task-Characterizations for Semi-Structured Text

One main challenge was to design the tasks for the report-level triage due to lack of precedent to draw from. Our data was mostly textual, however, there are a few in-depth evaluations of the proposed approaches in the text visualization community [AL19] which do not fit our single-report analysis. Due to lack of work on designing and evaluating visualization for document-level triage, we adapted tasks from many areas including cybersecurity [TRYB07], digital libraries [BBM*06, Loi12], overviews [HH11], and visualization [BMG*10, SSK16] and aligned them with analysts' workflow.

7. Conclusion

We present a task-characterization, design, and evaluation targeting triage workflows with semi-structured text data in the drug incident response domain. SumRe is a visualization-enabled "gist-based" summary that transforms incident information into a compact visual form following a model of how analysts approach triage, and drawing on visualization design principles to emphasize the cues that analysts were found to refer to most when processing incidents. To evaluate SumRe, we describe a crowdsourced study with medical experts that compares SumRe to a Tabular baseline, with tasks targeting aspects of the triage workflow, assessing participants' experience and performance. The results of these efforts suggest triage workflows can be effectively augmented with visualization-enabled designs, providing new means for analysts to quickly and efficiently process the numerous reports they deal with on a daily basis.

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

References

[AKG*10] AULA A., KHAN R. M., GUAN Z., FONTES P., HONG P.: A comparison of visual and textual page previews in judging the helpfulness of web pages. In *Proceedings of the 19th International Conference on World Wide Web* (2010), pp. 51–60. 2, 7

[AL19] ALHARBI M., LARAMEE R. S.: Sos textvis: An extended survey of surveys on text visualization. *Computers 8*, 1 (2019), 17. 10

[ALK*11] AMERSHI S., LEE B., KAPOOR A., MAHAJAN R., CHRISTIAN B.: Cuet: human-guided fast and accurate network alarm triage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), pp. 157–166. 3

[BBB*18] BLASCHECK T., BESANÇON L., BEZERIANOS A., LEE B., ISENBERG P.: Glanceable visualization: Studies of data comparison performance on smartwatches. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 630–640. 7

[BBK*15] BORKIN M. A., BYLINSKII Z., KIM N. W., BAINBRIDGE C. M., YEH C. S., BORKIN D., PFISTER H., OLIVA A.: Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 519–528. 5, 10

[BBM*06] BADI R., BAE S., MOORE J. M., MEINTANIS K., ZACCHI A., HSIEH H., SHIPMAN F., MARSHALL C. C.: Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (2006), pp. 218–225. 3, 10

[BKC*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (STARs)* (2013), pp. 39–63. 5

[BMG*10] BATEMAN S., MANDRYK R. L., GUTWIN C., GENEST A., MCDINE D., BROOKS C.: Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), pp. 2573–2582. 7, 10

[BNL14] BERGVALL T., NORÉN G. N., LINDQUIST M.: vigigrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. *Drug safety 37*, 1 (2014), 65–77. 4, 7, 10

[Bur11] BUREAU C. F. P.: CFPB. https://www.consumerfinance.gov/, 2011. Accessed: 2020-4-3. 1

[BvHH*16] BÖHM R., VON HEHN L., HERDEGEN T., KLEIN H.-J., BRUHN O., PETRI H., HÖCKER J.: Openvigil fda–inspection of us american adverse drug events pharmacovigilance data and novel clinical applications. *PloS one 11*, 6 (2016), e0157753. 3

[CAS13] CAPRA R., ARGUELLO J., SCHOLER F.: Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (2013), pp. 399–408. 2

[CCP09] COLLINS C., CARPENDALE S., PENN G.: Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 1039–1046. 2

[Cum13] CUMMING G.: *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013. 8

[CWDH09] CHEN Y., WANG L., DONG M., HUA J.: Exemplar-based visualization of large document corpus (infovis2009-1115). *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1161–1168. 2

[CZGR09] CHANG R., ZIEMKIEWICZ C., GREEN T. M., RIBARSKY W.: Defining insight for visual analytics. *IEEE Computer Graphics and Applications 29*, 2 (2009), 14–17. 10

[Dra16] DRAGICEVIC P.: Fair statistical communication in hci. In *Modern statistical methods for HCI*. Springer, 2016, pp. 291–330. 8

[DZG*07] DON A., ZHELEVA E., GREGORY M., TARKAN S., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (2007), pp. 213–222. 2

[FA15] FOOD U., ADMINISTRATION D.: FDA Adverse Event Reporting System (FAERS) data statistics. https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers, 2015. Accessed: 2019-01-11. 3, 6

[FPB16] FELIX C., PANDEY A. V., BERTINI E.: Texttile: an interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 161–170. 1, 2

[GLK*12] GÖRG C., LIU Z., KIHM J., CHOO J., PARK H., STASKO J.: Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics 19*, 10 (2012), 1646–1663. 1, 2

[HH11] HORNBÆK K., HERTZUM M.: The notion of overview in information visualization. *International Journal of Human-Computer Studies 69*, 7-8 (2011), 509–525. 7, 10

[HHT14] HAGGERTY J., HAGGERTY S., TAYLOR M.: Forensic triage of email network narratives through visualisation. *Information Management and Computer Security* (2014). 3

[HVG08] HÄRMARK L., VAN GROOTHEEST A.: Pharmacovigilance: methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology 64*, 8 (2008), 743–752. 1

[JA10] JEETU G., ANUSHA G.: Pharmacovigilance: a worldwide master key for drug safety monitoring. *Journal of Young Pharmacists 2*, 3 (2010), 315–320. 10

[JWS*05] JONKER D., WRIGHT W., SCHROH D., PROULX P., CORT B., ET AL.: Information triage with trist. In *2005 International Conference on Intelligence Analysis* (2005), Citeseer, pp. 2–4. 3

[JXYXJ*15] JIAN-XIANG W., YUN-XIA Z., JUN S., HOU-MING X., MING L., YUE-HONG S.: Adrvis: an information visualization platform for adverse drug reactions. *International Journal of U-and E-Service, Science and Technology 8*, 10 (2015), 139–150. 3

[KHGW07] KUO B. Y., HENTRICH T., GOOD B. M., WILKINSON M. D.: Tag clouds for summarizing web search results. In *Proceedings of the 16th International conference on World Wide Web* (2007), ACM, pp. 1203–1204. 2

[KHXM*15] KASS-HOUT T. A., XU Z., MOHEBBI M., NELSEN H., BAKER A., LEVINE J., JOHANSON E., BRIGHT R. A.: Openfda: an innovative platform providing access to a wealth of fda's publicly available data. *Journal of the American Medical Informatics Association 23*, 3 (2015), 596–600. 3

[KJW*14] KOCH S., JOHN M., WÖRNER M., MÜLLER A., ERTL T.: Varifocalreader—in-depth visual analysis of large text documents. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1723–1732. 2

[KQR*19] KAKAR T., QIN X., RUNDENSTEINER E. A., HARRISON L., SAHOO S. K., DE S.: Diva: Towards validation of hypothesized drug-drug interactions via visual analysis. In *Eurographics* (2019). 3

[KWMJ*15] KARIMI S., WANG C., METKE-JIMENEZ A., GAIRE R., PARIS C.: Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR) 47*, 4 (2015), 56. 3

[Loi12] LOIZIDES F.: *Understanding and conceptualising the document triage process through information seekers' visual and navigational attention*. PhD thesis, City University London, 2012. 10

[LPC98] LAZAROU J., POMERANZ B. H., COREY P. N.: Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Journal of the American Medical Informatics Association 279*, 15 (1998), 1200–1205. 1

[LSL03] LIU H., SELKER T., LIEBERMAN H.: Visualizing the affective structure of a text document. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems* (2003), ACM, pp. 740–741. 2

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

[LZP*12] LIU S., ZHOU M. X., PAN S., SONG Y., QIAN W., CAI W., LIAN X.: Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST) 3*, 2 (2012), 25. 2

[Mag14] MAGUIRE E. J.: *Systematising glyph design for visualization*. PhD thesis, Oxford University, UK, 2014. 2, 5

[Mat06] MATTHEWS T.: Designing and evaluating glanceable peripheral displays. In *Proceedings of the 6th conference on Designing Interactive systems* (2006), pp. 343–345. 5

[MR12] MARAIS K. B., ROBICHAUD M. R.: Analysis of trends in aviation maintenance risk: An empirical approach. *Reliability Engineering & System Safety 106* (2012), 104–118. 1, 10

[NBS05] NEUSTAEDTER C., BRUSH A. B., SMITH M. A.: Beyond" from" and" received" exploring the dynamics of email triage. In *CHI'05 extended abstracts on Human factors in Computing Systems* (2005), pp. 1977–1980. 3

[PS18] PALAN S., SCHITTER C.: Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance 17* (2018), 22–27. 6

[SES16] SAKET B., ENDERT A., STASKO J.: Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (2016), pp. 133–142. 8

[SND04] SARAIYA P., NORTH C., DUCA K.: An evaluation of microarray visualization tools for biological insight. In *IEEE Symposium on Information Visualization* (2004), IEEE, pp. 1–8. 9

[SOR*09] STROBELT H., OELKE D., ROHRDANTZ C., STOFFEL A., KEIM D. A., DEUSSEN O.: Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1145–1152. 2

[SSK16] SAKET B., SCHEIDEGGER C., KOBOUROV S.: Comparing node-link and node-link-group visualizations from an enjoyment perspective. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 41–50. 10

[STKO13] SAKAEDA T., TAMON A., KADOYAMA K., OKUNO Y.: Data mining of the public version of the fda adverse event reporting system. *International Journal of Medical Sciences 10*, 7 (2013), 796. 3

[TRYB07] THOMPSON R. S., RANTANEN E. M., YURCIK W., BAILEY B. P.: Command line or pretty lines?: comparing textual and visual interfaces for intrusion detection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2007), ACM, p. 1205. 7, 10

[VHWV09] VAN HAM F., WATTENBERG M., VIÉGAS F. B.: Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1169–1176. 2

[VW08] VIÉGAS F. B., WATTENBERG M.: Timelines tag clouds and the case for vernacular visualization. *Interactions 15*, 4 (2008), 49–52. 2

[WQK*17] WUNNAVA S., QIN X., KAKAR T., SOCRATES V., WALLACE A., RUNDENSTEINER E.: Towards transforming fda adverse event narratives into actionable structured data for improved pharmacovigilance. In *Proceedings of the Symposium on Applied Computing* (2017), ACM, pp. 777–782. 3, 7

[WQK*20] WUNNAVA S., QIN X., KAKAR T., KONG X., RUNDENSTEINER E.: A dual-attention network for joint named entity recognition and sentence classification of adverse drug events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (2020), pp. 3414–3423. 7

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1221–1228. 2

[YBH14] YILDIRIM P., BLOICE M., HOLZINGER A.: Knowledge discovery and visualization of clusters for erythromycin related adverse events in the fda drug adverse event reporting system. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, 2014, pp. 101–116. 3

© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.