Child & Youth Care Forum
https://doi.org/10.1007/s10566-019-09496-7

ORIGINAL PAPER

Understanding Open Access Data Using Visuals: Integrating Prospective Studies of Children's Responses to Natural Disasters

Hazel J. Shah[1] · Betty S. Lai[2] · Audrey J. Leroux[3] · Annette M. La Greca[4] · Courtney A. Colgan[2] · Julia Medzhitova[2]

Abstract

Background As access to open data is increasing, researchers gain the opportunity to build integrated datasets and to conduct more powerful statistical analyses. However, using open access data presents challenges for researchers in understanding the data. Visuals allow researchers to address these challenges by facilitating a greater understanding of the information available.

Objectives This paper illustrates how visuals can address the challenges that researchers face when using open access data, such as: (1) becoming familiar with the data, (2) identifying patterns and trends within the data, and (3) determining how to integrate data from multiple studies.

Method This paper uses data from an integrative data analysis study that combined data from prospective studies of children's responses to four natural disasters: Hurricane Andrew, Hurricane Charley, Hurricane Katrina, and Hurricane Ike. The integrated dataset assessed hurricane exposure, posttraumatic stress symptoms, anxiety, social support, and life events among 1707 participants (53.61% female). The children's ages ranged from 7 to 16 years ($M = 9.61$, $SD = 1.60$).

Results Visuals serve as an effective method for understanding new and unfamiliar datasets.

Conclusions In response to the growth of open access data, researchers must develop the skills necessary to create informative visuals. Most research-based graduate programs do not require programming-based courses for graduation. More opportunities for training in programming languages need to be offered so that future researchers are better prepared to understand new data. This paper discusses implications of current graduate course requirements and standard journal practices on how researchers visualize data.

Keywords Open access data · Integrated datasets · Visuals · Patterns · Hurricane · Natural disaster

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10566-019-09496-7) contains supplementary material, which is available to authorized users.

✉ Courtney A. Colgan
colganc@bc.edu

Extended author information available on the last page of the article

Published online: 01 March 2019

Introduction

Advancements in technology as well as initiatives towards transparency have resulted in a data revolution supporting open data (Molloy 2011). Access to open data provides opportunities to build integrated datasets and to conduct more powerful statistical analyses (Kitchin 2014). However, when researchers use open access data, unfamiliarity can lead to difficulties in understanding the data (Vis 2013). In order to take advantage of open access data, researchers must learn how to familiarize themselves with the information to be able to resolve these problems. Studies have shown that visuals facilitate a greater understanding of what is available in the data (Simon 2014). This paper focuses on how visuals may help address common challenges that researchers face when using open access datasets.

Establishing open access databases has become a priority for international governments and research funders. Large, open access datasets are increasing in availability (Kitchin 2014). In 2013, the President of the United States issued an executive order to support making government data open and accessible to the public (Executive Order No. 13642 2013). This order was supplemented with the Open Data Memorandum (The White House 2009), which required federal agencies to abide by an Open Data Policy. This policy defined data as an asset and required it to be "available, discoverable, and usable" (The White House 2009). At the same time, governments around the world began defining and implementing open data strategies to increase access to data (Huijboom and Van Den Broek 2011).

Research agencies are dedicated to improving the accessibility and reuse of their data. In compliance with the Open Data Memorandum, the National Science Foundation (NSF) has committed to expanding, enriching, and opening existing and future agency data (National Science Foundation 2018). NSF prioritizes the creation of open access datasets when allocating funding (National Science Foundation 2018). Additionally, the National Institutes of Health (NIH) has increased availability of data generated from funded health research (Walport et al. 2011). NIH supports 73 multidisciplinary data repositories that provide access to data for reuse (National Institutes of Health 2017). Considering the multitude of open access databases supported by national institutions, the availability of open access data is growing.

However, researchers encounter various challenges when using open access data, and this paper addresses three key challenges. *First, researchers often have limited background knowledge of the new data.* As the collective pool of open access data is increasing, understanding data is becoming both more important and more challenging (Kitchin 2014). Increased access to data allows researchers to answer questions and conduct studies that were previously inconceivable (Skiba 2014). However, researchers must first become familiar with the data in order to discern what questions to ask (Harwood and Mayer 2016). Visuals may be used to better acquaint researchers with new datasets by more effectively summarizing the data.

A second challenge for researchers is the need to identify patterns and trends within an open access dataset. If the dataset is large or if the researcher combines multiple smaller datasets, then identifying trends and relationships between variables becomes difficult. Visuals can be used to identify these patterns and are also beneficial for understanding the composition of a large or unfamiliar dataset (Liu et al. 2017). Visuals can replace mental calculations and enhance understanding as researchers are able to infer patterns among the data at a glance (Heer et al. 2010). Researchers are able to use plots to identify trends within and among variables. Further, visuals can serve as diagnostic tools in the early detection of errors in the data, prior to conducting statistical analysis (Keim 2002). For

example, visuals may be used to examine distributions of variables, plot descriptive statistics, and check for potential violations of statistical assumptions.

A third challenge for researchers may be the desire to integrate data from an open access data base with their own study or other open access studies in order to conduct an Integrative Data Analysis (IDA). Visuals may serve as a tool to understand how these datasets will integrate across studies. Integrated datasets combine individual-level data from multiple resources to form a pooled collection of data (Lenzerini 2002). Such datasets provide a strong foundation for building a cumulative knowledge base (Curran and Hussong 2009). Pooled data are more flexible, powerful, and diverse (Brincks et al. 2018; Maxwell 2004). Compared to meta-analysis, IDA allows researchers to use the combined data to estimate new parameters as well as to examine outcomes based on a broader range of risk factors (Brincks et al. 2018). Researchers are able to test whether a set of findings are consistent across multiple independent samples without having to conduct additional studies (Curran and Hussong 2009). Integrated datasets allow researchers to assess the influence of differences in sampling, geography, assessment methods, and measurement on the replicability of the data (Curran and Hussong 2009). Combining analogous datasets from open access databases is valuable in analyzing the replicability of studies by focusing on between-study differences.

Understanding what datasets to combine and how to integrate the studies can be accomplished through the use of visuals. Visuals provide an efficient method of comparison by enhancing the ability to comprehend and process large-scale collections of data (Chen 2017). Visualizing the data will help determine how a study compares with other independent studies and how it can be combined to form a unified dataset.

Accordingly, the objective of this paper is to illustrate the value of visuals in addressing three key challenges of using open access datasets. Using an example dataset that integrated multiple studies, the paper discusses how visuals can address the following challenges:

1. Becoming familiar with the data.
2. Identifying patterns and trends within the data.
3. Determining how to integrate data from multiple studies.

Example Data

The example dataset was created for a National Institute of Mental Health grant, Award #1R03MH113849-01 (Principal Investigator Betty Lai, Ph.D.; Co-Principal Investigator Annette M. La Greca, Ph.D.), referred to as the Child Disaster Data Integration (CDDI) project. The CDDI project used an integrated dataset that combined data from four prospective studies of children's responses to a hurricane. These hurricanes were: Hurricane Andrew (1992; La Greca et al. 1996, 2013a), Hurricane Charley (2004; La Greca et al. 2010), Hurricane Katrina (2005; Lai et al. 2015a, b; Self-Brown et al. 2014) and Hurricane Ike (2008; La Greca et al. 2013b; Lai et al. 2013, 2014). Each study examined children's trajectories of psychological distress following the hurricane. The integrated dataset included individual data for posttraumatic stress symptoms (PTSS) and risk factors in the domains of child characteristics, exposure, loss/disruption events, stressors, and social support. The purpose of the CDDI project was to integrate data across four studies in order to determine trajectories and associated risk factors of children's mental health outcomes

following devastating hurricanes. Use of these datasets for this study was considered exempt by the Georgia State University Institutional Review Board under exemption 7.5.4.

Method

Participants

The integrated dataset contained 1707 participants (53.61% female) from the four studies ($n_{\text{Andrew}} = 568$, $n_{\text{Charley}} = 384$, $n_{\text{Katrina}} = 426$, $n_{\text{Ike}} = 329$). Students were recruited from elementary and middle schools in Florida, Louisiana, and Texas that were in close proximity to each hurricane's path of destruction. Participants' ages ranged from 7 to 16 years ($M = 9.61$, $SD = 1.60$) at their baseline assessment.

Study Design

The four individual studies were approved by the University of Miami or Louisiana State University Institutional Review Boards. Children were recruited for the studies through letters that were sent home to their parents. Active parental consent and written child assent were required for study participation in all four hurricane studies. Questionnaires were verbally administered in group settings as children followed along and marked their responses. Research assistants were available in the room to help facilitate and answer questions. Baseline data collection for each hurricane study occurred between 3 and 9 months after the hurricane made landfall. Subsequent assessments ranged from 7 to 26 months post-hurricane.

Time

Timepoints were established as the number of months post-hurricane that an assessment occurred for each hurricane study. The presence of a participant at any point of data collection was indicated through a series of time buckets, formatted as dummy codes (e.g., TIME03 = 1 or 0). Participants were considered present if they had non-missing data for at least one measure-related or non-demographic-related question at the given timepoint. Assessment timepoints were determined using existing documentation based on the original hurricane studies. In instances where assessment periods ranged over multiple months, the midpoint of that range was used to designate the time of data collection. The final combination of time buckets included data from 3, 5, 7, 8, 9, 10, 14.5, 15, 20.5, 21, and 26 months post-hurricane.

Measures

The four individual studies assessed post-hurricane reactions in children. Hurricane exposure and PTSS were measured in all four hurricane studies. Anxiety, social support, and life events were only assessed in some of the studies (as described below).

Hurricane Exposure

The Hurricane-Related Traumatic Experiences (HURTE; Vernberg et al. 1996) and Hurricane-Related Traumatic Experiences—Revised (HURTE-R; La Greca et al. 2010) questionnaires were used to assess children's perceived and actual life-threat during the hurricane as well as immediate and ongoing loss/disruption in the months post-hurricane. All four hurricane studies used either the HURTE or HURTE-R questionnaire to measure exposure to hurricane-related stressors. Both questionnaires were found to be reliable; the Cronbach's alpha was 0.98 for the HURTE (Spell et al. 2008) and 0.73 for the HURTE-R (Danzi and La Greca 2016).

Perceived/Actual Life-Threat

Perceived life-threat was assessed with a single item (i.e., "At any time during the hurricane, did you think you might die?"), while actual life-threat was measured via six Yes/No questions (e.g., "Did you get hurt during the hurricane?"). An actual life-threat summary score was calculated by summing the Yes responses of the six items, resulting in a possible total score range of 0–6.

Loss/Disruption

Exposure to immediate loss/disruption after the hurricane was measured using ten Yes/No questions (e.g., "Was your home damaged badly or destroyed by the hurricane?"), while ongoing loss/disruption since the hurricane was measured using six Yes/No questions (e.g., "Has almost all the damage to your house from the hurricane now been fixed?"). Three of the six ongoing loss/disruption items (items 1, 2, and 6) were reverse-coded to accurately depict loss/disruption. Summary scores for both immediate and ongoing loss/disruption were calculated by summing the items within each subscale, creating possible summary score ranges of 0–10 and 0–6, respectively.

Posttraumatic Stress Symptoms

The Posttraumatic Stress Disorder Reaction Index (PTSD-RI; Frederick 1985; Steinberg et al. 2004) evaluated children's PTSS using diagnostic criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM; Steinberg et al. 2004). Data collected for the Hurricane Andrew study used a 20-item questionnaire based on the DSM-III-R diagnostic criteria to measure children's reactions to the hurricane (e.g., "Do you get scared, afraid, or upset when you think about this event?"). Each item measured the frequency of PTSS in the months following the hurricane with scores of 0–4 (0 = none of the time, 2 = some of the time, 4 = most of the time). The Hurricane Charley, Hurricane Katrina, and Hurricane Ike studies used a 22-item version of the PTSD-RI (Revision 1) using DSM-IV-TR (Text Revision) criteria to assess 17 PTSD symptoms. Although the questions using the DSM-IV-TR were formatted slightly differently (e.g., "When something reminds me of what happened, I get very upset, afraid or sad"), responses to the items were measured on the same 0, 2, 4 scale as the DSM-III-R.

A total severity score was calculated for the DSM-IV-TR-based PTSD-RI using 18 items based on the 17 symptoms. Only the larger score between item 10 (i.e., "I have

trouble feeling happiness or love") and item 11 (i.e., "I have trouble feeling sadness or anger") was used; therefore, the severity score totaled the scores of 17 of the 18 items. The total possible score ranged from 0 to 68, with higher scores indicating greater PTSD severity. The DSM-IV-TR criteria also established subscales within the PTSD-RI for Arousal, Re-experiencing, and Avoidance. Of the 17 items, five symptoms were classified as Arousal, five were classified as Re-experiencing, and seven were classified as Avoidance. In addition to the total severity score, a truncated summary score was calculated using 10 items that were congruent between the DSM-III-R and DSM-IV-TR criteria for PTSD, resulting in a truncated severity score range from 0 to 40. The truncated summary score was generated for all four hurricane studies to facilitate comparability of the PTSD-RI data. The seven items from the total summary score that were not used for the truncated summary score included two Arousal items, one Re-experiencing item, and four avoidance items.

Previous studies have shown that the DSM-III-R and DSM-IV-TR versions of the PTSD-RI have high test–retest reliabilities, scoring 0.94 (Pynoos et al. 1987) and 0.84 (Steinberg et al. 2004), respectively. Internal consistency for the DSM-III-R Reaction Index used in the Hurricane Andrew data was high with a Cronbach alpha value of 0.89 (Vernberg et al. 1996). Internal consistency for the DSM-IV-TR Reaction Index varied for the individual studies, with Cronbach alpha values of 0.83 (La Greca et al. 2010) 0.92 (Lai et al. 2015b), and 0.88 (Lai et al. 2013).

Anxiety

The Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds and Richmond 1997) measured the degree and quality of anxiety experienced by children and adolescents post-hurricane (Reynolds and Richmond 2008). The inventory assessed the presence of anxiety-related symptoms using a 28 Yes/No item questionnaire (e.g., "I worry when I go to bed at night"). A total summary score was calculated by summing the Yes responses of the 28 items, with a potential total score range of 0–28. With an internal consistency Cronbach alpha value of 0.86 (La Greca et al. 2013a, b) the RCMAS was used to assess anxiety in the Hurricane Andrew and Hurricane Ike studies.

Social Support

The Social Support Scale for Children (SSSC; Harter 1985) measured the degree of support or positive regard offered to children by their parents, teachers, classmates, and close friends. Children responded to the 24-item measure by choosing between two alternatives (e.g., "Some kids have parents who don't really understand them but other kids have parents who really do understand them") and then selecting the degree of accuracy by indicating if the statements were "really true" or "sort of true" for them. Each item was scored from one to four, with a higher score representing higher levels of support. The sum of the 24 items was used as a total summary score for the measure, with values ranging from 24 to 96. The Hurricane Andrew, Hurricane Ike, and Hurricane Katrina studies used the SSSC to determine the level of social support provided to children. Internal consistency within the SSSC varied by hurricane study, with Cronbach alpha values ranging from 0.73 (Self-Brown et al. 2013) to 0.88 (La Greca et al. 2013a).

Life Events

The Life Events Schedule (LES; Johnson 1986) assessed what major life events a child experienced in the months following a traumatic event (i.e., the hurricane). A shortened version of the LES was administered, which included 14 Yes/No items regarding the occurrence of major life changes (e.g., "Death of a parent," "Birth of a sibling," "Divorce of parents," "Hospitalization for illness or injury"). The total number of major life events was summed across the 14 items for a summary score range of 0–14. The Hurricane Andrew, Hurricane Charley, and Hurricane Ike studies used the LES to determine the number of major life events experienced by children post-hurricane. Previous studies demonstrated the test–retest reliability of the LES measure to be 0.72 when used with children (Greenberg et al. 1983; Romero et al. 2009).

Visualization Approach

For the CDDI project, the individuals were assessed at varying timepoints post-hurricane. Consequently, the measures for each individual study were reformatted to reflect the exact number of months post-hurricane that the assessment occurred. Such inherent identifiers were essential in understanding the various layers of information that were provided by an integrated dataset. These layers contributed to the value of visuals as the graphs and plots conveyed a multitude of information.

All visuals were created using the RStudio 1.1.423 software program (RStudio Team 2018) and required installation of the ggplot2 package (Wickham 2009). Downloading and installing R 3.4.3 (R Core Team 2017) was also required for use of the RStudio interface. Both of the open-source license programs for R and RStudio Desktop were free to download. The list of variables used in the R code and the R code for creating the visuals are included as part of the online supplemental materials (Online Resource 1). The online supplemental materials also include color analogs of the black-and-white figures featured in this article (Online Resource 2), along with the R code used to develop the color analogs (Online Resource 1).

The following sections outline how visuals are used to address the three challenges faced by researchers. After identifying the challenge with respect to the CDDI dataset, the paper describes how visuals were created to address each challenge. The insights gained from generating these visuals are then discussed for each challenge. All of the visuals are provided as figures. The number associated with each figure is indicative of the challenge with which it is associated (e.g., Fig. 1a is used to illustrate the first challenge).

Challenge 1: Becoming Familiar with the Data

Becoming familiar with a new dataset involves understanding the unit of analysis; for the CDDI dataset, the unit of analysis was the individual within each hurricane study. In an open access dataset, it is important for researchers to know what information participants provide and when. As the number of participants in a longitudinal study increases, attrition becomes more probable, and participants drop out of the study at various timepoints (Amico 2009). For the CDDI study, each of the four hurricane studies contributed data from multiple timepoints where hundreds of participants were assessed at a given number

Fig. 1 **a** Illustration of how many participants were included in each case study and at how many months post-hurricane they were assessed, **b** the presence of participants over time, separated by study. Each participant represents one value on the y-axis, and their participant identifier was unique to their respective hurricane. The horizontal axis represents participants' movement in and out of the data set over time, if read from left to right], **c** the combination of the data in **a** and **b** to show the presence of participants over time and to provide a basic summary of how many participants were or were not assessed at the various time points in each study]

of months post-hurricane. These timepoints were not consistent throughout the studies. The participants from one hurricane study did not contribute data at timepoints from another hurricane study. For instance, an individual assessed at 5 months post-hurricane (T1 for the Katrina study) did not provide data at the 3 months post-hurricane assessment (T1 for the Andrew study). Consequently, missingness was introduced into the study by design (Brincks et al. 2018). A visual was created to help researchers understand what participants provided data at what timepoints and the way in which they moved in and out of the data.

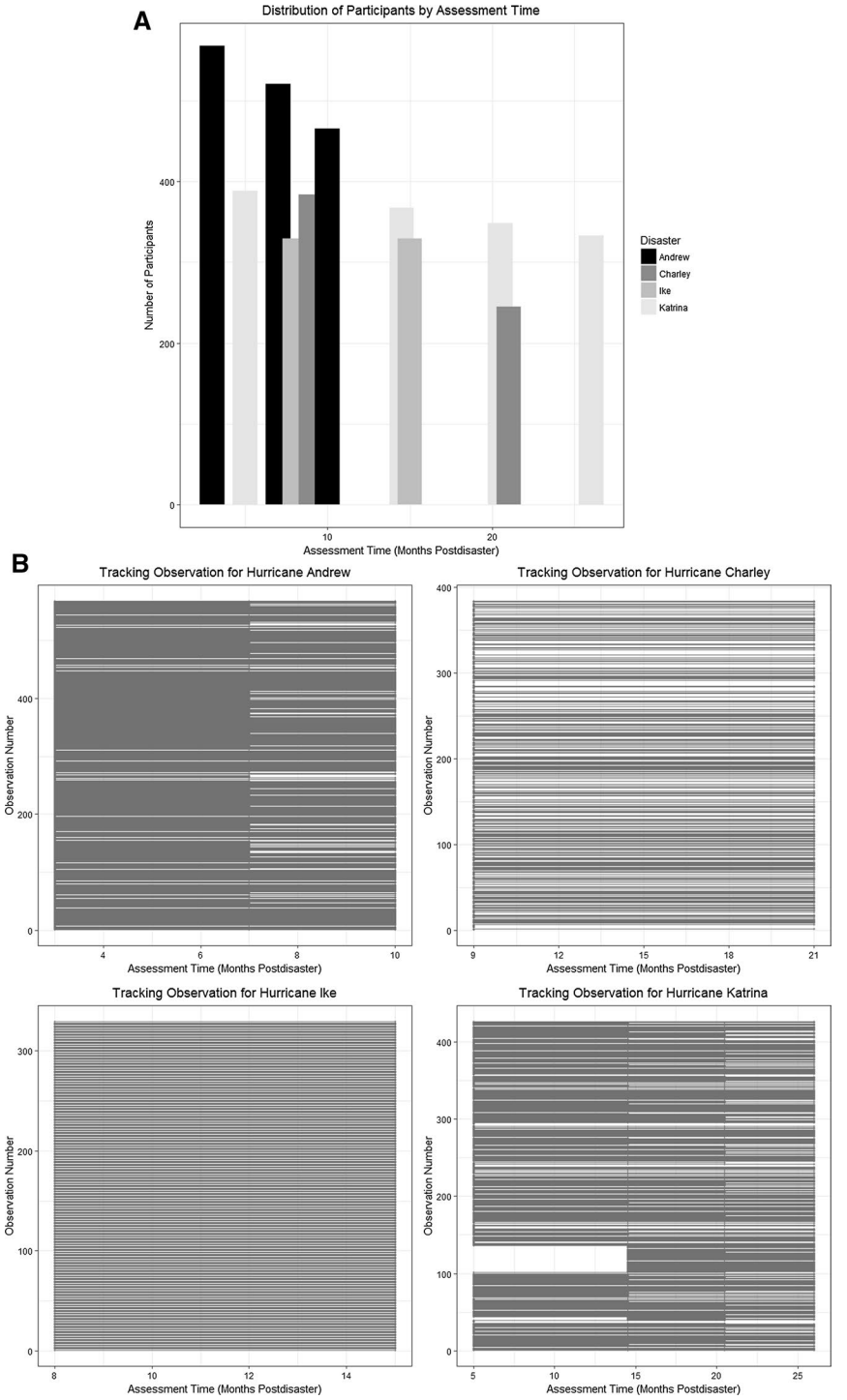
Illustrating the Challenge

The standard bar chart in Fig. 1a is used to show how many participants were included in each study and at how many months post-hurricane they were assessed. By designating each hurricane study with a specific shade of grey, the number of participants contributed by each study is evaluated. Figure 1b shows the presence of participants over time, separated by study. Each participant represents one value on the y-axis, and their participant identifier is unique to their respective hurricane. Viewing the plot from left to right, a participant's presence is tracked within each hurricane study. Figure 1c merges Fig. 1a, b to show the presence of individuals over time and to provide a basic summary of how many participants were or were not assessed at the various timepoints.

Insights from Visuals

Visualizing the data is beneficial in tracking participants and understanding individual participants' contribution to the overall study. While the basic bar chart (Fig. 1a) is useful in showing how the number of participants in the Andrew study changed from 3 to 7 to 10 months, it is unknown whether the same participants were being assessed or if new individuals were introduced to the study. For example, it was possible for a participant to return to the study for the last assessment after missing the 7-month assessment. By converting to scatter plots (Fig. 1b), the presence of participants over time is better visualized. For example, it is clear from the scatter plots that participant retention was high over the course of the Hurricane Ike study, but less so for the Hurricane Katrina study. In the Hurricane Katrina study, a number of participants were also missing at the first time point and then included in the data set at future timepoints. However, creating separate plots for each of the four hurricanes is less conducive to understanding the overall contribution of participants to the integrated dataset.

Visuals in this case provide an important illustration of insights that may be gained via visualization techniques. Figure 1c combines the value gained from both Fig. 1a, b to provide a more succinct and informative visual that simultaneously depicts the distribution and retention of participants from each of the four hurricane studies across their various timepoints. Figure 1c visualizes missingness throughout the data, which is represented by gaps in the vertical lines at each assessment timepoint. For example, there is a large gap in


A
Distribution of Participants by Assessment Time
Number of Participants
0
200
400
10
20
Assessment Time (Months Postdisaster)
Disaster
Andrew
Charley
Ike
Katrina
B
Tracking Observation for Hurricane Andrew
Observation Number
0
200
400
4
6
8
10
Assessment Time (Months Postdisaster)
Tracking Observation for Hurricane Charley
Observation Number
0
100
200
300
400
9
12
15
18
21
Assessment Time (Months Postdisaster)
Tracking Observation for Hurricane Ike
Observation Number
0
100
200
300
8
10
12
14
Assessment Time (Months Postdisaster)
Tracking Observation for Hurricane Katrina
Observation Number
0
100
200
300
400
5
10
15
20
25
Assessment Time (Months Postdisaster)

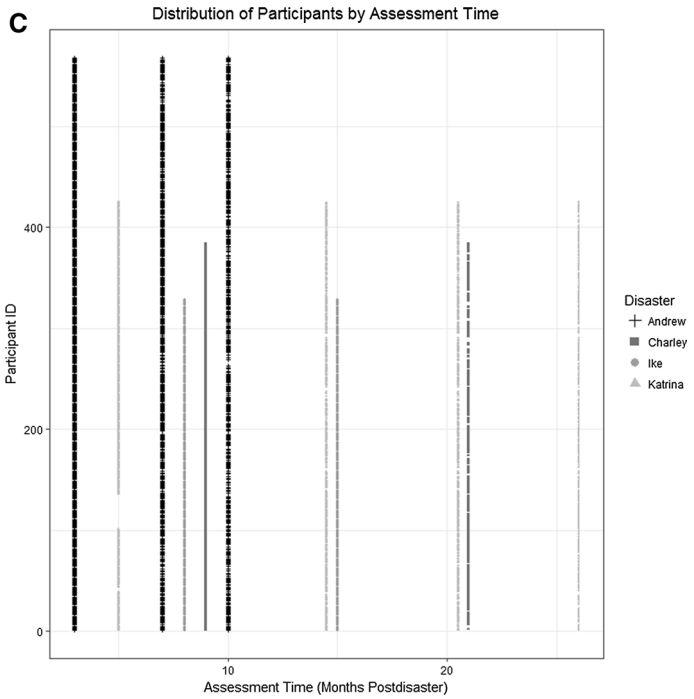


Fig. 1 (continued)

the vertical line at the first timepoint of the Hurricane Katrina study. This gap means that multiple participants were missing data for this first timepoint. These gaps are not visible at the later timepoints of the Katrina study, however, indicating that these participants were assessed at the later timepoints.

Failure to conduct a visual analysis might have caused the research team to miss an important and unique aspect of this dataset: not all Katrina participants entered the study at the first timepoint. By familiarizing themselves with the data through visuals, researchers can be better prepared to understand the original studies and to determine how to account for missing data. Missing data, particularly in longitudinal studies, may introduce problems during analysis (Graham 2009).

Challenge 2: Identifying Patterns and Trends Within the Data

When analyzing a new dataset, researchers must identify patterns and trends among the data. However, it becomes difficult to summarize the data at a glance when there are thousands of observations, multiple measures of interest, and a wide range of assessment timepoints. In the CDDI dataset, the researchers analyzed the summary scores from measures assessing various risk factors including exposure, posttraumatic stress, anxiety, social support, and life events. Because participants were assessed at multiple timepoints, the distribution of summary scores over time was analyzed. Additionally, the relationship between select risk factors was assessed using visuals.

Illustrating the Challenge

The scatter plot depicted in Fig. 2a shows the distribution of summary scores by measure at each timepoint that the measure was assessed. Different symbols are used to distinguish between data points for the four hurricane studies. The thickness of the data point is indicative of how many participants had that summary score at the given timepoint; thicker data points show that more participants received that summary score. For example, the thickness of the data points depicting exposure in the Katrina study (triangle shapes) indicate that the majority of participants in this study maintained an exposure summary score of two across all of the timepoints. Overall trends within each measure can be evaluated using the integrated data. Figure 2b illustrates patterns and relationships between perceived life threat, actual life threat, and PTSS. This visual shows the correlation between the risk factor of actual life threat (from the HURTE or HURTE-R) and PTSS at the baseline

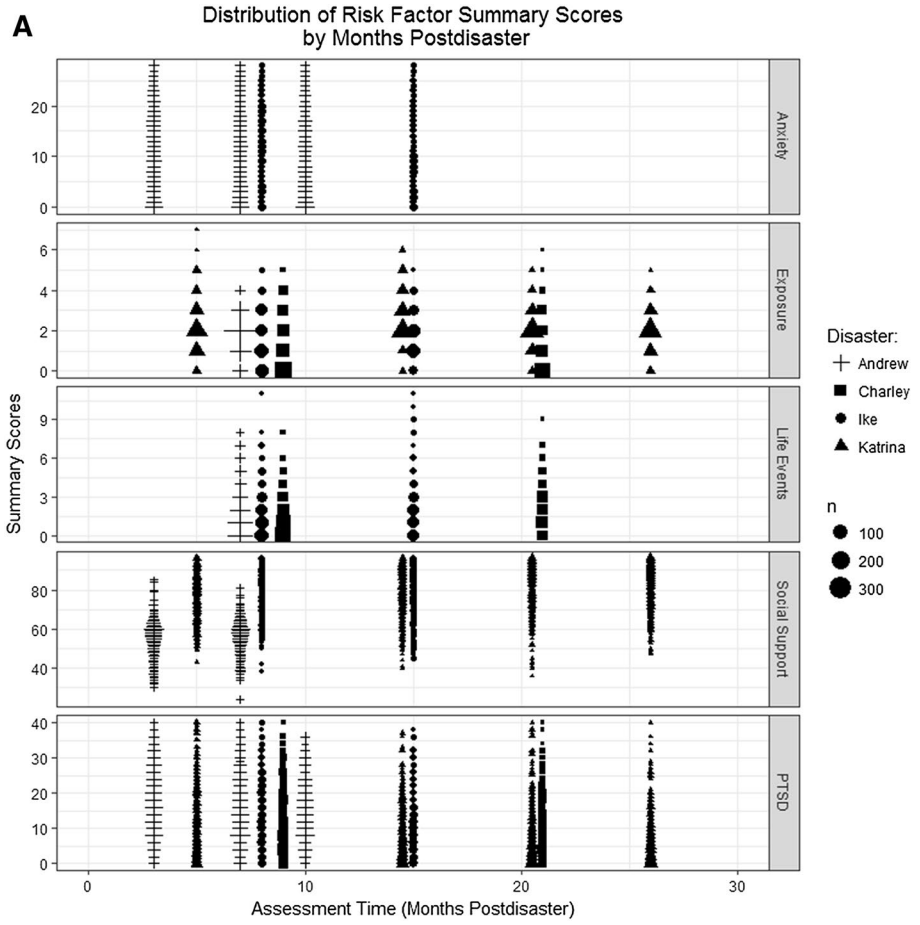


Fig. 2 **a** The distribution of summary scores by measure at each timepoint that the measure was assessed in each study], **b** the correlation between actual life threat and PTSS by depicting participants' responses to the HURTE or HURTE-R and truncated PTSD-RI questionnaires at the baseline assessment for each study]

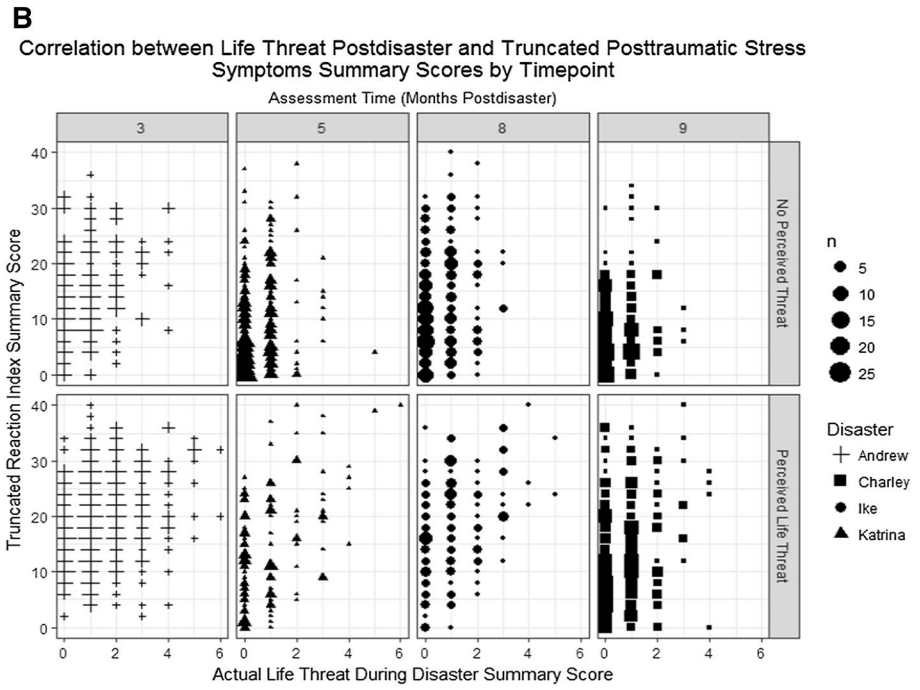


Fig. 2 (continued)

assessment for each study, stratified by whether the child perceived a life threat (bottom row) or not (top row).

Insights from Visuals

Visuals allow researchers to identify potential patterns and trends across multiple risk factors. While basic descriptive statistics could have been calculated for each of the measures within the master dataset, Fig. 2a is more effective in providing a cursory overview of changes and patterns in the available data. For instance, the number of life events slightly increased with time, while the average PTSD-RI summary score slightly decreased over time across all studies. Further, social support summary scores were generally concentrated towards the higher end of the range, while anxiety summary scores were relatively equally distributed throughout the range with a slight concentration at the lower scores. The scatter plot is valuable in visualizing the spread of the data within each summary score range and identifying trends of scores over time.

Visuals also depict patterns and relationships between measures. According to Fig. 2b, children's perception of a life threat seems to be on par with their experiences of actual life threat. Those who perceived a life threat (depicted in the bottom row of the figure) experienced greater actual life threats during the hurricane than those who did not perceive a life threat (depicted in the top row of the figure). Children who perceived a life threat also reported greater PTSS as measured by the PTSD-RI measure. These visuals suggest that there may be an interaction between perceived life threat, actual life threat, and PTSS. This

is not a question that was included in the original overall study questions, but it does raise important questions about whether children who perceive their lives to be threatened may have materially different experiences and responses to disasters. These are key questions that were gained through visualizations.

Further, the visualization indicates graphic similarities between the scatterplot distributions of the Hurricane Andrew and Hurricane Ike studies, as well as between the Hurricane Charley and Hurricane Katrina studies. Researchers should therefore consider potential factors that could account for these similarities. Through the use of visuals, researchers are able to identify potential patterns within the data that could provide insight into future analysis.

Challenge 3: Determining how to Integrate Data from Multiple Studies

The CDDI dataset combined studies initiated in 1992 (Hurricane Andrew), 2004 (Hurricane Charley), 2005 (Hurricane Katrina), and 2008 (Hurricane Ike). Over the span of 16 years, the measure used to evaluate PTSS changed. The DSM-III-R version of the PTSD-RI was used for the Hurricane Andrew study, while the DSM-IV-TR version was used for the other three studies. Although the PTSD-RI continued to assess posttraumatic stress, the questions used in the measure differed between versions. This complicated the comparison of PTSS in children from each study and the overall integration of the measure. While the newer DSM-IV-TR version used 17 of 22 questions to compute a summary score, only 10 of those questions overlapped with the older DSM-III-R version of the PTSD-RI. Visuals were useful in determining how much information was lost in truncating the measure versus how much was gained from integrating the Hurricane Andrew study.

Illustrating the Challenge

Figure 3a–d depict the relationships between various HURTE and HURTE-R summary scores of hurricane exposure and PTSD-RI summary scores in the months following the hurricanes. Figure 3a, b show the correlation between exposure (i.e., actual life threat during the hurricane, immediate loss/disruption after the hurricane, and ongoing loss/disruption since the hurricane) and PTSD-RI summary scores (truncated and total, respectively). The gradient color of the points represents the progression of time in months post-hurricane. Figure 3c, d focus on the ongoing loss/disruption since the hurricane in relation to PTSS. All timepoints of assessment are included. Figure 3a, c use the truncated PTSD-RI summary score of 10 items while Fig. 3b, d use the total PTSD-RI summary score of 17 items.

Insights from Visuals

Comparing the truncated PTSD-RI summary score using the 10 comparable items versus the total PTSD-RI summary score using the 17 items, the scatter plots show that the relative distribution of scores is similar across the hurricanes. Although the range of scores is greater for the 17 items, the spread of scores is visually congruent between the plots in Fig. 3a, b and between the plots in Fig. 3c, d. Because the distribution of scores is visually comparable, the benefit of including the Andrew study data may outweigh the cost of excluding items from the study. This visual persuaded the investigators to retain the Andrew study—an older, but critical, dataset—since the inclusion of these data does not

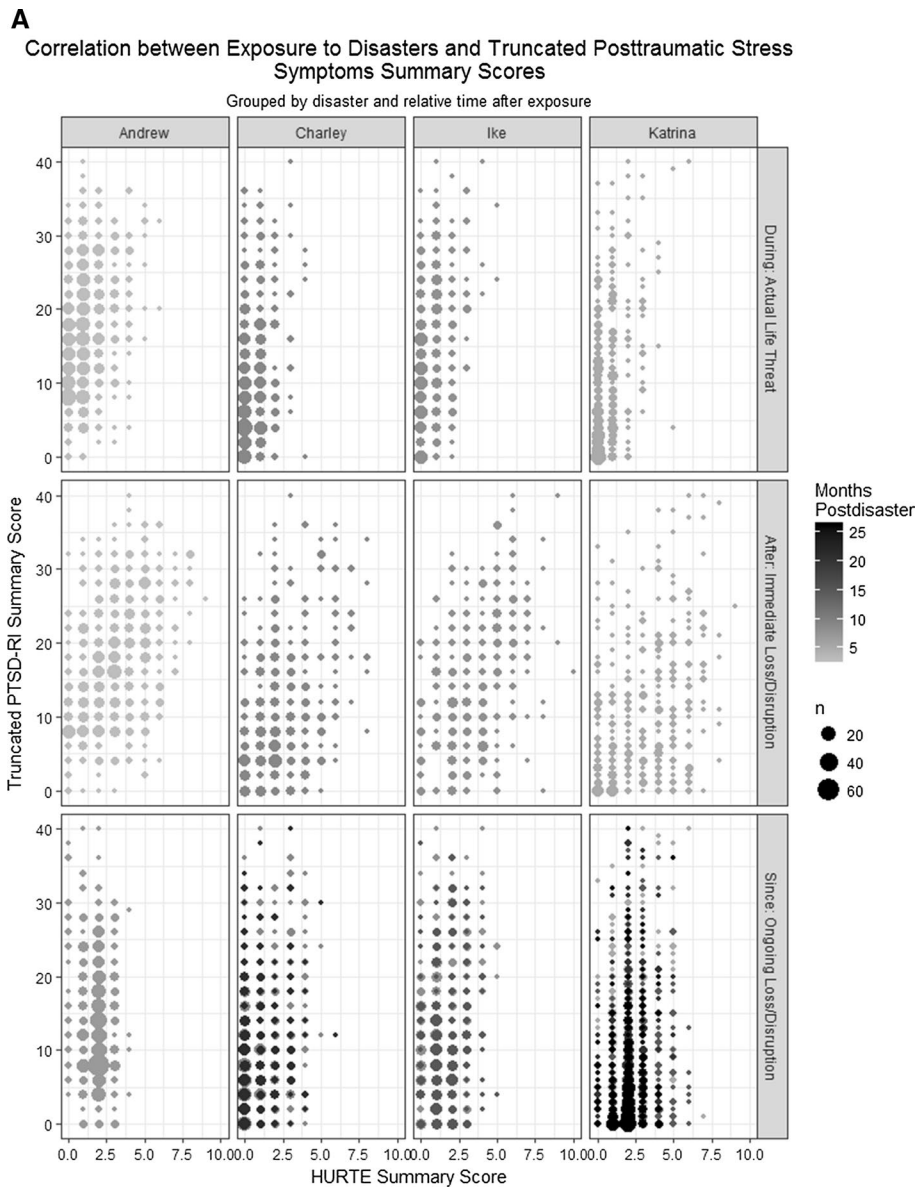


Fig. 3 **a** The correlation between exposure (measured by actual life threat during the hurricane, immediate loss/disruption after the hurricane, and ongoing loss/disruption since the hurricane) with truncated PTSD-RI summary scores of 10 items, grouped by disaster and relative time after exposure], **b** the correlation between exposure (measured by actual life threat during the hurricane, immediate loss/disruption after the hurricane, and ongoing loss/disruption since the hurricane) with total PTSD-RI summary scores of 17 items, grouped by disaster and relative time after exposure], **c** the experience of ongoing loss/disruption since the hurricane in relation to PTSS based on truncated PTSD-RI scores of 10 items], **d** the experience of ongoing loss/disruption since the hurricane in relation to PTSS based on total PTSD-RI scores of 17 items]

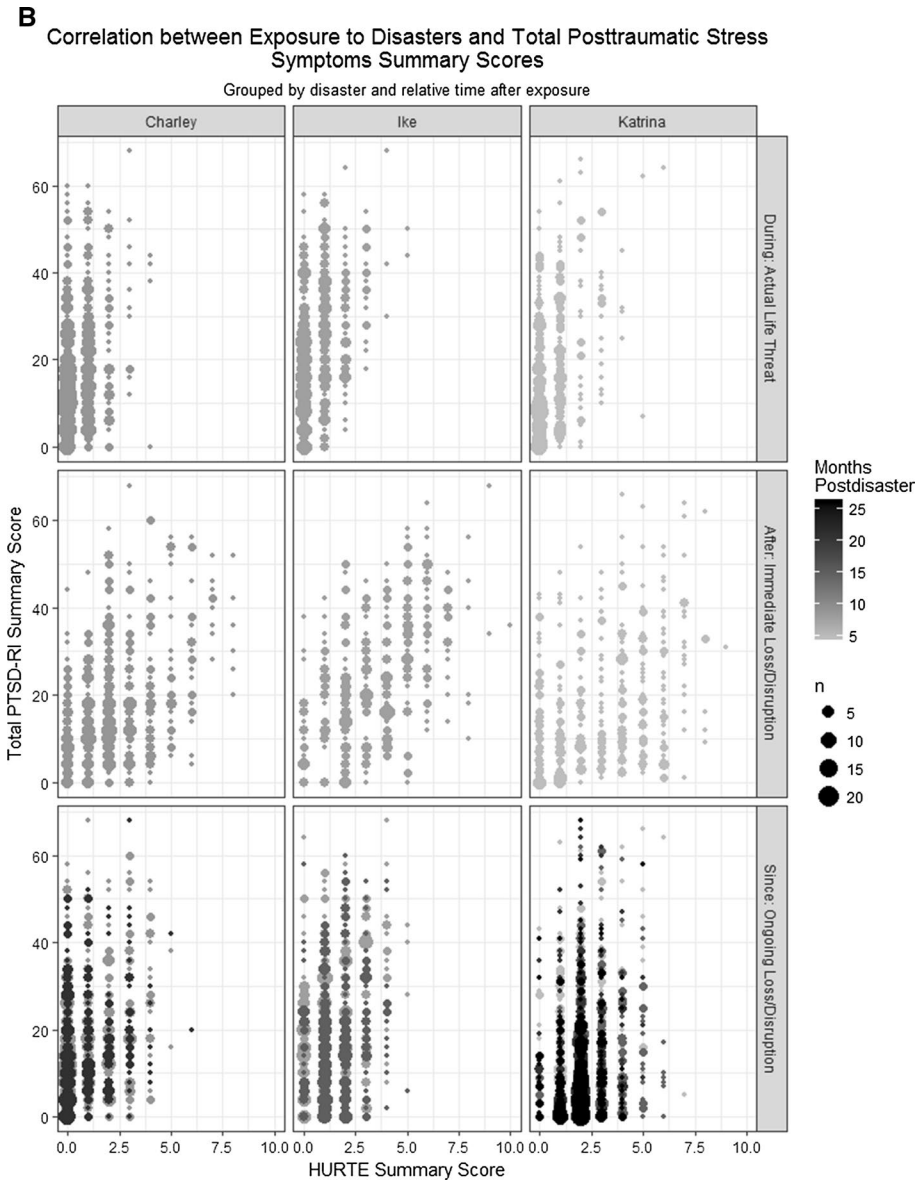


Fig. 3 (continued)

appear to affect the relationships between the variables. The Andrew study data add value to the integrated study as the data show how responses to hurricanes may have changed over time. Further, as explored in these visuals, the issue of whether dataset source moderates overall study findings is critical and warrants further exploration in the final analyses for the overall study.

Addressing heterogeneity in outcome measures is a critical component of harmonizing data for IDA (Brincks et al. 2018). Researchers may refer to these plots in determining

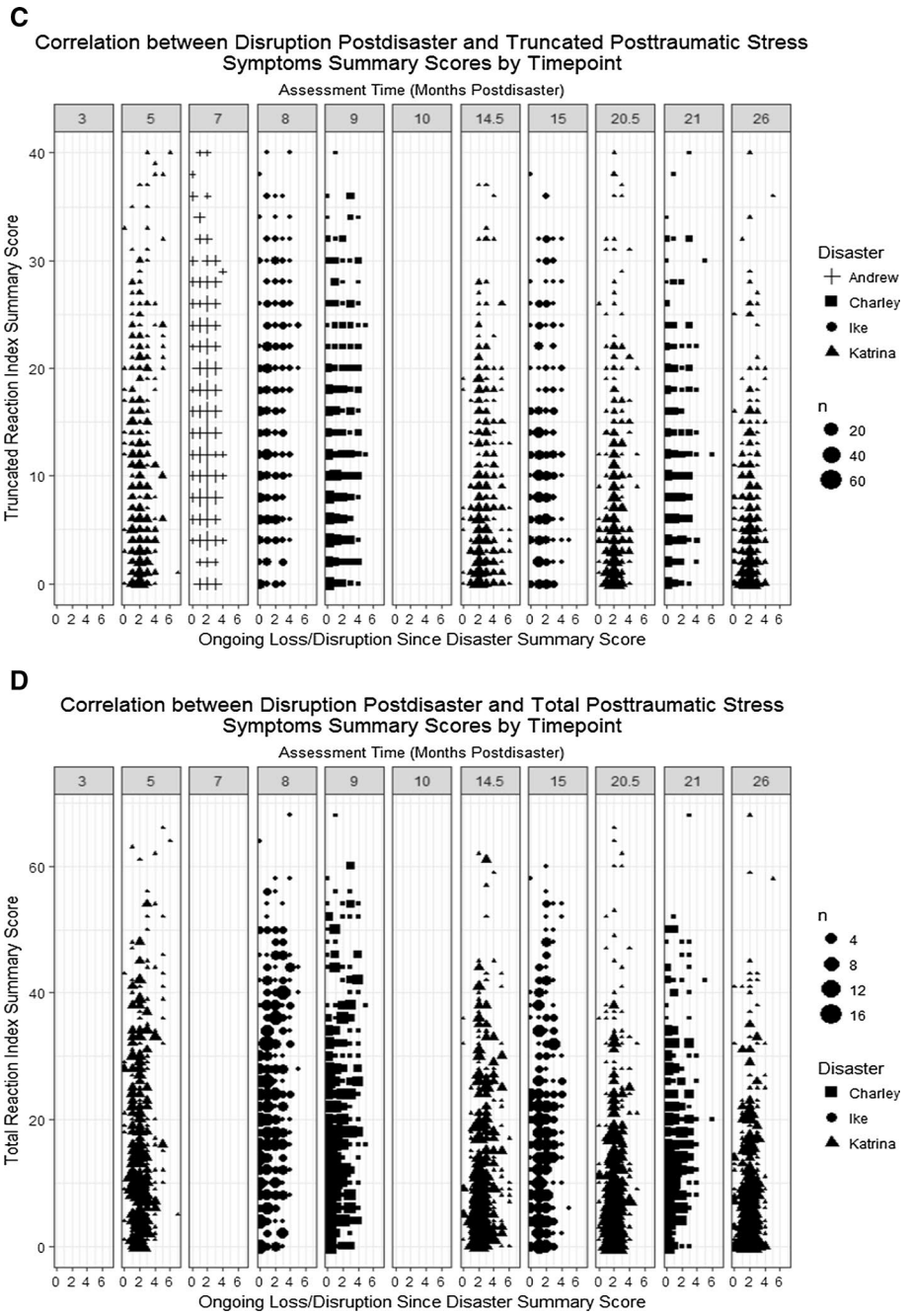


Fig. 3 (continued)

whether to use the truncated or total summary scores for analysis. These visuals depict the tension between what information is available in an integrated dataset versus what information is substantive for researchers. By visualizing the data, researchers can be better prepared to make decisions regarding the integration of the individual studies.

Discussion

The importance of visualizing data is growing in tandem with the increasing availability of open access data. Researchers must learn how to efficiently understand new data that are becoming available. Particularly for open access datasets, visuals are key in helping researchers become familiar with the data at hand. Creating effective and informative visuals allows researchers to consider potential patterns within and among risk factors and outcome measures. By visualizing the data, researchers are better able to comprehend relationships between variables in an integrated dataset and are consequently better prepared for statistical analysis.

Recommendations

Despite the value of visualizing data, many researchers are not trained in programming languages used to produce effective visuals. While there are a variety of options for creating visuals, R was used in this paper due to its wide flexibility of colors, graphs, and symbols (Wickham 2009). R is also free to use, and there is a wealth of online resources available to help researchers to develop custom visuals to represent their dataset. R can be used to create both simple graphics and advanced statistical graphics, depending upon the needs of the researcher.

Despite these benefits, R is not commonly taught in most graduate programs. For instance, the graduation requirements for some of the top clinical psychology graduate schools such as University of California, Los Angeles (2017), University of North Carolina at Chapel Hill (2017), and Emory University Department of Psychology (2017) require fundamental statistics courses but do not require courses that train students in programming languages such as SAS or R. While certain faculty members may choose to train their students in the use of R within statistical courses outside of formal courses in statistical programming, it is not clear to what extent it is being taught in graduate programs. Without such training, it is more difficult for researchers to develop the skills necessary to create visuals. Existing courses that train researchers in the statistical software program R are often expensive. For instance, Stats Camp provides a statistical methods training seminar in R Programming at the graduate and post-graduate level over the course of 5 days at a cost of $1095 for students and $1795 for professionals (Stats Camp 2018).

More accessible opportunities for programming training need to be offered for researchers. As an initial step towards addressing these issues, we included with this article an online appendix with the R code used to create the visuals shown in this paper (Online Resource 1). The code is annotated throughout to indicate how altering the code will create alternate visuals, in order to demonstrate how the issues described here would apply to other datasets. For individual researchers interested in accessing open data, one initial resource for open data would be the Inter-University Consortium for Political and Social Research, which contains a data archive of over 250,000 files of research (ICPSR 2018).

Additionally, the creation and use of effective visuals in research is impeded by current journal standards. Standard manuscript requirements generally allow black-and-white figures, but journals may charge authors excessive fees to print in color. This may dissuade researchers from spending time creating color visuals. For example, in the *Journal of the American Medical Association* (JAMA Network 2018), the fee for printing in color ranges from $1580 per page for a full page with matched color to $4925 per page for five colors (JAMA Network 2017). The *Journal of Consulting and Clinical Psychology* charges $900 for one figure in color, an additional $600 for the second figure, and $450 for each subsequent figure (Journal of Consulting and Clinical Psychology 2018). The *Journal of Pediatric Psychology* charges $600 per color page (Journal of Pediatric Psychology 2018). These fees deter authors from generating colorful visuals that may help both the author and the reader better understand the data. Compared to color figures, black-and-white figures are more difficult to comprehend; for instance, it is difficult to distinguish between the various shades of grey in Fig. 1a. The color figures included as part of the online supplement for this article (Online Resource 2) are more effective in illustrating patterns among the data and distinguishing between the hurricane studies.

Due to the current costs of color figures, it is important to consider an alternative way to present layers of data that does not rely on the use of colors. One alternative to presenting color-coded data is the use of shapes to distinguish between variables in black-and-white figures (see Fig. 2a). Shapes are preferable to different shades of grey, which are often difficult to distinguish between. R also offers a black-and-white theme for figures that removes the grey background and creates a greater contrast between the plot and the background.

Future Directions

The visuals presented in this paper are intended to bring about a basic understanding of new open access data. These figures represent just a few examples of how visuals can enhance the understanding of large datasets. With more time and experience there is the potential for researchers to create a wider variety of statistically advanced visuals. For example, the next step in data visualization may include plotting regression lines and latent growth models (Brincks et al. 2018). Given their cross-disciplinary applicability and potential for simplifying the data comprehension and utilization process, there is no limit to the ways that visuals can enhance researchers' current work with integrated datasets.

Of central concern to the future of data visualization use are the aforementioned issues of training and accessibility. In order to encourage researchers to make use of open access data, adjustments must be made in research-based graduate programs as well as in free-standing statistical training organizations. Currently, courses through which graduate students can gain instruction in the use of software programs such as R are rarely built into the core curriculum or offered at all. Graduate programs must encourage students to learn and utilize these tools by incorporating them into existing statistical training curricula.

Further, the cost of seeking training outside of graduate programs with organizations such as Stats Camp discourages and prevents most graduate students with limited financial resources from pursuing such training. In order to address this barrier, graduate programs might consider subsidizing the cost of these trainings. A subsidization strategy may enable graduate programs to offer this valuable skillset to students without making changes to existing curricula or program requirements. With ongoing promotion and implementation of this type of tool, a wider array of visuals can be created that will further expand the knowledge gained from open access datasets.

Funding This study was funded through a grant from the National Institute of Mental Health (1R03MH113849-01). Research time for this paper was partially supported by National Science Foundation Grant No. 1634234. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Institute of Mental Health or the National Science Foundation.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study. Active parental consent and written child assent were required for study participation.

References

Amico, K. R. (2009). Percent total attrition: A poor metric for study rigor in hosted intervention designs. *American Journal of Public Health, 99*(9), 1567–1575. https://doi.org/10.2105/AJPH.2008.134767.

Brincks, A., Montag, S., Howe, G. W., Shi, H., Siddique, J., Soyeon, A., et al. (2018). Addressing methodologic challenges and minimizing threats to validity in synthesizing findings from individual-level data across longitudinal randomized trials. *Prevention Science, 19,* S60–S73. https://doi.org/10.1007/s11121-017-0769-1.

Chen, H. M. (2017). Real-world uses for information visualization in libraries. *Library Technology Reports, 53*(3), 21.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–100. https://doi.org/10.1037/a0015914.

Danzi, B. A., & La Greca, A. M. (2016). DSM-IV, DSM-5, and ICD-11: Identifying children with posttraumatic stress disorder after disasters. *Journal of Child Psychology and Psychiatry, 57*(12), 1444–1452. https://doi.org/10.1111/jcpp.12631.

Emory University Department of Psychology. (2017). *Courses*. Retrieved August 25, 2018 from http://psychology.emory.edu/home/graduate/clinical/courses.html.

Executive Order No. 13642, Making Open and Machine Readable the New Default for Government Information, Signed: May 9, 2013.

Frederick, C. J. (1985). Selected foci in the spectrum of posttraumatic stress disorders. In J. Laube & S. A. Murphy (Eds.), *Perspectives on disaster recovery* (pp. 110–131). Norwalk, CT: Appleton & Lange.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530.

Greenberg, M. T., Leitch, C. J., & Siegel, J. M. (1983). The nature and importance of attachment relationships to parents and peers during adolescence. *Journal of Youth and Adolescence, 12*(5), 373–386. https://doi.org/10.1007/BF02088721.

Harter, S. (1985). *Manual for the social support scale for children*. Denver, CO: University of Denver.

Harwood, A., & Mayer, A. (2016). Big data and semantic technology: A future for data integration, exploration and visualisation. *Statistical Journal of the IAOS, 32*(4), 613. https://doi.org/10.3233/SJI-160989.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM, 53*(6), 59–67. https://doi.org/10.1145/1743546.1743567.

Huijboom, N., & Van Den Broek, T. (2011). Open data: An international comparison of strategies. *European Journal of ePractice, 12,* 1–13.

Inter-University Consortium of Political and Social Research (ICPSR). (2018). *About the organization*. Retrieved August 25, 2018 from https://www.icpsr.umich.edu/icpsrweb/content/about.

JAMA Network. (2017). *Print edition rate card*. Retrieved August 25, 2018 from https://jamanetwork.com/DocumentLibrary/Advertising/jama_rates_2017.pdf.

JAMA Network. (2018). *For authors*. Retrieved August 25, 2018 from https://jamanetwork.com/journals/jama/pages/for-authors.

Johnson, J. H. (1986). *Life events as stressors in childhood and adolescence*. Newbury Park, CA: Sage Publications.

Journal of Consulting and Clinical Psychology. (2018). *Manuscript submission*. Retrieved August 25, 2018 from http://www.apa.org/pubs/journals/ccp/?tab=4.

Journal of Pediatric Psychology. (2018). *Instructions to authors*. Retrieved August 25, 2018 from https://academic.oup.com/jpepsy/pages/msprep_submission.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1–8. https://doi.org/10.1109/2945.981847.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. Los Angeles, CA: Sage Publications.

La Greca, A. M., Lai, B. S., Joormann, J., Auslander, B. B., & Short, M. A. (2013a). Children's risk and resilience following a natural disaster: Genetic vulnerability, posttraumatic stress, and depression. *Journal of Affective Disorders, 151*(3), 860–867. https://doi.org/10.1016/j.jad.2013.07.024.

La Greca, A. M., Lai, B., Llabre, M. M., Silverman, W. K., Vernberg, E. M., & Prinstein, M. J. (2013b). Children's postdisaster trajectories of posttraumatic stress symptoms: Predicting chronic distress. *Child & Youth Care Forum, 42,* 351–359.

La Greca, A. M., Lai, B., Silverman, W. K., & Jaccard, J. (2010). Hurricane-related exposure experiences and stressors, other life events, and social support: Concurrent and prospective impact on children' s persistent posttraumatic stress symptoms. *Journal of Consulting and Clinical Psychology, 78*(6), 794–805. https://doi.org/10.1037/a0020775.

La Greca, A. M., Silverman, W. K., Vernberg, E. M., & Prinstein, M. J. (1996). Symptoms of posttraumatic stress in children after Hurricane Andrew: A prospective study. *Journal of Consulting and Clinical Psychology, 64*(4), 712–723. https://doi.org/10.1037/0022-006X.64.4.712.

Lai, B. S., Beaulieu, B., Ogokeh, C., Self-Brown, S., & Kelley, M. L. (2015a). Mother and child reports of hurricane related stressors: Data from a sample of families exposed to Hurricane Katrina. *Child & Youth Care Forum, 44*(4), 549–565. https://doi.org/10.1007/s10566-014-9289-3.

Lai, B. S., La Greca, A. M., Auslander, B. A., & Short, M. B. (2013). Children's symptoms of posttraumatic stress and depression after a natural disaster: Comorbidity and risk factors. *Journal of Affective Disorders, 146*(1), 71–78. https://doi.org/10.1016/j.jad.2012.08.041.

Lai, B. S., La Greca, A. M., & Llabre, M. M. (2014). Children's sedentary activity after hurricane exposure. *Psychological Trauma: Theory, Research, Practice, and Policy, 6*(3), 280–289. https://doi.org/10.1037/a0033331.

Lai, B. S., Tiwari, A., Beaulieu, B. A., Self-Brown, S., & Kelley, M. (2015b). Hurricane Katrina: Maternal depression trajectories and child outcomes. *Current Psychology, 34*(3), 515–523. https://doi.org/10.1007/s12144-015-9338-6.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the symposium on principles of database systems (PODS)* (pp. 233–246). Madison, WI: ACM.

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2017). Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics, 23*(3), 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147.

Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. *PLoS Biology, 9*(12), e1001195. https://doi.org/10.1371/journal.pbio.1001195.

National Institutes of Health. (2017). *NIH data sharing repositories*. Retrieved August 25, 2018 from https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html.

National Science Foundation. (2018). *Open data at NSF*. Retrieved August 25, 2018 from https://www.nsf.gov/data/.

Pynoos, R. S., Frederick, C., Nader, K., Arroyo, W., Steinberg, A., Eth, S., et al. (1987). Life threat and posttraumatic stress in school-age children. *Archives of General Psychiatry, 44*(12), 1057–1063. https://doi.org/10.1001/archpsyc.1987.01800240031005.

R Core Team. (2017). *R: A language and environment for statistical computing*. Retrieved August 25, 2018 from http://www.R-project.org/.

Reynolds, C. R., & Richmond, B. O. (1997). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 25*(1), 15–20. https://doi.org/10.1023/A:1025751206600.

Reynolds, C. R., & Richmond, B. O. (2008). *RCMAS-2: Revised children's manifest anxiety scale* (2nd ed.). Torrance, CA: Western Psychological Services.

Romero, S., Birmaher, B., Axelson, D. A., Iosif, A. M., Williamson, D. E., Gill, M. K., et al. (2009). Negative life events in children and adolescents with bipolar disorder. *The Journal of Clinical Psychiatry, 70*(10), 1452–1460. https://doi.org/10.4088/JCP.08m04948gre.

RStudio Team. (2018). *RStudio: Integrated development environment for R*. Retrieved August 25, 2018 from http://www.rstudio.com/.

Self-Brown, S., Lai, B. S., Harbin, S., & Kelley, M. L. (2014). Maternal posttraumatic stress disorder symptom trajectories: The long-term impact on youth following Hurricane Katrina. *International Journal of Public Health, 59*(6), 957–965. https://doi.org/10.1007/s00038-014-0596-0.

Self-Brown, S., Lai, B. S., Thompson, J. E., McGill, T., & Kelley, M. L. (2013). Posttraumatic stress disorder symptom trajectories in Hurricane Katrina affected youth. *Journal of Affective Disorders, 147*(1–3), 198–204. https://doi.org/10.1016/j.jad.2012.11.002.

Simon, P. (2014). *The visual organization: Data visualization, big data, and the quest for better decisions*. Hoboken, NJ: Wiley.

Skiba, D. J. (2014). The connected age: Big data & data visualization. *Nursing Education Perspectives (National League for Nursing), 35*(4), 267–269.

Spell, A. W., Kelley, M. L., Wang, J., Self-Brown, S., Davidson, K. L., Pellegrin, A., et al. (2008). The moderating effects of maternal psychopathology on children's adjustment post-Hurricane Katrina. *Journal of Clinical Child & Adolescent Psychology, 37*(3), 553–563. https://doi.org/10.1080/15374410802148210.

Stats Camp. (2018). *R programming for data science*. Retrieved August 25, 2018 from https://www.statscamp.org/summer-camp/statistical-programming-data-analysis.

Steinberg, A., Brymer, M., Decker, K., & Pynoos, R. (2004). The University of California post-traumatic stress disorder reaction index. *Current Psychiatry Reports, 6,* 96–100. https://doi.org/10.1007/s11920-004-0048-2.

The White House. (2009). *Transparency and open government: Memorandum for the heads of executive departments and agencies*. Washington, DC: OMB.

University of California, Los Angeles. (2017). *Graduate program in psychology handbook (2017–2018)*. Retrieved August 25, 2018, from https://ucla.app.box.com/s/m37p3vq4euus53gxwqumk0ni7irnfsqe.

University of North Carolina at Chapel Hill. (2017). *Doctoral program in clinical psychology program handbook*. Retrieved August 25, 2018 from https://clinicalpsych.unc.edu/files/2015/06/Clinical-Handbook.pdf.

Vernberg, E. M., La Greca, A. M., Silverman, W. K., & Prinstein, M. J. (1996). Prediction of posttraumatic stress symptoms in children after Hurricane Andrew. *Journal of Abnormal Psychology, 105*(2), 237–248. https://doi.org/10.1037/0021-843X.105.2.237.

Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. *First Monday, 18*(10). Retrieved August 25, 2018 from http://firstmonday.org/ojs/index.php/fm/article/view/4878.

Walport, M., Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *Lancet, 377*(9765), 537–539. https://doi.org/10.1016/S0140-6736(11)61211-7.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Affiliations

Hazel J. Shah[1] · Betty S. Lai[2] · Audrey J. Leroux[3] · Annette M. La Greca[4] · Courtney A. Colgan[2] · Julia Medzhitova[2]

[1] Centers for Disease Control and Prevention, 1600 Clifton Road NE, C-09, Atlanta, GA 30333, USA

[2] Lynch School of Education and Human Development, Boston College, 316A Campion Hall, 140 Commonwealth Ave, Chestnut Hill, MA 02467, USA

[3] Department of Educational Policy Studies, Georgia State University, P.O. Box 3977, Atlanta, GA 30302, USA

[4] Department of Psychology, University of Miami, P.O. Box 249229, Coral Gables, FL 33124, USA